Synthese (2015) 192:4031–4055
DOI 10.1007/s11229-015-0729-x

Deflationary truth and the ontology of expressions

Carlo Nicolai[1]

Received: 25 September 2014 / Accepted: 14 March 2015 / Published online: 24 April 2015

Abstract The existence of a close connection between results on axiomatic truth and the analysis of truth-theoretic deflationism is nowadays widely recognized. The first attempt to make such link precise can be traced back to the so-called conservativeness argument due to Leon Horsten, Stewart Shapiro and Jeffrey Ketland: by employing standard Gödelian phenomena, they concluded that deflationism is untenable as any adequate theory of truth leads to consequences that were not achievable by the base theory alone. In the paper I highlight, as Shapiro and Ketland, the irreducible nature of truth axioms with respect to their base theories. But, I argue, this does not immediately delineate a notion of truth playing a substantial role in philosophical or scientific explanations. I first offer a refinement of Hartry Field's reaction to the conservativeness argument by distinguishing between metatheoretic and object-theoretic consequences of the theory of truth and address some possible rejoinders. In the resulting picture, truth is an irreducible tool for metatheoretic ascent. How robust is this characterizaton? I test it by considering: (i) a recent example, due to Leon Horsten, of the alleged explanatory role played by the truth predicate in the derivation of Fitch's paradox; (ii) an essential weakening of theories of truth analyzed in the first part of the paper.

Keywords Truth theoretic deflationism · Axiomatic theories of truth · Conservativeness argument · Relative interpretability

Io stimo più il trovar un vero, benché di cosa leggiera, che 'l disputar lungamente delle massime questioni senza conseguir verità nissuna.

Galileo Galilei.

✉ Carlo Nicolai
carlo.nicolai6@gmail.com; carlo.nicolai@philosophy.ox.ac.uk

[1] Faculty of Philosophy, University of Oxford, Woodstock Road, Oxford, UK

Science craves truth. This, it may be said, is the expression of an indispensable and ultimately useful psychological attitude of the scientist. Or, from a darker perspective, a rhetorical residual in language of a lost epistemological paradise. Science craves truth*s*, says the deflationist. Asserting that a sentence is true is tantamount to asserting the sentence itself. The rules capturing the disquotational character of the truth predicate can be seen as meaning postulates that offer a sufficient, although possibly partial, analysis of the concept of truth.

There is more about deflationary truth, however: the truth predicate enables one to perform linguistic operations, above all blind ascriptions, and expressing facts otherwise not expressible, such as the ones encapsulated in general claims. And it is even less controversial to state what the concept of truth, according to the deflationist, is not: it is not a predicate in the usual sense, as it does not pick out a unique set of objects forming its extension, and it is not a notion endowed with explanatory power.

It is no surprise, then, that axiomatic theories of truth have been invoked to investigate the logical properties of a deflationary notion of truth. On the one hand they capture the indispensability of the truth predicate for certain linguistic purposes: a mathematical object theory containing a sufficient amount of syntax is extended with axioms corresponding to these meaning postulates and governing a notion (a predicate) that, so characterized, is not definable in the object theory. On the other, the axioms represent our best chance to understand the concept of truth: a partial account of it is given without appealing to a stronger metalanguage.

But is it really so? Can axiomatic truth address questions and clarify arguments on the nature of truth? Research on axiomatic truth began in fact with different aims: a clear instrumentalist flavour (Feferman 1984, 1991), or with a direct application to paradoxes (Cantini 1989; Friedman and Sheard 1987; Reinhardt 1986). The socratic question, *what is truth?*, did not represent the leitmotif of the development of axiomatic truth. A well-known effort to emphasize this link was represented by the so-called *conservativeness argument*. According to the argument, a deflationary acceptable theory of truth has to be a conservative extension of its base theory.

This work should be seen as a philosophical spin-off of two rather technical papers (Leigh and Nicolai 2013; Nicolai 2015). The next, introductory section describes formal and philosophical preliminaries. In Sect. 2 I elaborate on Leigh and Nicolai (2013) by defending Field's response to the conservativeness argument. A clear characterization of compositional, typed truth will arise from this analysis. In the final section, this understanding of typed truth will be tested by looking at variations of the original setting. In the paper I will appeal to formal results without giving full proofs: more details can be found in the above-mentioned works of the author.

1 Reductions and truth theoretic content

1.1 Disquotational and compositional truth

The systems of truth that will be considered in this work will be constructed over theories containing a sufficient amount of arithmetic. Throughout this section, Peano Arithmetic represents our base theory; we thus assume a canonical Gödel numbering of $\mathcal{L}_{PA}$-expressions that associates to any string e its unique Gödel number $\#e$.[1] We will denote with $\ulcorner e \urcorner$ the term of $\mathcal{L}_{PA}$ representing $\#e$ in *PA*. By expanding $\mathcal{L}_{PA}$ with a unary truth predicate T allowed to appear into the instances of induction schema, and by adding to *PA* formulated in the new language axioms of the form

$$T\ulcorner \sigma \urcorner \leftrightarrow \sigma \qquad Tb$$

for all sentences σ of $\mathcal{L}_{PA}$, we obtain the theory *TB*. *TB* seems to capture an important feature that the deflationist ascribes to truth, its disquotational character. Unfortunately, *TB* fares much worse in terms of capturing general claims: every generalization provable in *TB* is only a finite generalization (cf. Halbach 2011, Theorem 7.6).[2]

Once the inductive clauses of Tarski's definition of truth are added to *TB* as new axioms, however, many general claims become provable.[3] By letting $s, t, \ldots$, $\varphi, \psi, \chi \ldots$ and $\sigma, \tau, \ldots$, possibly with indices, be metavariables ranging over closed terms, formulas and sentences of $\mathcal{L}_{PA}$, the theory *CT* is thus obtained by adding to *PA* formulated in $\mathcal{L}_{PA} \cup \{T\}$ — again with T allowed into instances of the induction schema — the following sentences:[4]

$$\forall \ulcorner s \urcorner, \ulcorner t \urcorner \big(T\ulcorner s = t \urcorner \leftrightarrow val(\ulcorner s \urcorner) = val(\ulcorner t \urcorner)\big) \qquad \text{T1}$$

$$\forall \ulcorner \sigma \urcorner (T\ulcorner \neg\sigma \urcorner \leftrightarrow \neg T\ulcorner \sigma \urcorner) \qquad \text{T2}$$

$$\forall \ulcorner \sigma \urcorner \forall \ulcorner \tau \urcorner (T\ulcorner \sigma \wedge \tau \urcorner \leftrightarrow T\ulcorner \sigma \urcorner \wedge T\ulcorner \tau \urcorner) \qquad \text{T3}$$

$$\forall \ulcorner \varphi \urcorner (Sent_{\mathcal{L}_{PA}}(\ulcorner \forall v\varphi \urcorner) \rightarrow (T\ulcorner \forall v\varphi \urcorner \leftrightarrow \forall \ulcorner t \urcorner\, T\, sub(\ulcorner \varphi \urcorner, \ulcorner v \urcorner, \ulcorner t \urcorner))) \qquad \text{T4}$$

Quantification over codes of closed terms, formulas and sentences of $\mathcal{L}_{PA}$ should be understood in the standard way. Moreover, in T1 and T4 *val* and *sub* have to be understood as the standard, *PA*-definable evaluation and substitution functions: crucially, they are devised to yield

$$val(\ulcorner s \urcorner) = s \text{ and } sub(\ulcorner \varphi(v) \urcorner, \ulcorner v \urcorner, \ulcorner z \urcorner) = \ulcorner \varphi(z) \urcorner$$

[1] There is the temptation, vastly satisfied by the philosophical literature on the argument, to reason in terms of arbitrary base theories. This looks like a hasty move: many results mentioned below that hold of *PA* are no more than conjectures for weaker mathematical base theories, for instance. An important feature of the theories that will be introduced in Sects. 2 and 3 is given by the possibility of reasoning in terms of arbitrary base theories. For further details on coding and notational conventions we refer to Leigh and Nicolai (2013) or Halbach (2011). In particular, we only consider natural codings.

[2] Furthermore, *TB* is (proof-theoretically) conservative over *PA* (for the definition of conservativeness, cf. Sect. 1.4): this is achieved by noticing that any proof of a $\mathcal{L}_{PA}$-sentence σ in *TB* can be transformed in a proof of σ in *PA* by replacing the finitely many instances of *Tb* with a partial truth predicate definable in *PA*. Despite the conservativity of *TB* over *PA*, it is not the cases that every model of *PA* can be expanded to a model of *TB*. Models of *PA* that expand to models of *TB* are in fact such that the theory of the model $Th(\mathcal{M})$ (i.e. the set of standard sentences true in $\mathcal{M}$) is coded by some $a \in |\mathcal{M}|$. One can easily find models of *PA* that do not enjoy this property. For a proof of the (proof-theoretic) conservativity of *TB* over *PA*, we refer again to Halbach (2011).

[3] Including the general principle of contradiction — essentially, that the extension of the truth predicate is consistent — whose unprovability was one of the sources of Tarksi's skepticism on *TB*. Cf. Tarski (1956a).

[4] Assuming that $\mathcal{L}_B$ features only $\neg, \wedge, \forall$ as logical constants.

for any $\mathcal{L}_{PA}$-formula $\varphi(v)$, closed term s and term z.

