Truth is Simple

Leon Horsten
University of Bristol
leon.horsten@bristol.ac.uk

Graham E. Leigh
University of Gothenburg
graham.leigh@gu.se

Even though disquotationalism is not correct as it is usually formulated, a deep insight lies behind it. Specifically, it can be argued that, modulo implicit commitment to reflection principles, all there is to the notion of truth is given by a simple, natural collection of truth-biconditionals.

> In contrast to an arbitrary procedure for moving from A_k to A_{k+1}, a reflection principle provides that the axioms of A_{k+1} shall express a certain trust in the system of axioms A_k.
>
> Solomon Feferman (1962, p. 261)

1. Introduction

John Burgess published a paper with the title 'The Truth is Never Simple' (Burgess 1986). What he meant was that the *extension* of the truth predicate in a typed, and even more so in a type-free approach, is complicated. This cannot be disputed. But we argue that the *intension* of the truth predicate is simple, in the sense that the content of the concept of truth is given by a simple and natural collection of truth-biconditionals. In other words, we claim that some form of *disquotationalism* must be in some sense correct. From a logical point of view, this takes us to the area of proof-theoretic approaches to truth, and away from the area of model-theoretic approaches to truth, which was the focus of Burgess (1986).

Arguments by Shapiro (1998) and Ketland (1999), based on observations by Tarski, have shown that certain standard formulations of disquotationalism are untenable. The fact that truth is compositional cannot be fully accounted for by disquotational axioms alone. Moreover, disquotational principles alone do not seem to do *justice* to the role that truth plays in metamathematical reasoning. In particular, compositional truth principles can be used to show that

Mind, Vol. 126 . 501 . January 2017
doi:10.1093/mind/fzv184 Advance Access publication 8 November 2016

reflection principles hold, and thus to justify reflection principles, whereas disquotational principles are too weak to do this.

Our position in this article is that disquotational principles none the less capture the core content of the concept of truth. When reflection principles are applied to (proof-theoretically weak) disquotational principles against the background of a weak syntax theory, strong compositional theories result. And when we are committed to a weak disquotational theory of truth, then we are *implicitly committed* to reflection principles for it. Therefore the compositionality of truth is implicitly contained in disquotational principles.

Shapiro and Ketland argue that reflection principles are justified by appeal to compositional truth principles. This road is patently not open to us. So the onus is on us to deliver an alternative account of our epistemic warrant for reflection principles. We shall provide this by appealing to Tyler Burge's distinction between *justification* for and *entitlement* to beliefs. We argue that we are commonly in a situation where we are entitled to rely on and even believe in reflection principles, even though we do not have a justification for them.

The structure of this article is straightforward. We first discuss traditional forms of disquotationalism. Then we turn to the critique of disquotationalism by Shapiro and Ketland. Subsequently we outline how compositionality follows from disquotational theories, modulo reflection principles. (Sketches of proofs are relegated to a technical appendix.) We then compare our view with proposals by Field and Halbach that seek somehow to derive the compositional nature of truth from disquotational principles. Finally, we give an epistemological account of the notion of implicit commitment to reflection principles.

In the literature, the discussion of the relation between disquotationalism and reflection principles is mostly restricted to a typed setting, where Tarski's distinction between metalanguage (the language of truth) and object language (the language of the background syntax theory, which we identify with a weak arithmetical theory) is maintained. In this article, we shall not only be occupied with typed theories, but will give an account of the relation between disquotationalism and compositionality in a type-free environment as well.

We assume that the reader is familiar with the basic theories of arithmetic *EA* (Elementary Arithmetic) and *PA* (Peano Arithmetic).[1]

[1] For the purposes of this paper we assume *PA* is formulated with an additional function symbol 2^x for binary exponentiation and the axioms $2^0 = 1$ and $\forall x(2^{x+1} = 2^x + 2^x)$; *EA* is then the sub-theory of *PA* in which induction is restricted to bounded (Δ_0) formulae.

These will be used as background theories of syntax, modulo coding. In the interest of readability, we shall be somewhat sloppy with the details of coding (except in the technical appendix), and generally identify a formula ϕ with its code $\ulcorner\phi\urcorner$. We also assume familiarity with the most important typed and type-free disquotational and compositional theories of truth. Specifically, we assume that the reader is acquainted with *TB* (Tarski-biconditionals), *UTB* (*uniform* Tarski-biconditionals), *CT* (Compositional Truth), and *KF* (Kripke-Feferman). Precise definitions and extended discussions of these theories can be found in the appendix at the end of this paper and in Halbach (2011).

Reflection principles play a central role in this article. In particular, we shall be concerned with:

(1) *Local reflection principles* (Rfn_S), which are schematic principles of the form $Prov_S(\phi) \rightarrow \phi$, where ϕ ranges over a collection of sentences of some language $\mathcal{L}$, S is a theory in $\mathcal{L}$, and $Prov_S$ is a canonical provability predicate for S.

(2) *Uniform reflection principles* (RFN_S), which are schematic principles of the form $\forall x : Prov_S(\phi(x)) \rightarrow \phi(x)$, where $\phi(x)$ now ranges over a collection of formulae with one free variable of some language.

(3) *Global reflection principles* (GRP_S), which are axioms of the form

$$\forall \phi \in \mathcal{L} : Prov_S(\phi) \rightarrow T\phi$$

For a brief discussion of these types of reflection principles, see Halbach (2011, ch. 22).

2. From Tarski to disquotationalism

Tarski (1935, pp. 187ff.) famously proposed *Convention T* as a (material) adequacy condition on a definition of truth for a language $\mathcal{L}$:

> A truth definition for $\mathcal{L}$ must entail, for every sentence $\phi \in \mathcal{L}$, the *Tarski biconditional*
>
> ‘ϕ’ is true $\leftrightarrow \phi$

Tarski's theorem of the undefinability of truth entails that a sufficiently strong theory cannot consistently contain a truth definition for its own language that satisfies Convention T. But natural language

seems maximally expressive, in the sense that whatever can be expressed at all can be expressed in natural language. So no matter how strong a consistent theory we formulate in natural language, it cannot contain a truth predicate that satisfies Convention T. Yet natural language does contain a concept of truth.

Tarski concluded from this that the concept of truth in natural language must therefore be incoherent (Tarski, 1935, p. 264). But at the same time it is clear that there are extensive fragments of natural language, and scientific theories that do not contain the concept of truth, that are coherent. For such languages *cum* theories, a truth definition can be constructed in a more expansive metalanguage: Tarski showed us how.

In the 1950s and 1960s, under the influence of Wittgenstein and his *ordinary language* philosophy, natural language was to some extent rehabilitated in analytical philosophy (Soames 2003). This rehabilitation extended to the concept of truth. Moreover, ordinary language philosophers drew attention away from precise extensions of concepts and instead emphasized the *use* of concepts in ordinary language. Applied to the concept of truth, this means that, *pace* Tarski, giving a definition of the extension of the concept of truth in natural language is not called for. We should try to capture the use of the concept of truth without trying to define truth.

Tarski's Convention T provides key information about the use of the concept of truth. It is a device for *quotation* (right-to-left) and for *disquotation* (left-to-right). If this is indeed the central (and perhaps the only) function of the concept of truth, then its use can be captured in an infinite collection of *axioms*. They are called *truth-biconditionals* (or *Tarski-biconditionals*).

Ultimately, we are interested in providing a satisfactory theory of truth for natural language. But, as is customary in truth theory today, in order to test disquotationalism (and ideas like it), we will work with a toy language, a miniature version of natural language. The miniature language that we opt for is $\mathcal{L}_T$, which consists of the language of arithmetic $\mathcal{L}_{PA}$ plus a primitive truth predicate (T).

Suppose now that we want to describe the use of the concept of truth for $\mathcal{L}_T$. Then, inspired by Convention T, and determined to keep metalanguage and object language separate, we can suggest the restricted truth-biconditionals for this purpose:

Axiom 1 (TB) For all sentences $\phi \in \mathcal{L}_{PA}$: $T(\phi) \leftrightarrow \phi$

This axiom scheme to a significant extent captures the disquotational role of the concept of truth.

As was mentioned in the introduction, TB presupposes a background theory of syntax. Ultimately, we will opt for a very weak background theory, namely, Elementary Arithmetic (*EA*).[2] But more often than not, in the literature, a stronger background is presupposed, namely, Peano Arithmetic. For the time being, *TB* will denote the restricted Tarski-biconditionals with an unspecified background arithmetical theory, which can be *EA*, *PA* without the truth predicate appearing in the induction scheme (henceforth simply *PA*), or *PA* with induction extended to the language including the truth predicate (PA_T). When choice of background theory matters, we will say so and be more definite. In particular, later in this article we will officially take *EA* as our background syntax theory. In fact, one of the aims of the present article is to sidestep discussions about the background syntax theory.[3]

Some authors believe that free paramaters should be allowed in the disquotational axiom. This yields the stronger axiom scheme:

Axiom 2 (UTB). For all formulae $\phi(\vec{x}) \in \mathcal{L}_{PA}$: $\forall\vec{x}[T(\phi(\vec{x})) \leftrightarrow \phi(\vec{x})]$

Again, without risk of confusion we can denote by *UTB* the theory resulting from adding axiom 2 to a background syntax theory.

It is not hard to see that axiom 2 entails the distribution of truth over the universal quantifier: $\forall x T(\phi(x)) \leftrightarrow T(\forall x \phi(x))$. The truth of this statement is directly supported by an intuition that says that truth is compositional. But axiom 2 somewhat exceeds the immediate content of disquotationalism. (Indeed, it turns out that the theory *TB* does not entail axiom 2.) So we will take the content of the disquotational intuition to be captured by a scheme that quantifies over *sentences* rather than over formulae with free variables.

At first sight, it appears that the theory *TB* is *incomplete*: the restricted truth-biconditionals say nothing about the disquotational role of sentences that themselves contain the concept of truth.[4] So, *a fortiori*, the restricted Tarski-biconditionals cannot be seen as a *definition* of truth for $\mathcal{L}_T$. But Tarskian considerations suggest that this may be a

[2] *EA* is defined in the appendix, definition 1.

[3] More about this below: see §5.

[4] As mentioned earlier, later in this article we will also consider truth theories that speak to the disquotational role of sentences that themselves contain the concept of truth.

virtue rather than a vice. Tarski's proof of the undefinability of truth is based on instantiating the liar sentence into the Tarski-biconditional scheme. But the liar sentence itself contains the concept of truth. In sum, according to Tarski's diagnosis of the liar paradox, we are well advised not to allow the concept of truth inside the truth-biconditionals. This is known in the literature as *typing* the concept of truth. And this is of course precisely what happens in TB.

We thus arrive at *TB* as a natural disquotational theory of truth. Disquotationalists hold that quotation and disquotation are the only functions of the concept of truth: the (restricted) Tarski-biconditionals are all there is to truth (Williams 1988, p. 424). Quine put it as follows:

> The truth predicate is a device of disquotation. We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. (Quine 1970, p. 307)

A version of disquotationalism in terms of propositions instead of sentences is defended in Horwich (1998). He emphasizes that truth is a *simple concept*:

> [The Tarski-biconditionals] could be explained only by principles that are simpler and more unified than they are—principles concerning propositional elements and the conditions in which truth emerges from combining them. But the single respect in which the body of minimal axioms is not already perfectly simple is that there are so many of them—infinitely many; and no alleged explication could improve on this feature. For there are infinitely many constituents to take into account: so any characterization of them will also need infinitely many axioms. (Horwich 1998, p. 51)

We shall argue that in this Horwich is basically correct: the core content of the concept of truth is captured by a natural and simple collection of Tarski-biconditionals.

3. Compositionality and the limitations of disquotationalism

There is a fly in the ointment—as Tarski already knew. It seems fundamental to our concept of truth that it is a *compositional* notion: the concept of truth distributes over the logical symbols. For instance, it seems that the following statement is acceptable:

$$\forall \phi, \psi \in \mathcal{L}_{PA} : T(\phi \wedge \psi) \leftrightarrow [T(\phi) \wedge T(\psi)]$$

Tarski rejected *TB* as a theory of truth because of its deductive weakness:

> The value of [the theorem that *TB* is consistent] is considerably diminished by the fact that [the Tarski-biconditionals] have a very restricted deductive power. A theory of truth founded on them would be a highly incomplete system, which would lack the most important and most fruitful general theorems. (Tarski 1935, p. 257)

Indeed, it is not hard to see that *TB* does not entail compositional truth principles such as principle (1). A proof from *TB* (or *UTB*) can only use finitely many Tarski-biconditionals, whilst the compositional principles make a claim about infinitely many sentences. Every instance of the aforementioned compositional principles can be proved from *TB*; the universal closure cannot. This is why *TB* is incomplete as a theory of truth.

Moreover, there are statements from the background theory *S* that we expect a theory of truth for *S* to prove. For a sufficiently strong background theory *S*, we expect a theory of truth to prove the Gödel sentence G_S of *S* (Shapiro 1998; Ketland 1999, §6) and the local reflection principle Rfn_S for *S*. When we set set $S \equiv PA$, and take *PA* as our background arithmetical theory, we should expect *TB* to prove G_{PA} and Rfn_{PA}, but it does not, because *TB* is proof-theoretically conservative over *PA* (Halbach 2011, ch. 7).

We expect a truth theory for *PA* to prove G_{PA} and Rfn_{PA} because their truth can be established by *truth-theoretic reasoning*. Let us consider Rfn_{PA} first. The axioms of *PA* and of logic are all true. The rules of inference preserve truth. Therefore every theorem of *PA* is true. Now consider G_{PA}. By the diagonal lemma, this sentence is true if and only if it is unprovable in *PA*. Suppose G_{PA} were provable in *PA*. Then, since Rfn_{PA} holds, G_{PA} would be true. But by the diagonal lemma that means that it would be unprovable in *PA*. Contradiction. So G_{PA} is unprovable in *PA*. So by the diagonal lemma again, G_{PA} is true.

From this, truth theorists tend to draw the conclusion that the compositionality intuition about truth is not reducible to the disquotational intuition (Horsten 2011, p. 70). An option that is taken by many at this point is to take the compositional truth principles as *basic* axioms of truth. This option is taken in, for instance, Shapiro (1998) and Ketland (1999). This leads to the proposal of accepting, instead of *TB*, the compositional theory *CT* (which includes principles such as (1) as *axioms*) as a basic theory of truth.