1.2 Compositionality, reflective adequacy and generalizations

CT satisfies a requirement generally imposed on axiomatic truth: its truth predicate is compositional. The deflationist's attitude towards Tarski's inductive clauses, however, is still a matter of dispute. Horwich (1998) and Field (1994) seem to be content with the fact that the T-schema is capable of recovering each instance of those principles. If a compact logic is employed, however, it can promptly be seen that the infinite set of instances of a compositional principle does not enable one to derive the corresponding universally quantified statement.[5] Field (1999, 2005) offers more sympathetic remarks concerning compositional principles. In the first case, as we shall see, Field labels as compatible with deflationism a theory of truth based on them; in the second case it is held the view that the universally quantified versions of compositional principles can be derived from the T-schema together with a special rule of substitution (Field 2005, pp.13–14). Moreover, compositionality is nowadays widely recognized as an important adequacy requirement for axiomatic theories of truth [cf. the lists of desiderata recently proposed by Halbach and Horsten (2005) and Leitgeb (2007)]; if axiomatic truth is relevant to deflationism at all, therefore, theories that feature compositional principles as axioms, such as *CT*, have to be preferred to other choices.[6]

Another requirement generally imposed to theories of truth we are interested in and, as it shall be clear later, a premise of the conservativeness argument, is what has been dubbed *Reflective Adequacy* in Ketland (1999): an adequate theory of truth should be capable of declaring all theorems of the base theory true. In the language under consideration, this translates into the provability, in the theory of truth, of the *global reflection principle* for *PA*:

$$\forall \ulcorner \sigma \urcorner (Bew_{PA}(\ulcorner \sigma \urcorner) \rightarrow T\ulcorner \sigma \urcorner), \qquad GR_{PA}$$

where $Bew_{PA}(x)$ is a canonical provability predicate for *PA*. GR_{PA} expresses in full the belief in the soundness of *PA* and it is, according to Shapiro (1998) and Ketland (1999), an example of a basic, correct generalization that ought to be provable in any adequate theory of truth, included the ones endorsed by the deflationist. *CT* proves GR_{PA} by a formalization of Tarski's semantic consistency proof, capturing the crucial steps of our metatheoretic approach to the acceptance of B: the induction schema of *PA* extended to the truth predicate, together with T1–T4, enables one to prove that all

5 This fact was used against deflationism by Gupta (1993) and Shapiro (1998).

6 This is not to say, however, that a theory of truth that counts as compositional has to display axioms in the style of the truth axioms of *CT*. There may be ways to recover compositionality from other principles, such as reflection principles in the case of typed truth (this was pointed out to the author by Graham Leigh), or disquotation itself in the case of type-free truth. Halbach's theory *PUTB*, described for instance in Halbach (2011, Chap. 19), can in fact (partially) recover compositional axioms of Feferman's axiomatization of Kripke's theory of truth *KF*.

axioms of PA are true, that all rules of inference preserve truth, and to conclude that all theorems of PA are true.[7]

Neil Tennant questioned *Reflective Adequacy* by arguing that the full strength of CT is not necessary to express our belief in the soundness of PA (Tennant 2002); by adding to PA the local reflection schema

$$Bew_{PA}(\ulcorner \sigma \urcorner) \rightarrow \sigma$$

for all sentences σ of $\mathcal{L}_{PA}$, we obtain an extension of PA that is capable of capturing our trust in the proof machinery of PA.

As it is pointed out by Cieśliński (2010), this dispensability of the truth predicate does not give justice to the role assigned by deflationism to the truth predicate: it is a logico-linguistic device to express and *use* general claims capturing 'basic, sound' infinite sets of sentences. One would certainly accept $Bew_{PA}(\ulcorner \sigma_i \urcorner) \rightarrow \sigma_i$ for all $i \in \omega$. What the deflationist claims is that, in the presence of a truth predicate, we can fully capture or 'express' this infinite set of sentences by means GR_{PA}.[8] The central question thus becomes how to make sense of the way in which a deflationist might 'express' a (basic, sound) general claim. A straightforward answer is to equate 'expressing' with 'proving': this is the way chosen by Cieśliński.

> What is the point of having generalizations (A/N: such as GR_{PA}) if we do not have the slightest idea of how to arrive at them and how to use them in proofs? (Cieśliński 2010, p. 418)

The translation of 'expressing general claims' with 'proving their formalizations in a suitable theory of truth' may look too restrictive; after decades of reflection, however, it is still not clear how a more refined rendering could look like.[9] For the sake of Sect. 2 I will follow Cieśliński's 'provability' understanding of the way in which the 'expressive indispensability' of the truth predicate in capturing general claims should be understood. In Sect. 3.2 I will introduce a more liberal variant of Cieśliński's approach.

7 A proof can be found in Halbach (2011, pp. 104–105).

8 The standard reference for this form of truth theoretic deflationism is Horwich (1998, p. 32).

9 Volker Halbach has suggested, for instance, to understand 'expressing' as 'replacing an infinite list of sentences with a single one', a sort of finite reaxiomatization. For instance, PA+ all instances of Rfn_{PA} proves the same arithmetical sentences as $TB + GR_{PA}$. Halbach's example seems to suggest that, given a background (deflationary acceptable) axiomatization of the truth predicate, the single sentence σ capturing the infinite set S, when added to the base theory, should prove the same sentences in the base language as the base theory plus S (Halbach 1999). Richard Heck has shown, on the other hand, that if instead of just infinite conjunctions one considers both infinite conjunctions and disjunctions, then the equivalence breaks down (Heck 2004). Furthermore, although this strong interpretation may not be completely satisfactory, it may be also observed that it is assumed in the discussion of less controversial cases. It seems in fact that one of the reasons behind the rejection of a purely disquotational truth predicate lies in the deductive weakness of theories in the style of TB. This point was clearly made by Tarski (1956a). More recently, the monographs by Halbach and Horsten also highlight this point (Halbach 2011; Horsten 2012).

1.3 The conservativeness argument and Field's strategy

The doctrine, ascribable to Horwich (1998), that truth plays no explanatory role in philosophical and scientific argumentation has been recast as the requirement, for the theory of truth, to be a *conservative extension* of the base theory: U is a *conservative extension* of V with respect to a class of formulas Γ of $\mathcal{L}_V$ if U is an extension of V and, moreover, any formula φ in Γ provable in U is also provable in V.

The so-called conservativeness argument links a deflationary conception of the truth predicate with the conservativeness of an adequate theory of truth that axiomatizes it over a suitable base theory B: in a nutshell, if B is extended with axioms 'essential to truth',[10] the resulting theory of truth B' has to be conservative over B, to be compatible with deflationism. If not, the axioms for truth would be responsible for nontrivial consequences not involving truth; but a notion that displays such an evident explanatory power would hardly be considered 'thin' or 'insubstantial'. Since any adequate theory of truth is nonconservative over the base theory, deflationism is simply untenable. Shapiro (1998) and Ketland (1999) represent the standard reference for the application of the conservativeness requirement to deflationism, although already Horsten (1995) formulated some remarks in this direction. Criticisms of this argument were put forward by Field (1999), Tennant (2002) and Cieśliński (2010). I will not treat Tennant's and Cieśliński's reactions in detail; my proposal attempts to shift the discussion on other grounds.

Shapiro and Ketland's challenge to deflationism displays the following structure:

To be deflationary acceptable, a theory of truth must be conservative over the object theory (P1)

Any satisfactory theory of truth fulfils *Reflective Adequacy* w.r.t. the object theory (P2)

Reflective Adequacy entails the nonconservativeness of the theory of truth (P3)

There are no adequate, deflationary theories of truth. (C)

(P1) and (P2) are normative in character, whereas (P3) is a consequence of (P2) depending on the choice of a suitable object theory: already with the axioms of *TB* available, GR_{PA} entails $\neg Bew_{PA}(\ulcorner 0 = 1 \urcorner)$, that is, a canonical consistency statement for *PA*.

My previous remarks should now suggest how to properly understand (P1) and (P2). There are two main adequacy criteria that add up to conservativeness: compositionality and the fulfilment of *Reflective Adequacy*. *CT*, according to the proponents of the conservativeness argument, is an adequate, compositional axiomatization of the truth predicate. If conservativeness is the sort of reduction that the deflationist seeks and she accepts (P1), *CT* cannot be a viable option.

10 Cfr. Shapiro (1998, p. 497).

Field (1999) has detected, in the argument just sketched, a shift in the understanding of the expression 'theory of truth' when moving from (P1) to (P2). Field's point is that the expression 'theory of truth' is employed in an equivocal way in the argument. If one focuses on *CT*, Field continues, she has to realise that it is not constructed by adding to *PA* purely truth theoretic axioms only, as Shapiro and Ketland require. The presence of the induction schema of *PA* extended to the truth predicate is essential in two crucial steps of the proof: to prove that all instances of the non-extended schema of induction are true, and to perform the overall induction on the length of the derivation of *PA* that leads to the explicit assertion of soundness GR_{PA}. By extending the induction schema to contain semantic vocabulary, according to Field, we are enhancing the mathematical power of our theory of truth: the extended induction is not truth-theoretic in character, but it relies on the well-foundedness of the structure of natural numbers and it is a deeply mathematical principle (cf. Field 1999, p. 536). The deflationist who endorses (P1), according to Field's view, is not led to conclude that deflationism is untenable as the entire argument (P1)–(C) does not apply to the intended case of *CT*. Moreover, there are conservative theories of truth that satisfy some weaker desiderata, such as proving the T-biconditionals and some important general claims (cf. footnote 9), or displaying a finite set of compositional axioms. The theory $CT\restriction$ in particular, that is *CT* with arithmetical induction only, seems to be Field's best candidate. It is in fact well-know that $CT\restriction$ is a conservative extension of *PA*, although not every model of *PA* can be expanded to a model of $CT\restriction$.[11]

Field's strategy does not seem to be completely convincing: *Reflective Adequacy* still seems to be a desirable feature for a theory of truth. The claim that all theorems of *PA* are true, in fact, once *PA* is accepted as base theory,[12] is a correct general claim not expressible in *PA*: now if expressing a general claim is understood in our strong sense of provability, then an adequate theory of truth should have GR_{PA} among its consequences. Perhaps *Reflective Adequacy* might be weakened or shaped in a different way (cf., for instance, Sect. 3.2), but Field does not advance any proposal in this direction. If $CT\restriction$ is Field's best candidate, it is a fact that it cannot fulfil *Reflective Adequacy* as it stands, due to its ($\mathcal{L}_{PA}$-)conservativeness over *PA*.

Field's diagnosis only questioned the possibility of considering the extended induction schema, which led *CT* to satisfy *Reflective Adequacy*, an *essentially truth theoretic* principle, and I shall show shortly how he was right in this respect. The importance of satisfying *Reflective Adequacy* by an adequate theory of truth, however, is left untouched by his analysis. This led truth theorists to definitive conclusions:

> If the truth theory is also required to *prove* [...] generalizations, then conservativeness over the base theory fades away as well. The truth theory then allows for

[11] In Kotlarski et al. (1981) it was shown in fact that a nonstandard countable model of *PA* can be expanded to a model of $CT\restriction$ if and only if it is recursively saturated. The claim that not every model of *PA* can be expanded to a model of $CT\restriction$ thus follows from the observation that there are models of *PA* that are not recursively saturated. The conservativeness of $CT\restriction$ follows from the fact that every consistent first-order theory has a recursively saturated model. A new, fascinating proof of the conservativeness of $CT\restriction$ is contained in Enayat and Visser (2013).

[12] I.e. once we accept the truth of its axioms: cfr. Sect. 2.4.

> substantial conclusions that are usually established by invoking arithmetically definable sets. The deflationist must therefore accept that his truth theory has mathematically 'substantial' consequences and it is by no means conservative. (Halbach 2001a, p. 189)

In Sect. 2 I present a refinement of Field's strategy. There are in fact theories of truth satisfying *Reflective Adequacy* and displaying a truth predicate with deflationary traits that are not affected by the conservativeness argument in the usual or modified forms.

1.4 Relative interpretability

A relative interpretation of a theory U in a theory V, in nuce, is a translation of $\mathcal{L}_U$ in $\mathcal{L}_V$ that preserves provability and the logical structure of formulas.[13] Some authors have argued that a theory of truth that is conservative but not relatively interpretable in the object theory is particularly attractive for the deflationist (Fischer 2010; Horsten 2012):

> The notion of relative interpretability provides a way to articulate the double demand that truth should be conservative yet not reducible. Maybe we should look for a truth theory that is conservative but not interpretable into its arithmetical background theory. This would give rise to a new version of deflationism about truth. This form of deflationism would hold that conservativeness is an adequacy condition for theories of truth. But it would claim at the same time that truth is not reducible to the background theory. *Truth is a notion that we did not implicitly have before it was introduced; however, this has been done. Its inferential behavior cannot be simulated by our background theory.*[14] (Horsten 2012, pp. 94–95)

In this work, I will also provide examples of theories of truth that are conservative but non interpretable in their background mathematical theories: to be precise, I will not only provide some examples of theories with such properties, but a uniform and widely applicable method to construct them. This will become clear in Sect. 3.2.

[13] More precisely, let U and V be theories in predicate logic with identity and in a relational language. A *relative translation* of $\mathcal{L}_U$ into $\mathcal{L}_V$ can be described as a pair (δ, F) where δ is a $\mathcal{L}_V$-formula with one free variable—the domain of the translation—and F is a (finite) mapping that takes n-ary relation symbols of $\mathcal{L}_U$ and gives back formulas of $\mathcal{L}_V$ with n free variables. The translation extends to the mapping ρ:

- $(R(x_1, \ldots, x_n))^\rho :\leftrightarrow F(R)(x_1, \ldots, x_n)$;
- ρ commutes with propositional connectives;
- $(\forall x \varphi(x))^\rho :\leftrightarrow \forall x(\delta(x) \rightarrow \varphi^\rho)$ and $(\exists x \varphi(x))^\rho :\leftrightarrow \exists x(\delta(x) \wedge \varphi^\rho)$.

An *interpretation* K is then specified by a triple (U, ρ, V) such that for all sentences σ of $\mathcal{L}_{\mathsf{U}}$,

$$U \vdash \sigma \Rightarrow V \vdash \sigma^\rho.$$

We notice that this definition entails $V \vdash \exists x\, \delta(x)$.

[14] My emphasis.

As Horsten's passage highlights, a theory of truth that is not relatively interpretable in the base syntactic or mathematical theory displays a notion of truth that, although 'adequate' in the light of the conservativeness requirement, is conceptually not reducible to the syntactic or mathematical resources of the object theory: the notion of truth, according to this understanding, has no explanatory power but it is expressively irreducible.

2 Objects of truth versus object theory

I deliberately restricted my attention to mathematical base theories: a deflationary theory of truth, says one of the premises of the conservativeness argument, ought to be conservative over the underlying mathematics. Some authors argued that if the conservativeness requirement (P1) for deflationary truth theories has to be taken seriously, one should focus on their conservativeness over the underlying ontology of expressions and not over the ontology of mathematical objects. Among them, Volker Halbach in a previous issue of this journal (Halbach 2001a). In Shapiro's own words:

> So, at best, the proper conclusion of [A/N the conservativeness argument] is that a deflationist should hold that truth is conservative over the underlying syntax. (Shapiro 2002, p. 110)

In the customary setting, however, this is no different from requiring the theory of truth to be conservative over the underlying mathematics.[15] Let us consider the familiar case of *CT*: according to the received view, *CT* has two main groups of axioms, the axioms of *PA* (including the its logical axioms) and the truth axioms — regardless of where we place the extended induction schema. If confronted with the original Tarskian picture of the metatheory that *CT* tries to capture, we see that something is missing: Tarski distinguishes between an object theory, *PA* in this case, and a theory of expressions that formalizes syntactic operations and notions concerning the object theory (Cf. Tarski 1956a, §2). In modern presentations, we employ the fact that the intended objects to which truth is ascribed — sentence types in this case — possess an ontological structure that is similar to that of natural numbers: it is known that the structure of natural numbers is isomorphic to the structure of strings of a finite or countable alphabet (Corcoran et al. 1974).

As the example of *CT* emphasizes, in the usual construction natural numbers play the role of the objects to which truth is ascribed: *PA* contains a theory of syntax. Hilbert's program, Gödel phenomena and contemporary logic are rooted in the possibility of coding syntactic objects as numbers or sets. For the purpose of our discussion, however, the inclusion of the domain of the objects of truth in the domain of discourse

[15] For instance:

> Therefore, we might try to defend the conservativeness of his theory of truth over a base thery of sentence types rather than logic. But what is a suitable theory of sentence types? There are good reasons for picking Peano arithmetic, although other theories might be suitable as well. Because the discussion so far has mainly focused on Peano arithmetic (which has some nice features as a base theory), I will concentrate on it; similar points, however, can be made for several other theories (like *ZF*). (Halbach 2001a, p. 182)

of the base theory appears too stringent; what if we could distinguish between a notion of conservativeness (of the theory of truth) over the underlying mathematical object theory and a notion of conservativeness (of the theory of truth) over the theory of sentence types that formalizes the syntax of the object theory? To accomplish this task an unconventional approach to the construction of axiomatic truth theories, already proposed and investigated in Heck (2015), in Halbach (2011), and studied in technical aspects in Leigh and Nicolai (2013) and Nicolai (2015), comes in handy.

2.1 Overview of the theories

I start with a first-order, consistent object theory B in a relational language. As additional technical assumptions, I stipulate that B comes with a Δ_1^b-definition (i.e. a p-time notion). For a description of the complexity classes Δ_i^b, we refer to Hájek and Pudlák (1993). In Sect. 3.2, but only there, it is required B to interpret a very weak arithmetical theory such as Q. Again the only logical constants of $\mathcal{L}_B$ are $\neg, \wedge, \forall$. The idea is to enlarge B to a many-sorted theory that is capable of talking about 'syntactic' aspects of B and to attribute truth to these syntactic entities. The crucial aspect of our construction lies in the fact that syntactic objects are not a subset of the domain of discourse of B, but they live in a disjoint domain: *in a sense, they are built-in the truth axioms*. Depending on which axiomatization of the truth predicate is employed, I will tweak the resources needed to formulate the theory of truth.

The intuitive picture of the construction is as follows: on the one hand we have an object-theoretic universe governed by the axioms of B relativized to a suitable sort. These may be numbers, sets, physical or even linguistic objects. On the other, we have a universe of objects that are exclusively employed to formalize the syntax of B. These are the bearers of truth ascriptions. The two universes are disjoint. Only when compositional truth axioms are introduced — as will be carefully explained shortly — a third universe of 'mixed' objects is needed.