The advantage of this proposal is threefold. First, *CT* is a simple and natural set of axioms for truth. As a truth theory, it is not as simple as *TB*, but it is still a simple theory of truth. That it is natural is shown by the fact that it has been embraced by leading philosophers such as Davidson (1967). Second, *TB* is a sub-theory of *CT* (Horsten 2011, ch. 6). So the disquotational intuition is contained in the compositional intuition as expressed by *CT*. Third, *CT* does entail metamathematical statements such as G_{PA} and Rfn_{PA} that can be recognized to be true on the basis of truth-theoretic reasoning (Ketland 1999).

The structure of the standard proof of the non-conservativeness of *CT* goes as follows. First, by an induction, in *CT*, on the length of proofs in *PA*, the *global reflection principle* GRP_{PA} for *PA* is proved. Then, in a second step, the truth-biconditionals for *PA* (which are provable in *CT*) are used to derive the local reflection principle Rfn_{PA}, which is of course unprovable in *PA*.

It has recently become clear that the moral of our discussion also extends to the type-free setting. In particular, it applies to Feferman's system *KF*, which is the most popular type-free compositional theory of truth.[5] In particular, is commonly recognized to be a very natural way of extending *CT* to a consistent type-free system.

The unrestricted truth-biconditionals are of course inconsistent. Until recently, it was not clear how to construct a natural consistent type-free disquotational theory of truth. However, the situation changed with the publication of Halbach (2009). In this article, Halbach proposes the *Positive Uniform Tarski-biconditionals* (PUTB), containing the universal closure of every truth-biconditional $T(\phi(\vec{x})) \leftrightarrow \phi(\vec{x})$, where ϕ is a formula of $\mathcal{L}_T$ in which the truth predicate only occurs in the scope of an even number of negation signs: call such formulae *truth-positive formulae*. Halbach proves that PA_T + PUTB is a sub-theory of *KF* and can define a truth predicate that satisfies the *KF* axioms.

As before in the typed setting,[6] the scheme PUTB exceeds the content of disquotationalism somewhat. So, for now, we will take as a natural candidate for a type-free disquotational scheme the following:

[5] Another popular, but decidedly less popular, type-free compositional theory of truth is the Friedman-Sheard system *FS* (Friedman and Sheard 1987). The main drawback of *FS* is that it is ω-inconsistent. For that reason, we leave it aside in the present article.

[6] See our comparison of *TB* with *UTB* in the previous section.

Axiom 3 (PTB). For all truth-positive sentences $\phi \in \mathcal{L}_T$: $T(\phi) \leftrightarrow \phi$

The result of adding axiom 3 to the background theory is called *PTB*, and the extension by the uniform version of axiom 3 is called *PUTB*. *PTB* is a weak truth theory: it is conservative over its background theory (Cieśliński 2011).[7] For the proof-theoretic strength of *PUTB* it matters whether induction is permitted for the truth predicate. If induction is not extended, then it is proof-theoretically conservative over the background theory; if induction on all formulae of $\mathcal{L}_T$ is chosen, then it is highly non-conservative over the background theory. Again, we need not be very concerned with this phenomenon in this paper for reasons that will be explained later. At any rate, neither *PTB* nor *PUTB* proves the compositional principles of truth, not even the typed ones (see, for example, Halbach 2011, lemma 19.20). The argument is structurally identical to the proof that *TB* and *UTB* cannot prove the compositional axioms of *CT*. So in the type-free setting too, the compositionality of truth seems to have more content than the disquotationality of truth.

4. Reflection and justification

Let us return to the typed setting. In §3 the conclusion was drawn that *TB* is too weak and should instead be replaced by *CT* (Shapiro 1998; Ketland 1999). Tennant (2002) takes it upon himself to defend the disquotationalist against this line of reasoning.

Tennant first points out that it is not the case that metamathematical justification crumbles if one is only allowed to make use of a conservative, disquotational concept of truth. For instance, it is a commonplace that one can prove G_{PA} from weak reflection principles such as Rfn_{PA}.[8] Secondly, weak reflection principles express '[one's] willingness …, via [Rfn_{PA}], to assert any theorem of [*PA*]' (Tennant 2002, p. 574).

Ketland, in his reply to Tennant, expresses dissatisfaction with Tennant's strategy for simply *postulating* reflection principles. He replies that reflection principles such as Rfn_{PA} ought instead to be *proved* (Ketland 2005, p. 82), and stresses that the truth theory *CT* does indeed prove such reflection principles. In other words, Ketland objects that Tennant's justification of G_{PA} is shallower than his own

[7] See also theorem 12 in the appendix

[8] In fact, Tennant stresses that in order to prove G_{PA} the full force of Rfn_{PA} is not needed. All that is needed is the schematic principle that asserts $Prov_{PA}(\phi) \rightarrow \phi$ for every primitive recursive formula ϕ of the language of *PA*.

justification, for some of the justifying principles (viz., the reflection axioms) that are postulated in Tennant's justification are themselves proved in Ketland's justification of G_{PA}.

Tennant rejects Ketland's request for a justification of reflection principles such as Rfn_{PA}. He writes:

> No further justification is needed for the new commitment made by expressing one's earlier commitments. As soon as one appreciates the process of reflection, and how its outcome is expressed by the reflection principle, one already has an explanation of why someone who accepts S should also accept all instances of the reflection principle. (Tennant 2005, p. 92)

Ketland finds this response unsatisfactory. He asks, first, what the reflective process of which the passage speaks is supposed to be. Tennant is not explicit about this, so Ketland offers an answer to this question. The reflective process consists in stepping back from one's practice and realizing that one is ready to accept every theorem of *PA* (Ketland 2010, p. 428). So the outcome of the reflection process, in Ketland's view, is the conclusion 'I am ready to accept every theorem of *PA*' (Ketland 2010, p. 433). Second, Ketland disputes that the realization of this disposition is adequately expressed by Rfn_{PA}:

> It should be noted that this is a non-standard claim. Usually, a reflection scheme like $[\mathrm{Rfn}_{PA}]$ is said to express the *soundness* of $[PA]$: that whatever $[PA]$ proves is *true*. And being *true* is not the same as being accepted. (Ketland 2010, p. 430)

All this underscores that the process of reflection is not well understood: it is not clear what it consists in, what its outcome is, and how that can be formally expressed (if it can be formally expressed at all). Ketland holds that reflection principles should be and can be justified by means of principles of truth, whereas Tennant regards reflection principles as the outcomes of reflective processes that need no further justification.

Cieśliński then enters the debate between Ketland and Tennant (Cieśliński 2010). First of all, he takes CT^-, the version of *CT* which has *PA* (induction only for $\mathcal{L}_{PA}$ formulae) as a background theory but where the truth predicate is not allowed in the induction scheme, as an unobjectionable base theory. The theory CT^- goes beyond *TB* in one sense, because *TB* cannot prove the compositional truth axioms. Thus CT^- cannot be seen as an expression of disquotationalism. But CT^- is weaker than *TB* in another sense: the truth predicate is not allowed in the induction scheme of CT^-, and hence the theory is proof-theoretically conservative over its background

theory (Kotlarski et al. 1981).[9] The background theory of *TB* may, however, feature the full induction scheme missing from CT^-. So in that sense, many deflationists who are not strict disquotationalists find CT^- unproblematic.