For the 'syntactic' theory that will be paired with our typed axiomatizations of the truth predicate, I opt for an easy choice and assume that a copy of Q is available inside the theory of truth and suitably relativized to the 'syntactic' sort. This gives me the means to develop the syntax of B in an intensionally correct way.[16]

The first system that I will consider is the equivalent of *TB* in the new setting. To make disquotation axioms meaningful, we only need to consider the two-sorted language $\mathcal{L}_D$: variables of the first sort, $x, y, z, \ldots$ range over elements in the domain of discourse of object-theoretic quantifiers; variables of the second, denoted by $i, j, k, l, \ldots$ (and often by expressions of the form $\ulcorner e \urcorner$, where e ranges over $\mathcal{L}_B$-expressions), range over objects in the domain of discourse of syntactic quantifiers. Nonlogical constants of $\mathcal{L}_D$ will be the nonlogical constants of $\mathcal{L}_B$ and the language of arithmetic relativized to the appropriate sorts and types, s for 'syntactic', o for 'object-

[16] Other choices are possible, and perhaps more motivated: in Nicolai (2014) it is considered a theory of finite sets, which naturally captures the informal development of the syntax of B in terms of finite trees. We recall that, although Q is extremely weak, there is an interpretation of Buss' theory S_2^1 in it on a definable cut. S_2^1 is sufficiently strong and extremely efficient for an intensional formalization of the syntax of B. This is the preferred choice in Nicolai (2015).

theoretic'. The truth predicate T will thus be a unary predicate of type s. Gödel corners acquire now a new meaning compared to the one they had in the definition of *TB* above. If σ is a formula of $\mathcal{L}_B$, $\ulcorner \sigma \urcorner$ is now a term of the language of *TBD*[*B*] whose nonlogical constants and variables belong to the 'syntactic' part of the language; in other words, no nonlogical constants and variables of $\mathcal{L}_B$ are present in $\ulcorner \sigma \urcorner$. By keeping σ, τ, χ to range over sentences of $\mathcal{L}_B$, one defines *TBD[B]* as the theory in $\mathcal{L}_D$ featuring (i) axioms of B (reativized to the sort o); (ii) axioms of Q (relativized to the sort s); (iii) the schema

$$T\ulcorner \sigma \urcorner \leftrightarrow \sigma \qquad \text{TBD}$$

for all $\mathcal{L}_B$-sentences σ.

This simple setting also enables one to introduce compositionality, although not full compositionality: the system *CD*[*B*] is defined by adding to *TBD*[*B*] the axioms:

$$\forall \ulcorner \sigma \urcorner (T\ulcorner \neg\sigma \urcorner \leftrightarrow \neg T\ulcorner \sigma \urcorner) \qquad \text{CD}\neg$$

$$\forall \ulcorner \sigma \urcorner, \ulcorner \tau \urcorner (T\ulcorner \sigma \wedge \tau \urcorner \leftrightarrow (T\ulcorner \sigma \urcorner \wedge T\ulcorner \tau \urcorner)) \qquad \text{CD}\wedge$$

There are obvious reasons, however, to consider more complex cases: one might want in fact to introduce parameters into the T-schema and to further analyze atomic and universally quantified statements. In this particular setting, truth is not sufficient anymore: I need to consider a binary satisfaction predicate *Sat* and introduce a further domain of variable assignments to be able to unravel the structure of atomic and quantified formulas. I thus enlarge $\mathcal{L}_D$ to $\mathcal{L}_D^+$ by adding to it new variables $a, b, c \ldots$ ranging over a further domain of 'mixed' objects, syntactic and object-theoretic at the same time. We assume only that (i) the domain of variable assignments is nonempty and (ii) that, given an assignment a, we can construct an assignment b such that a and b agree on what they assign on all variables except one. We call this theory *Seq*, without specifying its axioms.

The theory *UTBD*[*B*] of uniform diquotation in $\mathcal{L}_D^+$ will thus contain, besides the axioms of B, Q and *Seq* relativized to the appropriate sorts, the schema

$$\forall a \forall \ulcorner \vec{v_j} \urcorner (Sat(a, \ulcorner \varphi \urcorner (\vec{v_j}/\vec{v_i})) \leftrightarrow \varphi(\overrightarrow{a(\ulcorner v_j \urcorner)})) \qquad \text{(UTBD)}$$

for all $\mathcal{L}_B$-formulas $\varphi(v_i)$, where the vector notation informally indicates sequences of terms of the three-sorted language $\mathcal{L}_D^+$, and the syntactic term $\ulcorner \varphi \urcorner (\vec{v_j}/\vec{v_i})$ stands for the result of *formally* substituting the 'object-theoretic' variables $\vec{v_j}$ for the free variables $\vec{v_i}$ of the same sort in the $\mathcal{L}_B$-formula φ. The functional expression $a(\ulcorner v_j \urcorner)$ yields, intuitively, the object in the assignment a corresponding to the variable v_j. Once again, we notice that $\ulcorner v_j \urcorner$ is a term of the language $\mathcal{L}_D^+$ whose value lives in the universe of the bearers of truth, that v_j is an object-theoretic variable, and that the value of $a(\ulcorner v_j \urcorner)$ lives in the object-theoretic domain.

The fully compositional theory *CTD*[*B*] is obtained by reformulating *CD*[*B*] with the new satisfaction predicate and by adding the axioms for atomic and quantified formulas to it:

$$\forall a \forall \ulcorner v_i \urcorner (Sat(a, \ulcorner R\vec{v_i} \urcorner) \leftrightarrow R(\overrightarrow{a(\ulcorner v_i \urcorner)})) \qquad \text{CTD}R$$

$$\forall a \forall \ulcorner v_i \urcorner \forall \ulcorner \varphi \urcorner (Sat(a, \ulcorner \forall v_i \varphi \urcorner) \leftrightarrow \forall \ulcorner v_j \urcorner (\ulcorner v_j \urcorner \neq \ulcorner v_i \urcorner \to a(\ulcorner v_j \urcorner) = b(\ulcorner v_j \urcorner)) \to Sat(b, \ulcorner \varphi \urcorner)) \qquad \text{CTD}\forall$$

CTD∀ simply captures the standard, Tarskian clause for universally quantified formulas: $\forall v_i \varphi$ is satisfied by a sequence a if and only if it is satisfied by all sequences that differ from a *at most* in what they assign to the variable v_i.

For the present discussion, a special role is played by the theory $CTD^+[B]$, obtained by adding to $CTD[B]$ the induction schema

$$A(0^s) \wedge \forall k(A(k) \to A(S^s k)) \to \forall k A(k) \qquad \text{SIN}$$

where 0 and S are terms of 'syntactic type' and A is an arbitrary formula of $\mathcal{L}_D^+$. SIN is a syntactic induction principle: it enables one to perform metatheoretic reasoning on the object theory B regardless of the deductive power of the base theory. The role of SIN can be clarified via an example: if B is *PA*, it will feature an induction schema of the form

$$\varphi(0^o) \wedge \forall x(\varphi(x) \to \varphi(S^o x)) \to \forall x\, \varphi(x) \qquad \text{MIN}$$

where the notation 0^o and S^o suggest that the value of 0^o lives in the object theoretic domain as well as the values of the input and the output of S^o. The schematic variable φ in MIN only varies on the object-theoretic portion of the formulas of the language. No syntactic or semantic components of the language can appear in φ. SIN and MIN are completely different principles in $CTD^+[PA]$. As an anonymous referee has pointed out, moreover, the existence of a form of 'metamathematical' induction along the lines of SIN, as opposed to formal, object theoretic induction principles or schemata, such as MIN, was advocated by Hilbert in responding to Poincarè's attacks to his program. For more information on this point, we refer to Sieg (1999).

In the next section I outline consequences and properties of the theories just considered that are particularly relevant for the discussion at issue. Before this, however, I pause for a moment to briefly elaborate on the structure of the truth theories just introduced. Theories in the style of $CTD[B]$ (and $CTD^+[B]$) are *essentially typed* and the author has not found any plausible way extended this idea to a type-free setting: this is simply because the strict separation between sorts rules out self-applications of the satisfaction predicate. Can this be a threat for the entire project? As observed by McGee (2006), if the truth predicate is *just* a logico-linguistic device, then the fact that some of its applications are blocked or unavailable due to paradoxes is not terribly problematic, more or less like the fact that restrictions to the applicability of $\in$ does not represent a significant blow to the development of set theory and to the understanding of the notion of membership. Although type-free truth theories are more natural theories of truth, the conservativeness argument, and all the subsequent literature on this matter, only focus on typed truth. The reader might thus go on also without agreeing with McGee's defence of typed truth in this particular discussion.

2.2 Metatheoretic theorems and properties: a survey

If the object theory B is finitely axiomatized, $CTD^+[B]$ turns out to be adequate in the sense of *Reflective Adequacy*. By induction on the length of the derivations in B in fact, that is via a straightforward application of SIN, we can prove in $CTD^+[B]$ the $\mathcal{L}_B$-sentence

$$\forall a \forall k (Fml_{\mathcal{L}_B}(k) \wedge Bew_B(k) \rightarrow\ Sat(a, k)), \qquad GR_B$$

that is the statement 'all theorems of B are true' formulated in the new language. The 'syntactic' induction principle SIN, extended to the new vocabulary, suffices to prove the claim: since $CTD^+[B]$ proves the Tarski-biconditional for B, and B is finitely axiomatized, we obtain the truth of all axioms of B.[17] *CTDR*-*CTD*$\forall$, together with a clever choice of a proof system for B, tell us that if a proof of length n has a true conclusion, so has a proof of length $n+1$. SIN thus enables us to conclude GR_B. Standardly, by instantiating an absurdity ξ in the sense of B into GR_B, we obtain Con_B,[18] that is the *syntactic* assertion of the consistency of B.