Cieśliński agrees with Tennant (contra Ketland) that reflection principles are not themselves in need of justification. He furthermore holds that *uniform* reflection principles express the outcome of reflective processes. But he agrees with Ketland that even *global* reflection principles, which allow us to express the content of an infinite sequence of uniform reflection principles, ought to be provable.

Cieśliński shows that $CT^- + \mathrm{GRP}_{PA_0}$ does not prove RFN_{PA}, where we take PA_0 to stand for Peano Arithmetic formulated in the language *without* the truth predicate. In other words, adding a strong reflection principle for the *restricted* language $\mathcal{L}_{PA}$ to CT^- does not give us even uniform reflection for the *extended* language $\mathcal{L}_T$. This means that adding reflection principles for the restricted language cannot give us what we want.

Cieśliński considers the following uniform reflection principle:

Axiom 4 ($\mathrm{RFN}_\varnothing$) $\forall x : \mathit{Prov}_\varnothing(\phi(x)) \rightarrow \phi(x)$ for all formulae $\phi \in \mathcal{L}_T$

where $\mathit{Prov}_\varnothing$ expresses provability from the empty theory, that is, first-order logical theoremhood. So this principle expresses the truth of theorems of (first-order) logic in the extended language.

Then Cieśliński shows that in CT^-, the full induction axiom for the extended language $\mathcal{L}_T$ follows from axiom 4 (Cieśliński, 2010, p. 419):

Theorem 1. $CT^- + \mathrm{RFN}_\varnothing \vdash CT$

But we know that *CT* entails the global reflection principle for the extended language $\mathcal{L}_T$, which is all the reflection that one might want in a typed setting.

In the type-free setting, something similar can be said. The Kripke-Feferman theory without truth allowed in the induction scheme, often called KF^-, is conservative over its background theory *PA*. But if we add reflection over logic to KF^- in the extended language, then we obtain the full (and highly non-conservative) theory *KF* (Cieśliński, 2010, p. 421).

Now Cieśliński argues that axiom 4 is a reflection principle that is the outcome of a reflective process of the sort that Tennant describes:

[9] CT^- is not *semantically* conservative over its background theory (Lachlan 1981): not every model of arithmetic can be expanded to a model of CT^-. Neither is *TB* in the case that its background theory contains induction for the truth predicate (Strollo 2013).

if we reflect upon on our inferential practices, we see that we accept the laws of logic even in the extended language. Note that axiom 4, *unlike* Rfn_{PA}, is a uniform reflection principle. If one agrees with Ketland that the outcome of reflection is not captured by a local reflection principle, then one will *a fortiori* deny that it is captured by a uniform reflection principle. On the other hand, if one believes that the outcome of a reflection process is captured by a reflection principle, then it seems that a uniform reflection principle is as good a candidate as a local reflection principle. Also, even though it is essential that axiom 4 quantifies over all sentences in the extended language $\mathcal{L}$, the truth predicate does not play an 'active' role in this principle. In other words, truth does not play a *substantial* role in axiom 4.

The key element in Cieśliński's account is the suggestion that we can not only reflect on a theory in our background language, but that we can also reflect on our background theory in an *extended language*. Indeed, when we commit ourselves to logic, or to the principle of induction, we commit ourselves to open-ended schemes. In Feferman's terms, logic and Peano arithmetic are *schematic* theories (Feferman 1991, p. 2). Whenever a new bona fide predicate enters our language, we automatically extend these schemes to the new language.

In Cieśliński's process of passing from CT^- to CT via a reflection on logic, conservativeness over the background theory is lost. So from the point of view of a deflationist who holds that a theory of truth ought to be proof-theoretically conservative, the theory $CT^- + \mathrm{RFN}_{\varnothing}$ does not seem attractive as a basic truth theory. But it ought to be remembered that it is not proposed as a *basic* truth theory. The basic truth theory is the conservative theory CT^-. But, especially if, as Field once suggested, truth is a *logical* notion (Field 1999, p. 534), we seem to be implicitly committed to axiom 4.

So that means that, given the basic truth theory CT^-, we are implicitly committed to the full theory CT, and thus to the global reflection principle for arithmetic. In other words, modulo reflection on logic, truth can perform the metamathematical functions (for instance, justifying G_{PA}) that it is meant to.

5. Reflection and disquotation

A key aspect of Cieśliński's proposal is the following. Whereas Ketland proves reflection principles from truth principles, and Tennant argues

that reflection principles need no 'proof', Cieśliński reverses the explanatory direction, at least to some extent. Reflection principles allow us to derive a stronger truth theory (CT) from a weaker one (CT^-). In what follows, we will take Cieśliński's strategy much further. We will show, in fact, that strong compositional truth theories follow from local truth-biconditionals, via reflection principles.

According to the strict disquotationalist, our basic truth theory is *TB*. As far as the truth axioms go, *TB* is weaker than CT^-, which is taken by Cieśliński as the basic truth theory. But as far as the arithmetical principles go, *TB* may be stronger, that is, if it allows the truth predicate in the induction scheme. But, as we have argued earlier, it is reasonable to allow this, for Peano arithmetic is a *schematic* theory. Indeed, CT^- does not seem to represent a stable position. It extends the logical schemes so as to allow occurrences of the truth predicate, but does not extend the same privilege to the induction scheme.

Nevertheless, as announced earlier, we will sidestep the discussion about the background syntax theory and start with an arithmetical theory that does not contain strong schematic principles: Elementary Arithmetic (EA).[10] So from now on, we let TB denote the Tarski-biconditionals for the language of arithmetic, and fix *EA* as the background theory. Then our official disquotational starting point is $TB_0 = EA + \text{TB}$.[11]

Cieśliński's use of implicit commitment is very modest: it is restricted to the reliability of *logic*. But implicit commitment is a general notion. Whenever we have fully committed ourselves to a theory *S*, we are implicitly committed to accepting a reflection prinicple for *S* (Feferman 1991, p. 1). Since the disquotationalist's starting point is TB_0, she is implicitly committed to a reflection principle for TB_0. We have argued earlier that the reflective commitment extends at least to *uniform* reflection over the starting point. That is, the disquotationalist is implicitly committed to RFN_{TB_0}.

[10] It is entirely possible to consider a base theory that is completely free of schematic principles. One option is simply to take as a base theory $I\Sigma_1$, the extension of *EA* by induction for Σ^0_1 formulae, which is known to permit a finite axiomatization. If $I\Sigma_1$ is considered too strong as a background theory, it is possible to proceed using a finitely axiomatized sub-theory of *EA*: cf. footnote 9.2 below.

[11] We adopt the convention in this paper that theories notated using subscripts are formulated without induction for the extended language. Theories notated without subscripts, such as *TB*, *CT*, *KF*, etc., always contain induction for the whole language.

The disquotationalist can reflectively come to realize that she is implicitly committed to RFN_{TB_0} and explicitly accept her commitment to RFN_{TB_0}. Then she has accepted the stronger theory

$$EA + \mathrm{RFN}_{TB_0}$$

Let us call this theory TB_1. In the process, she has also extended her implicit commitment to $\mathrm{RFN}_{EA+\mathrm{RFN}_{TB_0}}$. She can reflectively come to realize this, and explicitly accept

$$EA + \mathrm{RFN}_{EA+\mathrm{RFN}_{TB_0}}$$

We shall call this theory TB_2. It can then be said that a commitment to TB_2 is implicit ('implicitly implicit') when one explicitly signs up to TB. This process TB_0, TB_1, $TB_2, \ldots$ can justifiably be iterated along a transfinite ordinal hierarchy that is as long as the length of verifiable well-orderings that can be generated in the process, in the spirit of Feferman (1991). This transfinite process results in a full description of the implicit commitment of *TB*, which is called the *reflexive closure* of TB. Indeed, in a similar fashion, reflection hierarchies of other truth theories can be constructed, such as $EA + \mathrm{UTB} = UTB_0, UTB_1, UTB_2, \ldots$ But for the purposes of this article, we shall be interested only in the two first movements of extending the basic truth theory by reflection principles.

It can be shown that:[12]

Theorem 2. $TB_1 \vdash UTB + \mathrm{Ind}(\mathcal{L}_T)$

where $\mathrm{Ind}(\mathcal{L}_T)$ is the induction scheme for the language $\mathcal{L}_T$.

So in the first reflective moment, the truth theory is strengthened from the local Tarski-biconditionals to the uniform version. And someone who accepts the disquotational axioms is at least implicitly committed to induction in the extended language. When a person *reflects* on this commitment implicit in *TB*, she explicitly accepts Peano Arithmetic in the extended language. In other words, for a reflective person, the discussion whether truth should be allowed in the induction scheme is immaterial. This is how we sidestep the discussion in the literature on the appropriate background syntax theory for an axiomatic truth theory.

[12] See appendix.

Moreover, it has also been shown (Halbach 2001, §4) that:

Theorem 3. *CT is identical to* UTB_1 *and a sub-theory of* TB_2.

This means that in the second reflective moment, the full compositionality of truth is obtained. In other words, the compositionality of truth is implicitly contained in the disquotational axioms. The compositional axioms for conjunction, disjunction and negation are obtained after just one act of reflection on the local biconditionals (that is, in TB_1); a second instance of reflection is necessary in order to derive the compositional quantifier axioms.

Let us now turn to the type-free environment. Here the positive Tarski-biconditionals (PTB) seem a possible starting point for the disquotationalist. From the compositional point of view, *KF* seems a natural theory. Again, we can consider the hierarchy $PTB_0 = EA + \text{PTB}$, PTB_1, $PTB_2, \ldots$ of truth theories generated from the positive truth-biconditionals by repeated application of uniform reflection. Then we have

Theorem 4. (Halbach 2009). *KF is interpretable in* PTB_1.

Theorem 5. (Halbach 2001). *The 'strictly positive' compositional axioms of KF are derivable in* PTB_2.

Theorem 4 has to be interpreted with care. Halbach's proof (see the appendix for an outline) shows that *KF* (in any reasonable formulation) can be interpreted in a very *simple* manner into *PUTB*, a sub-theory of PTB_1. As in the type-free scenario, it is only after two instances of reflection that the compositional axioms for quantifiers become provable from local biconditionals. Nonetheless, *KF* (in any formulation) is not a sub-theory of PTB_2. For example, in the formulation of *KF* with a single truth predicate, the compositional axioms involving negation (e.g. $T(\neg T(\psi)) \leftrightarrow T(\neg\psi)$) will not derivable. So we refrain from claiming that someone who accepts PTB_0 is implicitly committed to the compositional theory *KF*.

McGee (1992) has shown, using a diagonal argument, that *every* theory in $\mathcal{L}_T$ can be written as a collection of Tarski-biconditionals in $\mathcal{L}_T$. But the collection consisting of the positive Tarski-biconditionals forms a *natural* theory. Moreover, we know that it is consistent, because it forms a sub-theory of *KF* (Halbach 2009). Nonetheless, it would be desirable to have an argument that *motivates* taking the positive Tarski-biconditionals as one's basic disquotational truth axioms in a type-free setting. In keeping with the main tenets of

disquotationalism, we want such a motivation to start from a basic collection of Tarski-biconditional sentences.

Consider the language $\mathcal{L}_P$ (where '*P*' is short for 'partial'), which is just like the language $\mathcal{L}_T$ except that:

(1) for every atomic predicate A, there is an atomic dual $\bar{A}$ of A;

(2) the negation symbol is removed from the language.

Let the dual of T be denoted as F. Aside from the unrestricted truth-biconditionals for $\mathcal{L}_P$, we also consider the *falsity-biconditionals* for this language, which are the sentences of the form $F(\bar{\phi}) \leftrightarrow \phi$, where $\bar{\phi}$ is the result of replacing in ϕ each connective, quantifier and atomic predicate by its dual.[13] Consider the theory TFB_0 that consists of EA plus the unrestricted truth- and falsity-biconditionals in $\mathcal{L}_P$. The theory TFB_0 seems a good basic starting point. It is no more than a totally unrestricted collection of truth-biconditionals and falsity-biconditionals in a language without negation. In particular, PTB_0 is a sub-theory of TFB_0. Here the concept of negation is seen as the source, when combined with the truth predicate, of the semantic paradoxes.

Then it can be shown that:[14]

Theorem 6. $TFB_1 = EA + \mathrm{RFN}_{TFB_0} \vdash \mathrm{PUTB}$

Theorem 7. *KF is a sub-theory of* $UTFB_1$ *and is strictly contained in* TFB_2.

Here the provability relation $\vdash$ denotes (as always in this article) provability in *classical* predicate logic.

If we string these results together, then we see that in the type-free setting, the full compositional axioms of *KF* can be seen to be implicit (via two reflective moments) in the unrestricted truth- and falsity-biconditionals in a 'liar-proof', negation-less language.

In sum, both in the typed and in the untyped setting, the compositional content of the concept of truth can be taken to derive, through reflection principles, from a purely disquotational conception of truth. So perhaps truth is a very simple concept after all.

Of course, all of this leaves one fundamental question unanswered: how are reflection principles justified? This question was forcefully put on the agenda by Ketland (2005, p. 85). Indeed, it is surprising that

[13] For the precise definition of $\bar{\phi}$, see the appendix.

[14] See appendix.

even though hierarchies of reflection principles had been studied by Kreisel and Feferman since the 1950s, and that they suggested early on that we are implicitly committed to reflection principles for the theories that we accept,[15] philosophers of mathematics have hitherto largely failed to investigate the notion of *implicit commitment*, and have not spent much philosophical energy on analysing our warrant for reflection principles.

We have seen that in Ketland's view, reflection principles are justified by appeal to (compositional) truth principles. Since we instead want to motivate compositional truth from reflection principles (and natural disquotational principles), this road is not open to us. Instead, we argue that our epistemic warrant underwriting reflection principles is of a different kind.