Field's emphasis on the mathematical character of the induction axioms of CT can thus be refined by the new setting in which syntactic-truth theoretic reasoning is distinguishable from the reasoning carried out in the object theory. He was right in diagnosing a 'mixed' character of the induction axioms of CT: the usual construction, however, is an unsuitable setting for realizing the importance played by the theory of syntax in the argument. *All that we need to obtain GR_B and, consequently, Con_B, is an induction principle applying to syntactic objects, to the bearers of truth.*

If B is schematically or infinitely axiomatized, $CTD^+[B]$ does not have the resources to prove Con_B and not even the claim that all axioms of B are true.[19] One has to resort to the claim

$$\forall a \forall k (Ax_B(k) \rightarrow\ Sat\ (a, k)) \qquad AxT_B$$

expressing that all axioms of B are true. If instead of the theory $CTD^+[B]$, one focuses on the theory $CTD^+[B]+AxT_B$, she will immediately obtain a theory which is adequate in the sense of *Reflective Adequacy*. Once we assume the base case of our induction, in fact, we can again combine compositional axioms and syntactic induction SIN to obtain GR_B. For a discussion of the role of AxT_B, we refer to Sect. 2.4.

[17] In particular, by a straightforward metatheoretic induction on the complexity of an arbitrary formula $\varphi(v_i)$, we prove

$$\forall \ulcorner v_j \urcorner \forall a (Sat(a, \ulcorner \varphi \urcorner [v_j / v_i]) \leftrightarrow \varphi(a(\ulcorner v_j \urcorner))$$

for all formulas $\varphi(v_i)$ of the base language.

[18] The argument is as follows. We reason in CTD^+[B]: assuming, $Bew_B(\ulcorner \xi \urcorner)$, we get $Sat(a, \ulcorner \xi \urcorner)$ and thus ξ by the Tarski-biconditionals. But also $\neg\xi$. So $\neg Bew_B(\ulcorner \xi \urcorner)$.

[19] It is not possible to be absolutely general here: but from Leigh and Nicolai (2013), Nicolai (2015) one can extract the impossibility results in the cases in which B is PA or ZF, and thus for many of their extensions.

$CTD^+[B]$ and $CTD^+[B] + AxT_B$ prove the global reflection for B and also a consistency statement for it constructed in a natural way. They thus satisfy *Reflective Adequacy* and, according to (P3), this fact should entail the nonconservativeness of $CTD^+[B]$ and $CTD^+[B] + AxT_B$ over B. When it comes to $CTD^+[B]$, however, (P3) does not hold any more. In fact any model of B can be expanded to a model of $CTD^+[B]$ and $CTD^+[B] + AxT_B$, and thus $CTD^+[B]$ and $CTD^+[B] + AxT_B$ are conservative extensions of B.[20] Since syntactic and object-theoretic domains are disjoint, in fact, the 'special' satisfaction classes determined by the satisfaction predicates of $CTD^+[B]$ and $CTD^+[B]+AxT_B$ do not depend on the structure of the objects living in the 'object-theoretic' universe. Therefore from an arbitrary model of B we can always construct models of $CTD^+[B]$ and $CTD^+[B] + AxT_B$ that contain interpretations of the syntactic vocabulary over our disjoint domain of quantifiers of the syntax theory and of entities encoding sequences of objects in the sense of the domain of discourse of B. To understand the idea of proof, it is sufficient for the reader to realize that in models of $CTD^+[B]$ and $CTD^+[B] + AxT_B$, syntactic objects and sequences of nonstandard length can live together with a standard model of B and vice versa.[21] Yet from another point of view, the model expansion argument just sketched may be seen as a formalization of the informal metatheoretic reasoning on the object theory: in informal metamathematics, in fact, nonstandard objects are simply not present.

The conservativeness of the axiomatizations of the truth predicate with 'built-in' syntax over the object theory seems to be quite appealing for the deflationist. In the first place model-theoretic conservativeness is a stronger notion than proof-theoretic conservativeness. Some authors (McGee 2006, for instance) have in fact argued that this model expansion property is to be preferred over proof-theoretic conservativeness, as the theory of truth does not restrict the class of models of the object theories. Moreover, $CTD^+[B]$ and $CTD^+[B] + AxT_B$ satisfy all the adequacy requirements considered by Shapiro and Ketland,[22] *Reflective Adequacy* included. This fact appears to be particularly in line with Field's understanding of the role of the truth predicate:

> The main point of having the notion of truth, many deflationists say, is that it allows us to make fertile generalizations we could not otherwise make; where by a fertile generalization I mean one that has an impact on claims not involving the notion of truth. (Field 1999, p. 533)

We have already stressed how strong is the sense in which we are formalizing Field's 'making' of fertile generalizations. But even in our strong sense, the theories just introduced are able to obtain the general claims that the inflationist requires, and the syntactic consistency statement Con_B can be regarded as an example of a claim not involving the notion of truth on which GR_B has an impact on.

The thesis that truth does not play a substantial role in solving philosophical disputes refers mainly to Horwich's discussion of the role of the notion of truth in the theory of meaning. Contrary to Davidsonians who assign an epistemological priority to the

[20] The same argument straightforwardly applies to *TBD[B]* and *UTBD[B]*.

[21] For the details Leigh and Nicolai (2013, Theorem 3.11). The main strategy behind the conservativeness proof was first noticed by Volker Halbach in 2010.

[22] The same holds for $CTD[B] + AxT_B$.

theory of truth — indeed, Tarski's — in grasping the meaning of natural language expressions, Horwich claims that a theory of truth cannot explain *all* attributions of meaning, and thus it is ultimately not needed.[23] The debate on the conservativeness argument sketched above, as we have seen, was based on the assumption that this alleged insubstantiality of the truth predicate could be explained via the logical notion of conservativeness of the theory of truth over the object theory. Shapiro and Ketland suggest, with their argument, that truth does indeed play a substantial role in explanations, as by having a notion of truth one is led to learn more about one's target subject matter, natural numbers in the case of CT.[24]

The case of the theories introduced suggests that the notion of truth, as depicted by the axiomatizations considered, provides syntactic or metatheoretic explanations, but these should not be considered 'substantial', at least in the usual sense attributed to the term: *by invoking a theory of truth for the object theory B, we at best learn more about the syntactic universe belonging to B's metatheory, but not about its subject matter*. The Gödelian phenomena at the root of the conservativeness argument show that, when axioms for truth are added to B in the usual way, these syntactic consequences might be pushed back in B via the known isomorphism between numbers and expressions. In many cases, this relies on accidental features of B: only when B features induction schemata extendable to semantic vocabulary, this 'pushing back' is possible and full metamathematical reasoning is replicable in the theory of truth.[25]

In the next subsections I will consider two possible objections to the acceptance of $CTD^+[B]$ and variants thereof.

2.3 First rejoinder: conservativeness over the syntax

Once we consider axiomatizations of the truth predicate with built-in syntax, it is possible to rephrase the conservativeness argument in a refined environment. In Leigh and Nicolai (2013) the following strategy was outlined:[26] let B again be a consistent, object theory in predicate logic formulated in a relational language and let us expand this language with a new sort of variables and quantifiers ranging over the domain of natural numbers that, as before, play the role of our 'disjoint' syntactic universe: we call the resulting language $\mathcal{L}_{B^*}$. We shall confine ourselves to the case in which B is finitely axiomatized, although what we shall say naturally applies to $CTD^+[B]+AxT_B$ as well.

The theory B^* contains thus, besides the axioms of B, the basic axioms of PA and an induction principle open to all formulas of $\mathcal{L}_{B^*}$ but applicable — exactly as SIN — to variables of syntactic sort only. As we have seen above, $CTD^+[B]$ proves Con_B.[27] B^*, on the contrary, cannot prove Con_B. For a natural choice of B and of the

[23] Cf. Horwich (1998, pp. 68–70).

[24] Cf. Shapiro (1998, p. 499).

[25] In Leigh and Nicolai (2013, §1.4) the role of induction schemata in this context is explained in full detail.

[26] This new version of the argument was also suggested to the author by Richard Heck and Jeffrey Ketland in private communication.

[27] $CTD^+[B] + AxT_B$, if B is schematically axiomatized.

formalization of the syntax of B in B^*, in fact, we can interpret B^* in B in such a way that, if Con_B were provable in B^*, then a natural consistency statement Con_B^*, this time formulated in the language $\mathcal{L}_B$, would be provable in B, contradicting Gödel's Second Incompleteness Theorem.[28] Now if one focuses over B^* as new base theory, she will realise that $CTD^+[B]$ is now a nonconservative extension of B^* as it proves, by the argument sketched above, Con_B. Therefore also $CTD^+[B]$ would not count as a theory that is acceptable for the deflationist, as although it is a conservative extension of the base theory B, it is not conservative over the syntactic/object-theoretic bundle theory B^*.