6. Reflection and entitlement

We will now argue that when we are justified in believing a theory, we do not need extra justification for adopting a reflection principle for that theory. In such a situation, we are entitled to adopt a reflection principle without giving additional justification for accepting it. In arguing our point, we draw upon the distinction between *entitlement* and *justification* that was made in Burge (2003).

Our claim goes against the majority view. We have seen that Ketland and Shapiro request a justification for reflection principles. This is the common viewpoint. Also Volker Halbach, for instance, requires justification in this situation: 'the transition from a theory to a reflection principle for that theory requires an argument' (Halbach 2001, p. 1963).

Burge has invoked the distinction between between justification and entitlement in his writings on perception, memory, self-knowledge and logical reasoning (Burge 1993, 1998, 2003, 2007). We believe that this distinction can do similar work in the epistemology of proof-theoretic reflection.

To provide background for our discussion, let us start with perception. We trust our senses. That we do so is made manifest in our everyday beliefs and actions. Most of us *believe* that we trust our senses, but we do not *need* to believe that we trust our senses in order to trust our senses.

[15] See, for instance, Kreisel (1970).

Someone who has the required conceptual machinery can articulate our trust in our senses. Then it becomes a *reflection principle*:

What our senses tell us is (or tends to be) true.

Let us call this the *sense reflection principle*. We are *entitled* to believe in the trustworthiness of our senses. This entitlement does not require a justification of our belief in the trustworthiness of our senses. Indeed, perhaps the sceptic cannot be answered: perhaps there is no non-circular justification of our belief in the trustworthiness of our senses.

We can come to believe in the trustworthiness of our senses through a *reflective act*. When we do so, we acquire a priori belief in, and even knowledge of, the trustworthiness of our senses. The details of the reflective process are complicated. But, for our purposes, the important point is that this reflective process does not typically involve a justification of the trustworthiness of our senses. In sum, through a process of reflection we come to *know* the sense reflection principle. The latter is therefore not an axiom or basic principle, and it is also not justified on the basis of more basic principles.

Note also that we need not come to believe in the trustworthiness of our senses. After all, we do not normally *use* even instances of the sense reflection principle when we rely on our sensory experiences to form our beliefs about the world around us: we do not *reason* from our experiences to our beliefs.

Something similar can be said about our everyday reliance on preservative memory and our reliance on logical inference. We rely on memory in daily life, and we are usually entitled to do so without justification. When we remember something (in normal circumstances), we are entitled without further justification to self-attribute having a memory, but we can rely on our memory without doing so. When we reason, we rely on logical inference rules. Our normal entitlement of doing so is grounded in our knowledge of the logical constants (Burge 2007, p. 197).

There is an important difference between our reliance on perception and our reliance on logical reasoning. In the case of perception, there is always the possibility of *brute error*.[16] We may be in a situation where there is no malfunction of our perceptual faculties, yet we find ourselves in an inclement environment which causes our perceptual beliefs to go astray. There is no possibility of *brute* error in logical

[16] The question of the possibility of brute error in perception, self-knowledge, memory and logical reasoning is discussed in Burge (2007).

reasoning (Burge 2007, p. 196). If our reasoning faculties function as they should, and are appropriately rooted in our understanding of the logical concepts, then we cannot go astray in our reasoning. This is not to say, of course, that faulty reasoning is not possible; it is just that faulty reasoning cannot simply be due to the world not cooperating.

Mistakes in perception, memory and logical reasoning do occur, and we know that they do. But this does not mean that we are invariably entitled only to less than full acceptance of what perception, our memory, or our logical faculties deliver. When our faculties do function appropriately (and, where relevant, the world cooperates), we are entitled to unqualified acceptance of the beliefs based on the functioning of those faculties: there is then no epistemic obligation to hedge.

If Burge's distinction, within the category of epistemic warrant, between justification and entitlement, is applicable to any other domain, then it must be applicable to the domain of reflection principles for mathematical theories and truth theories. In particular, we seek to apply this distinction to the *uniform* reflection principle for *TB*. And we have seen that we seek to apply uniform reflection at least *twice*. How are we to think of this?

Consider a congregation of believers attending mass in church. They accept what is written in the Gospels: they see it as the word of God and they trust God. They have not reflected on their acceptance of what is written in the Gospels. The priest, however, in his sermon articulates the community's trust in God; he reflects on the doxastic mode of the congregation. The members of the congregation could perhaps not even do this by themselves: before the priest has articulated it for them, they may lack the necessary conceptual machinery. But the congregation simply extends its trust to the priest, and (if all is well) it is entitled to do so, without proof or argument. And the priest can even reflect on this extended trust in his sermon …

So it is with reflection principles for mathematical theories and for truth theories. We trust basic arithmetic (*EA*), classical logic, and a basic collection of truth-biconditionals, perhaps even including a set of falsity-biconditionals. This is evident from the way in which we use what we establish in basic arithmetic and in basic disquotational theories. Just as our perceptual states (as representational states) are integrated into our belief system (Burge 2003, p. 521), so are our *arithmetical proof states* integrated into our belief system. Indeed, we indispensably use arithmetical theorems in our best explanations of physical phenomena. For instance, the fact that cicadas have

prime-numbered periodical life-cycles (of thirteen or seventeen years) has been given an evolutionary explanation involving general facts about prime numbers (Baker 2005). Similarly, disquotational reasoning with the truth predicate is integrated into our belief system.

Reflection principles express our trust in theories. In this vein, Halbach (2001, p. 1963) states that our *trust* in *TB* is expressed by a uniform reflection principle for *TB*. We can (but do not need to) reflect on our trust in basic arithmetic or in a basic truth theory. This would result in an explicit acceptance of a reflection principle. The explicit theory that is trusted would thereby be enlarged. The trust in this extended theory is implicit in the acceptance of the extended theory. This trust can (but need not be) be articulated in a second reflective moment.

Compare this with reliance on *memory*. At some level of abstraction, memory can be seen as a box. When we rely on our memory, we take things from this box and integrate them into our system of occurrent beliefs. This is what it means to rely on memory. But *that* memory is reliable is not to be found in the memory-box. It could not be![17] Similarly, our basic theory of arithmetic can be seen as a box. When we do arithmetic (calculate and prove propositions), we take things from the box. We integrate these things into our belief system, and thus rely on them. But *that* basic arithmetic is reliable is not, and cannot be, found in the box.

Implicit commitment to reflection principles is implicit in our explicit and unqualified *acceptance* of a theory *S*. Unqualified acceptance of *S* is close to what Franzén (2004, p. 207) calls *acceptance of S as sound*. Franzén glosses acceptance of *S* as sound as 'accepting *that* the axioms of *S* are true' (2004, p. 213; our emphasis). But this phraseology suggests an explicit propositional attitude (acceptance) towards the truth of *S*, and this is more than what unqualified acceptance of *S* entails. Indeed, accepting *S* in an unqualified way is not the same as accepting a reflection principle for *S*. After all, someone might accept *S* without possessing the concept of truth (for *S*) at all.

Not everyone accepts basic arithmetic, classical logic and a basic set of truth-biconditionals in an unqualified way. Someone who is a constructivist, for instance, will not accept classical logic. Such a person can still *use* classical logic *instrumentally*. For instance, given the constructive provability of the Π_2-conservativeness of classical Peano

[17] For an extensive account of our epistemic warrant for relying on our memory, see Burge (1993, 1998).

arithmetic over Heyting arithmetic, the intuitionist may instrumentally use classical Peano arithmetic to derive intuitionistic arithmetical theorems of restricted complexity. But it will be manifested by the constructivist's assertive practice that she does not fully accept classical logic. Therefore she is not entitled to belief in the uniform reflection principle for classical basic arithmetic. In a related vein, it is possible to accept a theory S in a Hilbertian formalist spirit. This is what Franzén (2004, §14.4) calls *accepting S as consistent*. Someone who accepts S in this sense is implicitly entitled to the consistency of S, but not even to Rfn_S.

More generally, being *justified* in accepting a theory S is a necessary condition for being entitled to accept Rfn_S. Failing to be justified in accepting S can of course come about in multiple ways: S may be false, or one may lack sufficient justification for one of S's axioms… Being justified in accepting S is not quite enough to be entitled to accept Rfn_S, but not much more is needed. One additional thing that *is* needed is to understand that the canonical provability predicate $Prov_S$ expresses provability in S.

So, as is the case with reflection principles for sense experience, memory and logical reasoning, we may be in a situation where we think that we are entitled to accept a proof-theoretic reflection principle (or certain instances of it) when we are not. But, as before, this does not mean that our acceptance of proof-theoretic reflection principles can never be more than qualified acceptance.

Is there, as with sense perception, a possibility of brute error when we accept a proof-theoretic reflection principle for a mathematical theory? This question is difficult to answer. Some argue that our justification for our mathematical theories can be reduced to our understanding of the content of mathematical notions (such as the notion of natural number). If this is so—but this is a big 'if'—then brute error is presumably not possible.[18] If, on the other hand, the subject matters of mathematical theories are independently existing platonic realms of which we have fallible mathematical intuition which is somehow akin to sense perception, then brute error is at least to some extent possible. But this is an equally big 'if'. In sum, whether brute error is possible in relying on mathematical theories that we accept, and in explicitly accepting reflection principles for them, depends on larger issues in the

[18] Burge thinks that *simple* mathematical truths are immune to brute error in this way: see Burge (2007, p. 198).

philosophy of mathematics that are of immense importance, but to which we have nothing new to contribute in this article.

It is *possible* to justify uniform reflection for a theory such as *TB*. If one believes in Zermelo-Fraenkel set theory, for instance, and the basic satisfaction-biconditionals for this theory, then one can prove (and thus justify) the soundness of *TB*. The point is just that one *can* legitimately accept uniform reflection for *TB* without having a justification for it. This view seems in line with Feferman's view on the question of the epistemic warrant for reflection principles:

> The idea of an autonomous progression more nearly approximates the process of finding out what is implicit in accepting a basic system L_1, i.e., of what one ought to accept, *on the same fundamental grounds*, when one accepts L_1. (Feferman 1988, p. 131; our emphasis)

There are additional reasons why our warrant for reflection principles cannot be of any of the typical kinds. Reflection principles cannot be *deduced* from the theory upon which the reflection principle reflects. But it is also not possible to take reflection principles to be *axioms* in the sense that, for instance, the power set axiom is an axiom.

According to many, the power set axiom is justified by appeal to a form of *intuition*. We vaguely 'see' a model (the cumulative hierarchy of sets, generated in stages) and realize that the power set axiom is true in it. But the occurrence of the word 'true' in the previous sentence indicates that we do not want to say something similar for reflection principles. For it would mean motivating reflection principles by appeal to the notion of truth. Our position would thereby collapse into the Ketland-Shapiro position, and then the game would be lost. In the type-free setting, it seems clear that the Ketland-Shapiro line is not very plausible. Presumably the uniform reflection principle for our most encompassing theory (including our most encompassing truth theory) holds. So, according to the Ketland-Shapiro line, it should be *justified* on the basis of truth principles, and this justification should presumably take roughly the same form as the canonical justification the reflection principles of sub-theories (such as *TB*) of our overall theory. But for all too familiar Gödelian reasons, we have no truth principles from which we can justify the reflection principle for our most encompassing theory.

According to others, the power set axiom is 'analytic' in some full-blown, Gödel-like sense. But this does not seem to work for reflection principles. The required conceptual connection between provability *in a theory* and truth is just not available. The corresponding conceptual

connection between *informal* provability (not tied to a specific system) and truth *is* there, but that is an entirely different matter.

In any case, reflection principles are *too specific* (tied to a particular theory) to be candidates for being axioms in the true sense of the word, that is, basic principles.[19]

Nonetheless, there may be reason to think that *uniform* reflection principles derive from more basic principles. Kreisel once argued that we believe in the first-order induction scheme (for the language of arithmetic) *because* we believe in the second-order induction axiom (Kreisel 1967). A related train of thought would not go as far as this, but instead, say, argue that we believe in the first-order induction scheme because we believe in the *axiom*

$$\forall \phi \in \mathcal{L}_{PA} : (T\phi(0) \wedge T(\forall y.\ \phi(y) \rightarrow \phi(y + 1))) \rightarrow \forall x T\phi(x)$$

If this is plausible, then one might also say that we believe in all instances of the scheme RFN_S *because* we believe in the global reflection *axiom* GRP_S. Thus—or so the argument goes—global reflection *justifies* uniform reflection.

It is hard to evaluate this line of reasoning: what is the evidence for the 'because'? But even if it is found cogent, then the substance of what we have argued for in this article stands. What we have claimed for uniform reflection will then apply instead to global reflection. It will remain the case that the compositionality of truth is contained in disquotational principles for truth, modulo reflection principles; they will then just be global instead of uniform reflection principles.

7. Other roads from disquotation to compositionality

It is fair to say that the doctrine that disquotationalism cannot explain the compositional nature of truth counts as the *received* view. Nevertheless, we are not the only ones to dissent from it. There are some authors who have defended the view that there is a justificatory road from disquotational truth axioms to axioms that express the compositionality of truth. We will discuss three such views here—Field (2006), Halbach (2001) and Halbach (2002)—and explain how the viewpoint that we advocate differs from theirs.

Hartry Field (2006) has shown how compositional axioms can be derived from a specific schematic way of expressing the disquotational

[19] In a similar vein, Donald Martin (1998, p. 227) argues that determinacy principles are too specific to ever be candidates for being basic axioms.

viewpoint. In a nutshell, and simplifying greatly, his theory goes as follows. In Field's account, Tarski-biconditionals are expressed in the language of truth as $Tp \leftrightarrow p$, where p is a *schematic* letter. Then, aside from the usual rules for reasoning in languages with schematic letters, Field introduces a new rule for reasoning with schematic letters, which (roughly) says that if $\Phi(p)$ has been derived, then we are allowed to conclude $\forall \phi \in \mathcal{L} : \Phi(\phi)$, for some language $\mathcal{L}$. So, in particular, since in a schematic disquotational theory we can derive $\neg Tp \leftrightarrow T\neg p$, this rule of inference allows us to derive the compositionality of negation.