Despite its apparent attractiveness, this version of the conservativeness argument is misguided. To elaborate on this point, it is sufficient to look at how $CTD^+[B]$ is constructed. According to the argument sketched above the conservativeness of $CTD^+[B]$ over B is not the right notion of conservativeness we should be interested in: B^* is the right object theory to focus on. As a simple reaction, we notice that $CTD^+[B]$ is *not* a theory of truth for B^*, neither in the usual nor in the sense of the disentanglement of the syntactic objects from the mathematical domain. For the truth axioms of $CTD^+[B]$ only stipulate the conditions under which a (code of a) formula of $\mathcal{L}_B$ belongs to the satisfaction class forced by the theory of truth, but they do not tell us anything about the truth of sentences of $\mathcal{L}_{B^*}$. *A fortiori*, $CTD^+[B]$ does not satisfy adequacy requirements for theories of truth with respect to B^* such as compositionality or *Reflective Adequacy*.

A typed theory of truth for B^* can of course be easily constructed, but it would be radically different from theories in the style of CT or $CTD^+[B]$; in the 'entangled' case, since B^* has PA as its subtheory, we can formalize the syntax of the two-sorted language $\mathcal{L}_{B^*}$ in PA in the usual way, expand $\mathcal{L}_{B^*}$ with a unary truth predicate governed by axioms that stipulate its applicability to sentences of the entire language $\mathcal{L}_{B^*}$. An axiomatization of the truth predicate with 'built-in' syntax for B^*, on the other hand, will be constructed exactly as before, so a disjoint universe of codes of expressions of the language $\mathcal{L}_{B^*}$ will be provided together with a suitable theory of variable assignments for *both* syntactic and mathematical variables, and again the axioms for the satisfaction predicate will give the conditions of its applicability to sentences of $\mathcal{L}_{B^*}$ and not just of $\mathcal{L}_B$. But it should be clear how different this theory is from $CTD^+[B]$.

2.4 Second rejoinder: on the truth of the axioms of the base theory

Before proceeding any further, we address another possible objection. One may in fact wonder whether the unprovability of AxT_B in $CTD^+[B]$, when B is schematically axiomatized, is not an example of a big deficiency of the theory. AxT_B might appear a correct generalization that *ought* to be provable in an adequate theory of truth. In

[28] The details of the interpretation are straightforward as the syntactic part of B^* is essentially 'collapsed' in B. To carry out the translation of the induction axioms in the mixed language a sufficient amount of induction is required to be available in B. However, it seems harmless for the present argument to assume that B contains Σ_1-induction.

this respect $CTD^+[PA]$, for instance, would not be much different from $CT\restriction$.[29] This objection can be circumvented, though. The provability of AxT_B in the theory of truth, when B is schematically axiomatized, can only rely on an interaction of the object theoretic and syntactic universes.[30] As the case of CT clearly indicates, only when schemata of the object theory are extended to contain semantic vocabulary, one can obtain the truth of all their instances. But this is a quite accidental character of the theory: if one considers a theory which is infinitely axiomatized without being schematically axiomatized, the provability of the truth of all its axioms is unreachable — in an entangled or in a disentangled setting. The postulation of AxT_B should be considered as a reasonable assumption to add to our theory of truth, at least when one is considering — as we are — consistent base theories and a framework such as the one behind $CTD^+[B]$ whose rigidity is exchanged for new insights on the different conceptual areas encapsulated in usual truth theoretic constructions.

Moreover, even if one does not accept AxT_B as reasonable part of the theory of truth, we have already noticed that in the case of a finitely axiomatized base theory, the claim becomes provable from the compositional axioms alone. If B is schematically or infinitely axiomatized, one might consider a finite reaxiomatization of it using well-known tricks.[31] This is also justified by the fact that B does not play any role in the truth theoretic reasoning carried out on it, therefore focusing on an infinitely axiomatized B or on a finite reaxiomatization of it B^{fin} amounts to the same, from the perspective of the theory of truth. For this reason from now on I will mostly focus on the case of finitely axiomatized base theories B and confine extensions to different choices of B to a concluding remark.

3 How (compositional) truth plays a role

There seem thus to be evidences in support of the thesis that truth plays a substantial role in syntactic, metatheoretic explanations concerning the object theory B, but no substantial role concerning its subject matter. But how robust is this characterization? I address this question by considering two variations of the setting considered.

In Sect. 3.1 I take into account a recent proposal of Leon Horsten, who seems to provide an example of the substantial role played by the notion of truth in epistemology. If taken at face value, Horsten's example appears to dismiss our reading of the role

29 Namely the theory, we recall, that is defined as CT with arithmetical induction only.

30 In CT, for instance, one can only prove the truth of all axioms of PA by applying compositional truth axioms to a suitable instance of the induction schema of CT:

$$\begin{aligned}&T\,sub(\ulcorner\varphi\urcorner,\ulcorner v\urcorner,\underset{\cdot}{0})\wedge\\&\forall x(T\,sub(\ulcorner\varphi\urcorner,num(x),\ulcorner v\urcorner)\rightarrow T\,sub(\ulcorner\varphi\urcorner,num(Sx),\ulcorner v\urcorner))\rightarrow\\&\forall x(T\,sub(\ulcorner\varphi\urcorner,num(x),\ulcorner v\urcorner)\end{aligned}$$

for all $\varphi(v)$ of $\mathcal{L}_{PA}$, where $num(\cdot)$ is the PA-definable function that sends each number to its numeral. Quantification in the instance above is not over syntactic objects but over all numerals.

31 We might for instance treat schematic variables as 'second-order' variables, as we proceed when we move from PA to ACA_0, or to the method devised by Craig and Vaught (1958).

of a typed, compositional truth predicate. In Sect. 3.2, I investigate to what degree this characterization of the notion of truth as syntactically substantial behaves in the absence of a crucial principle: the syntactic induction schema SIN.

3.1 Horsten's example

Leon Horsten, in Horsten (2009) and Horsten (2012), gives a modification of Fitch's well-known derivation (Cfr. Fitch 1963) of strong verificationism ('all truths are known') from weak verificationism ('all truths are knowable'). Horsten works in a theory formulated in the modal logic **K** augmented with factivity and distributivity axioms for a predicate K for knowledge. More precisely, we refer to a system — that, as in Horsten (2009), we call F — containing (i) the modal logic **K**, (ii) axioms sufficient to formalize the syntax of the language $\mathcal{L}_F$, (iii) (propositional) axioms for a truth predicate with type restrictions — i.e. applying only to sentences of $\mathcal{L} = \mathcal{L}_F - \{T\}$ — in the style of T2 and T3 in Sect. 1.1, and (iv) axioms for factivity and distributivity of the predicate for knowledge:

$$\forall x\big(Sent_{\mathcal{L}}(x) \rightarrow \Box(T\underset{\cdot}{K}x \rightarrow Tx)\big) \tag{fc}$$

$$\forall x \forall y\Big(Sent_{\mathcal{L}}(x \underset{\cdot}{\wedge} y) \rightarrow \Box\big(T\underset{\cdot}{K}(x \underset{\cdot}{\wedge} y) \rightarrow (T\underset{\cdot}{K}x \wedge T\underset{\cdot}{K}y)\big)\Big) \tag{ds}$$

In this section it is useful to employ, as already done in (fc) and (ds), Feferman's dot convention for naming syntactic notions and operations. So, for instance, $\underset{\cdot}{\neg}$ becomes a function symbol, definable in the syntax theory, that represents the operation of prepending the negation symbol to an expression.

It is then shown that, over F, weak verificationism

$$\forall x\big(Sent_{\mathcal{L}}(x) \rightarrow (Tx \rightarrow \Diamond T\underset{\cdot}{K}x)\big). \tag{wv}$$

entails strong verificationism, that is the sentence

$$\forall x\big(Sent_{\mathcal{L}}(x) \rightarrow (Tx \rightarrow T\underset{\cdot}{K}x)\big). \tag{sv}$$

Horsten interprets this result as an evidence in favour of the claim that truth plays indeed a substantial role in philosophical argumentation, epistemology in particular, contrary to Horwich's tenet.

I agree with Horsten that truth seems to play a *significant* role in the argument, but not in the sense in which Horsten appears to intend it. To support my interpretation, I consider a variant of $CD[B]$ introduced in Sect. 2.1. We let $\mathcal{L}_B$ be an arbitrary first-order language with a fresh predicate K. Moreover, we require B to be formulated in the modal logic **K**.[32] Let us now expand $\mathcal{L}_B$ in the way explained above: we add a new sort of variables, intended to range over a disjoint domain of syntactic objects. For uniformity, we employ the same conventions concerning variables as

[32] For the sake of the present argument, indeed, it can well be *just* the modal logic **K** formulated in $\mathcal{L}_B$.

above. The theory F^* is thus obtained by reformulating Horsten's axioms in the new setting.[33] Again we extend **K** with syntactic axioms — axioms of Q will suffice — suitably relativized; in addition, as we only need two axioms for truth for negation and conjunction, a unary truth predicate can be employed instead of a binary satisfaction predicate. Quantification over codes of sentences acquire thus the usual, new meaning: we quantify over codes of sentences of the language $\mathcal{L}_B$ just described provided in the new universe of expressions. The axioms of F^*, beside the axioms of **K**, are:

$$\forall\ulcorner\sigma\urcorner(\Box(T\underset{\cdot}{K}\ulcorner\sigma\urcorner \to T\ulcorner\sigma\urcorner)) \tag{fc*}$$

$$\forall\ulcorner\sigma\urcorner,\ulcorner\chi\urcorner\Big(\Box\big(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\ulcorner\chi\urcorner) \to (T\underset{\cdot}{K}\ulcorner\sigma\urcorner \wedge T\underset{\cdot}{K}\ulcorner\chi\urcorner)\big)\Big) \tag{ds*}$$

$$\forall\ulcorner\sigma\urcorner\big(T\underset{\cdot}{\neg}\ulcorner\sigma\urcorner \leftrightarrow \neg T\ulcorner\sigma\urcorner\big) \tag{F*¬}$$

$$\forall\ulcorner\sigma\urcorner\forall\ulcorner\chi\urcorner\big(T\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\ulcorner\chi\urcorner \to T\ulcorner\sigma\urcorner \wedge T\ulcorner\chi\urcorner\big) \tag{F*∧}$$

A consistency proof for F^* can be easily given by interpreting F^* in $CD[B]$ — where B can be just first-order logic in a language with the knowledge predicate K — by erasing occurrences of $\Box$ in $\mathcal{L}_{F^*}$-formulas in F^*-proofs.

Again we can show that the new version of weak verificationism

$$\forall\ulcorner\sigma\urcorner\big(T\ulcorner\sigma\urcorner \to \Diamond T\underset{\cdot}{K}\ulcorner\sigma\urcorner\big) \tag{wv*}$$

will yield the corresponding formalization of strong verificationism. Here is the proof: By an instance of (ds*), we obtain, for an arbitrary sentence σ of $\mathcal{L}_B$,

$$\Box\big(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner) \to (T\underset{\cdot}{K}\ulcorner\sigma\urcorner \wedge T\underset{\cdot}{K}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner)\big) \tag{1}$$

Moreover, by employing the factivity principle (fc*) and ($F^*\neg$), one gets

$$\Box\big(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner) \to (T\underset{\cdot}{K}\ulcorner\sigma\urcorner \wedge \neg T\underset{\cdot}{K}\ulcorner\sigma\urcorner)\big); \tag{2}$$

but the results of reasoning in propositional logic about necessary truths are themselves necessary, so by definition of $\Diamond$ we obtain

$$\neg\Diamond\big(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner)\big). \tag{3}$$

On the other hand, an instance of (wv*) gives us

$$T(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner) \to \Diamond(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner)); \tag{4}$$

by propositional logic, $F^*\neg$, $F^*\wedge$, and by definition of $\Box$,

$$\neg\Diamond(T\underset{\cdot}{K}(\ulcorner\sigma\urcorner\underset{\cdot}{\wedge}\underset{\cdot}{\neg}\underset{\cdot}{K}\ulcorner\sigma\urcorner)) \to \neg(T\ulcorner\sigma\urcorner \wedge \neg T\underset{\cdot}{K}\ulcorner\sigma\urcorner). \tag{5}$$

33 We notice that F^* is formulated in the modal logic **K**, as well as the theory $TD[B]$ was formulated in the same logic as B, that is classical logic.

That is, by (3), we have proved in F^*

$$\forall k(Sent_{\mathcal{L}_B}(k) \rightarrow (Tk \rightarrow T\dot{K}k)) \tag{6}$$

The possibility of translating so easily Horsten's argument in a setting with 'disentangled' syntax highlights the contribution that the notion of truth gives to the argument, also in the case of F. As Horsten concedes, in fact, the role of the truth predicate in the argument mainly lies in its capability of amending the quantification over sentence letters proper of Fitch's original argument:

> Our common understanding of quantification is in terms of objectual quantification. A formula of the form $\exists p: \ p$ simply appears to be ill-formed, because an object is not a candidate for having a truth value. The received view is that from the conventional objectual quantification point of view, sense can be made of propositional quantification, using a truth predicate [...]. A sentence of the form $\exists p: \ p$ is then taken to be short for a sentence of the form $\exists x : x \in \mathcal{L} \wedge Tx$. If this line is adopted, then Fitch's argument is really an argument that involves a truth predicate. (Horsten 2009, p. 565)

If compared to the case of *CT* and of Shapiro and Ketland's rendering of the notion of 'substantiality' of the truth predicate, however, the role played by the truth predicate in Horsten's rendering of Fitch's argument seems to be even more basic. In the case of *CT* I started with a well-established and mathematically fruitful object theory, Peano Arithmetic; the addition of semantic vocabulary and of axioms governing its behaviour determined the possibility of arriving at new theorems, so enriching an already rich theory with more mathematical results. I then tried, for philosophical purposes, to make intelligible the distinction, proper of informal metatheory, between quantification over syntactic entities and quantification over natural numbers. As a result, we were no longer able to contribute to the subject matter of our object theory, although (quite surprisingly) our metamathematical reflection *on* the object theory proceeded undisturbed.

In the case of F I started from an arbitrary, first-order *language* equipped with a fresh predicate K. This predicate acquired 'meaning' only when syntactic and semantic vocabulary was added to the language and suitable axioms governing the interaction of a syntactic name for K and the truth predicate were adjoined to the theory. The truth predicate did not play a significant role in boosting our capability of doing epistemology; it played a more basic role, without it we would not be able to *formulate* the epistemological theory. The case of F^* is illuminating: the theory of truth and the ontology of expressions gave us all that we needed to carry out the reasoning involved in Fitch's paradox. The theory of truth played an indispensable expressive role, and it committed us only to a universe of linguistic objects. Far from dismissing our characterization of typed, compositional truth, it seems, Horsten's example appears to accentuate some of its traits.[34]

[34] There are some further worries related to Horsten's theory F. A first worry concerns the non uniform treatment of modal notions: intensional notions, either epistemic or alethic, seem to belong to a single syntactic category. If one opts for a predicate for knowledge, then one should also opt for a predicate

3.2 Measuring metatheoretic ascent

The arguments reported in Sect. 2.2 seem to suggest that a fundamental component of the metatheoretic realm, that the theories of truth in the style of $CTD^{+}[\cdot]$ are devised to capture, is represented by the capability of performing inductive reasoning involving semantic notions. The syntactic induction principle SIN, essential in the provability of GR_B and thus Con_B, appears to play a fundamental role in the construction of the theories. In this subsection we report and discuss a surprising fact: once *Reflective Adequacy* is generalized and from provability we move to relative interpretability as the privileged mean of reduction, compositional truth axioms are sufficient to reproduce inductive reasoning involving semantic notions. But even more is true: the formal facts reported below — essentially due to Richard Heck and to the author, building on Albert Visser's work — suggest that the logical resources that a compositional truth predicate adds to a suitable base theory B can be matched exactly with intensional metamathematical claims on the base theory B.

As before, let Con_B be an intensionally correct consistency statement whose properties are provable in the 'syntactic' part of our theory of truth. Con_B is the formalized assertion of consistency of B based on a p-time description of B (cf. p. 10).

Proposition 1 *Let B be finitely axiomatized. CTD[B] is mutually interpretable with* $Q + Con_B$.

Proof (*Sketch*) For a detailed proof I refer to Nicolai (2015).

To interpret $Q + Con_B$ in $CTD[B]$, one employs the fact that the latter proves the consistency of B on a cut.[35] In other words, $CTD[B]$ proves the relativization of Con_B to a $CTD[B]$-definable initial segment of the natural numbers as seen by the 'syntactic' part of $CTD[B]$.

For the other direction, we employ the fact that an arithmetized (term) model for B can be constructed in $Q+Con_B$. This 'model' comes with a truth predicate that works for sentences in a $Q+Con_B$-definable initial segment of the natural numbers. The interpretation is thus obtained by relativizing the syntactic quantifiers to this initial segment and by mapping the truth predicate of $CTD[B]$ into the truth predicate of the model. □

The result of adding truth axioms endowed with minimal syntactic resources yields a minimal theory of syntax plus an intensionally correct consistency statement for the base theory B. To see that Proposition 1 improves, let alone sustaining, the proposed characterization of typed, compositional axioms, I need to invoke a fascinating

Footnote 34 continued
for possibility. F and F^*, as we have seen, feature by contrast an operator for possibility (or necessity) and a predicate for knowledge. But if one treats also possibility as a predicate, then one has to revise the entire setting as a typed truth predicate does not suffice anymore — or, in the worse scenario, one might fall pray of paradoxes resulting from the interaction of modal notions. A second worry concerns the solution to the paradoxes that is adopted in this context: even in the traditional analysis of knowledge as justified true belief, a finite number of iterations of the truth (or satisfaction) predicate is generally required. In F and, *a fortiori* in F^*, this is not possible.

[35] This direction is due to Richard Heck.

strengthening of Gödel's Second Incompleteness Theorem due to Pudlák (1985) telling us that Con_B is irreducible to B under very minimal assumptions on B. In other words, not only any recursively enumerable theory B cannot prove Con_B; it cannot even interpret $Q + Con_B$.[36]

Corollary 1 *CTD[B] is not interpretable in B.*

Proof (*Proof Sketch*) If *CTD[B]* (or $CTD^+[B]$) were interpretable in B, also $Q+Con_B$ would be interpretable in it, contradicting Pudlák's strengthening of Gödel's second Incompleteness Theorem (cf. Hájek and Pudlák 1993 and above). □

Typed, compositional axioms with a minimal, self-sufficient machinery to talk about the syntax of the base theory B, produce a clear metamathematical gain: they equate a consistency statement for B. This is even more apparent if one looks at a refined formulation of Proposition 1 that can be found again in Nicolai (2015) Con_B can be seen as the unique Π_1-sentence π — verfiably in $I\Delta_0+\mathsf{exp}$ or EA (cf. Hájek and Pudlák 1993 for the definitions) — such that $CTD[B]$ is mutually interpretable with $Q + \pi$. This means that, from the perspective of a rather weak subsystem of Peano Arithmetic, an intensionally correct consistency statement for B is the unique solution to the equation with one variable (i.e. π)

$$CTD[B] \text{ is mutually interpretable with } Q + \pi.$$

In a sense, therefore, SIN appears to be dispensable for the purpose of detecting a non reducible character of compositional truth. A natural worry comes immediately to mind, though: does $CTD[B]$ satisfy the adequacy requirements set up by the proponents of the conservativeness argument? In Sect. 1, we mostly focused on compositionality and *Reflective Adequacy*. Now (typed) compositionality is obviously satisfied. Rather surprisingly, if we consider a more liberal version of *Reflective Adequacy*, $CTD[B]$ turns out to be adequate also in this sense. $CTD[B]$ in fact proves

$$\forall a\, \forall \ulcorner\varphi\urcorner\, (K(\ulcorner\varphi\urcorner) \wedge Bew_B(\ulcorner\varphi\urcorner) \Rightarrow\ Sat(a, \ulcorner\varphi\urcorner)) \tag{7}$$

where K is a $CTD[B]$-definable initial segment of the numbers living in the disjoint, 'syntactic' universe in which the syntax of B is formalized. In other words, *Reflective Adequacy* is satisfied relative to a definable cut K. Crucially, K contains all standard numbers and the numbers in K satisfy a copy of Q: therefore, the relativization to K of the provability predicate $Bew_B(\cdot)$, that we may call $Bew_B^K(\cdot)$, strongly represents, or binumerates, provability in B. If *Reflective Adequacy* requires the theory of truth explicitly assert the soundness of its base theory, *CTD[B]* satisfies this requirement precisely because all standard proofs lie in the cut K.[37]

The satisfiability of a relativized form of *Reflective Adequacy* in *CTD*[B] might also suggest a new approach to the problem of how to formally render the deflationist's stance on the role of the truth predicate as a device for generalizations. In Sect. 2

[36] We refer to Pudlák's paper or to Hájek and Pudlák (1993) for the proof.

[37] As an anonynous referee has pointed out, a similar justification of the proposed adequacy requirement has been independently proposed by Fischer and Horsten (2015).

we were content with a quite strong characterization: if we accept an infinite set of formulas $\mathcal{S}$, and this set is defined by an arithmetical formula $\Phi(x)$, then an adequate theory of truth should be able to *prove*

$$\forall \ulcorner \varphi \urcorner (\Phi(\ulcorner \varphi \urcorner) \Rightarrow \varphi \text{ is true}) \tag{8}$$

In the new rendering, the truth predicate of a deflationary acceptable theory of truth should be able to at least *interpret* (8) in such a way that the ‘meaning’ of the set $\mathcal{S}$ does not change after the translation: To achieve this one might require, as in the case of (7), that the formula defining $\mathcal{S}$ holds of a significant or relevant portion of the original domain. Several variations are also possible and we defer a more extensive treatment of this issue to a forthcoming work.

At any rate, once the attention is shifted to relative interpretability as the main form of reduction, we still have reasons to conclude that the notion of truth depicted by compositional, typed truth axioms is not reducible to the logical resources of any reasonable base theory.[38] As envisaged by Horsten’s passage in Sect. 1.4 and by Fischer (2010), this conclusion goes in the direction of any form of deflationism that accepts a theory of truth that is conservative but not interpretable in the object theory: The conservativeness of $CTD[B]$ immediately follows, in fact, from the argument given in Sect. 2.1. Still adhering to the parallelism drawn by Horsten, this form of deflationism would hold that the notion of truth does not play a substantial role in object-theoretic argumentation, but it is nonetheless irreducible to the conceptual apparatus assumed by the object theory.

The principles that govern the truth predicate, once laid down, give rise to a notion that properly enriches our conceptual apparatus: there is no object-theoretic or syntactic concept that can mimic the inferential behaviour of this notion of truth, as Corollary 1 shows. Also when our capability of performing metatheoretic inductive reasoning is threatened by the absence of SIN, compositional truth axioms give us the means to recover some portions of it. Once again, it seems, a compositional truth predicate essentially enriches our reflection *on* a portion of mathematics, science or philosophy without yielding new, substantial consequences *within* them.

Remark 1 As in Sect. 2.2, if B is schematically or infinitely axiomatized, Proposition 1 and Corollary 1 go through if we include, among the axioms of the theory of truth, the sentence AxT_B declaring the truth of all the axioms of B. For a discussion of this assumption, we refer to Sect. 2.4.

Acknowledgments This work was supported by the Art and Humanities Research Council UK AH/H039791/1 and by the *Analysis Trust*. I would like to thank Martin Fischer, Volker Halbach, Richard Heck, Leon Horsten, Jeffrey Ketland and two anonymous referees for their comments and suggestions.

References

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

[38] See Sect. 2.4 and the remark below for the assumptions on B.

Cieśliński, C. (2010). Truth, conservativeness, provability. *Mind*, *119*(474), 423–436.
Corcoran, J., William, F., & Maloney, M. (1974). String theory. *Journal of Symbolic Logic*, *39*(4), 625–637.
Craig, W., & Vaught, W. (1958). Finite axiomatizability using additional predicates. *The Journal of Symbolic Logic*, *23*, 289–308.
Enayat, A., & Visser, A. (2013). *New constructions of satisfaction classes*. Logic Preprints Series. Utrecht: University of Utrecht.
Feferman, S. (1984). Towards useful type-free theories. *Journal of Symbolic Logic*, *45*, 75–111.
Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*, 1–49.
Field, H. (1994). Deflationist views of meaning and content. *Mind*, *103*, 249–285.
Field, H. (1999). Deflating the conservativeness argument. *Journal of Philosophy*, *96*(10), 533–540.
Field, H. (2005). Compositional principles vs. schematic reasoning. *The Monist*, *89*(1), 9–27.
Fischer, M. (2010). Minimal truth and interpretability. *The Review of Symbolic Logic*, *2*, 799–815.
Fischer, M., & Horsten, L. (2015). The expressive power of truth. *The Review of Symbolic Logic*. doi:10.1017/S1755020314000392.
Fitch, R. (1963). A logical analysis of some value concept. *Journal of Symbolic Logic*, *28*, 135–142.
Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, *33*, 1–21.
Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, *21*, 57–81.
Hájek, P., & Pudlák, P. (1993). *Metamathematics of first-order arithmetic*. Berlin: Springer.
Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, *108*(429), 1–22.
Halbach, V. (2001a). How innocent is deflationism? *Synthese*, *126*, 167–194.
Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
Halbach, V., & Horsten, L. (2005). The deflationist's axioms for truth. In J. C. Beall & B. Armour-Garb (Eds.), *Deflationism and paradox* (pp. 203–217). Oxford: Oxford University Press.
Heck, R. (2004). Truth and disquotation. *Synthese, 142*, 317–352.
Heck, R. (2015). Consistency and the theory of truth. *The Review of Symbolic Logic*. doi:10.1017/S1755020314000549.
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 (Studies in the General Philosophy of Science: Vol. 3)* (pp. 173–187). Tilburg: Tilburg University Press.
Horsten, L. (2009). Levity. *Mind*, *118.471*, 555–581.
Horsten, L. (2012). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.
Horwich, P. (1998). *Truth* (2nd ed.). Oxford: Clarendon Press.
Ketland, J. (1999). Deflationism and Tarski's Paradise. *Mind*, *108*(429), 69–94.
Kotlarski, H., Krajewski, S., & Lachlan, A. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bullettin*, *24*, 283–293.
Leigh, G., & Nicolai, C. (2013). Axiomatic truth, syntax and metatheoretic reasoning. *The Review of Symbolic Logic*, *4*, 613–626.
Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, *2*(2), 276–290.
McGee, V. (2006). In praise of a free lunch: Why disquotationalists should embrace compositional semantics. In T. Bolander, et al. (Eds.), *Self reference*. Stanford: CSL Publications.
Nicolai, C. (2014). *Truth, deflationism and the ontology of expressions. An axiomatic study*. DPhil Thesis, Oxford.
Nicolai, C. (2015). A note on typed truth and consistency assertions. *Journal of Philosophical Logic*. doi:10.1007/s10992-015-9366-6.
Pudlák, P. (1985). Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, *50*(1985), 423–441.
Reinhardt, W. (1986). Some Remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, *15*, 219–251.
Sieg, W. (1999). Hilbert's Programs: 1917–1922. *The Bulletin of Symbolic Logic*, *5*, 1–44.
Shapiro, S. (1998). Truth and proof: Through thick and thin. *Journal of Philosophy, 95*, 493–521.
Shapiro, S. (2002). Deflation and conservation. In V. Halbach & L. Horsten (Eds.), *Principles of Truth*. Frankfurt: Dr. Hänsel-Hohenhausen.
Tennant, N. (2002). Deflationism and the Gödel phenomena. *Mind*, *111*, 551–582.

Tarski, A. (1956a). The concept of truth in formalized languages. In H. J. Woodger (Trans., Ed.), *Logic, semantic, metamathematics: Papers of Alfred Tarski from 1922–1938* (pp. 152–278). Oxford: Clarendon Press.