The admissibility of this new inference rule indicates that, in effect, a scheme of the form $Tp \leftrightarrow p$ can be read as a *sentence*,

$$\Pi p : Tp \leftrightarrow p,$$

which substitutionally quantifies over propositions of a certain kind. But it is fairly commonly held that substitutional quantification must be explained in terms of a compositional notion of truth. So Field's proposal is particularly vulnerable to Halbach's critique that 'a language embracing substitutional quantification can be considered as a notational variant of the language with the truth predicate satisfying the Tarskian clauses for truth' (Halbach 2002, §5).

In Halbach (2001), an account is given which bears some resemblance to the theory that we are proposing in the present article. Halbach starts by defining the notion of *truth-analyticity*: 'A is truth-analytic iff A is logically implied by the disquotation sentences' (Halbach 2001, p. 1962). Then he argues that RFN_{TB_0} expresses what is *analytical in the concept of truth*. Moreover, he claims that RFN_{TB_0}, rather than TB_0, should be taken to be the disquotationalist's truth theory:

> The formalization of the disquotationalist standpoint by a reflection scheme takes into account that the disquotationalist does not only claim the disquotation sentences, but that he also claims something about them, namely, that they govern the meaning of the truth predicate. (Halbach 2001, p. 1963)

And, as mentioned before, Halbach observes that within the context of Peano arithmetic, RFN_{TB} entails the typed compositional axioms for truth.

What are we to make of this? First of all, we observe that $PA + \mathrm{RFN}_{TB_0}$ is not a collection of truth-biconditionals. So it is hard to see how this is an expression of (rather than a redefinition of) disquotationalism. Second, it is not clear that RFN_{TB_0} can be seen

as a good formalization of the notion of truth-analyticity. Halbach writes:

> In general, I do not propose that axioms of a theory should be replaced by the uniform reflection principles for that theory. For the transition from a theory to a reflection principle for that theory requires an argument. In the case of disquotationalism this is provided by paying attention to the modal status of the disquotation sentences, i.e., by arguments for their analyticity. (Halbach 2001, p. 1963)

We argued in the previous section that if one is fully committed to a theory, then the transition from the theory to the reflection principle for it does not require an *argument*. But, aside from this, even if Halbach is right that truth-analyticity is the *justification* for RFN_{TB}, it does not follow that (and it does not seem correct that) RFN_{TB} is an acceptable *formalization* of truth-analyticity.

Halbach implicitly seems to take this point in his article 'Modalized Disquotationalism', which appeared somewhat later (Halbach 2002). In this article, the compositional axioms are derived from a theory that contains axioms and rules governing a *necessity predicate*, plus a disquotational scheme stating that the Tarski-biconditionals for the language of arithmetic are necessary. So this proposal can be seen as an alternative articulation of the idea that motivates the Halbach (2001) view. After all, nothing precludes interpreting the notion of necessity in Halbach (2002) as 'analytic in the Tarski-biconditionals'.[20] Nevertheless, this theory is also less than satisfactory. First, on the account under investigation it still cannot be said that the compositionality of truth is contained in disquotational axioms. Instead, it is a combined account of necessity and of truth that yields the compositionality of truth. Second, the proposed theory contains seemingly *ad hoc* restrictions on the principles of necessity. In particular, the proposed theory denies the principle that if a sentence necessarily holds, then it is true (Halbach, 2002, §2).[21]

[20] Perhaps this interpretation is suggested in an earlier article by Halbach. In Halbach (2000), he writes, '[O]ne might have the notion of provability from the disquotation sentences, that is, the notion of truth-analyticity, available without believing the soundness of the consequences. The distinctive feature of somebody who believes in the disquotation principle—like the disquotationalist—is his belief in the *soundness* of what I call truth-analyticity' (p. 169). So, 'Disquotationalism is *not* simply an axiomatization of truth but a theory about an axiomatic system of truth …[D]isquotationalism is not only a theory of truth but a theory of the modal status of the disquotation sentences as well' (p. 170).

[21] Of course, in the light of the Kaplan-Montague paradox for necessity predicates (Kaplan & Montague 1960), either the reflexivity axiom $\Box\phi \rightarrow \phi$ or the necessitation rule for $\Box$ has to be restricted in order to avoid contradiction.

8. Open questions

The conclusion that we have arrived at is that the concept of truth is simple because at bottom disquotational, but reflection is complicated. It has long been known that from a proof-theoretic point of view, reflection hierarchies are complicated and interesting.[22] But we have argued that the process of reflection is also *conceptually* complicated, and that it is intriguing from an epistemological point of view.

In this article we have explored the result of one or two reflective moments applied to a natural typed or type-free collection of truth-biconditionals. But, as we have mentioned, the full implicit content of a theory is not thereby revealed. Instead, it is given by the reflexive closure of the theory in the sense of Feferman. Thus it would seem important to have an informative description of the reflexive closure of *TB* and of *PTB*. Moreover, it is an interesting question whether the Friedman-Sheard theory (as first formulated in Friedman and Sheard 1987) can be obtained by reflection from a natural collection of Tarski-biconditionals.

9. Technical appendix

In this section we provide proofs of the main results outlined above. We begin by fixing notations and definitions referred to earlier.

Any theory of truth requires some background theory of syntax in which the basic syntactic operations corresponding to the manipulation of formulae can be formalized. This requires only a modicum of arithmetic, enough to define a Gödel coding of the prescribed language and simple operations on Gödel codes corresponding to the sentence-building operations and substitution. For this purpose we utilize the theory *EA* of *elementary arithmetic*:[23]

Definition 1. *EA is the theory of arithmetic comprising the basic axioms of PA, including exponentiation, and induction for* Δ_0*-formulae.*

[22] The *locus classicus* for this subject is Feferman (1962).

[23] It is worth noting that *EA* does not represent the *minimal* subsystem of arithmetic that can be used as a base theory for the results of the present paper. This accolade would likely go to Buss's theory of bounded arithmetic S^1_2, whose provably total functions are the polytime computable functions. Via an 'efficient' coding of $\mathcal{L}^+$ (see, for example, Hájek and Pudlák 1998, ch. 5), the relevant work, in particular the version of theorem 8 below, can be established with S^1_2 in place of *EA*. Moreover, it is known that S^1_2 permits a finite axiomatization (Ferreira and Ferreira 2013).

As with *PA*, we assume *EA* is formulated in the extended language $\mathcal{L}^+ = \mathcal{L}_T \cup \mathcal{L}_P$ comprising $\mathcal{L}_T$ as well as unary predicates T and F. Nevertheless the induction schema of both theories is restricted to the *truth-free* language (denoted $\mathcal{L}_{PA}$) in which neither predicate T nor F may occur. Thus for each theory it is only the rules and axioms of classical logic that apply to the extended language.

EA, although weak, is sufficiently strong to allow the formulation of the syntactic matters that are required for formalizing standard metatheoretic arguments of formal theories (sequences, Gödel coding, substitution, etc.). In particular, we can fix a Gödel coding $e \mapsto \ulcorner e \urcorner$ of expressions in the language $\mathcal{L}^+$ such that the functions

$$\begin{array}{ll} \wedge : (\ulcorner \phi \urcorner, \ulcorner \psi \urcorner) \mapsto \ulcorner \phi \wedge \psi \urcorner & \dot{\forall} : (\ulcorner v \urcorner, \ulcorner \phi \urcorner) \mapsto \ulcorner \forall v \phi \urcorner \\ \vee : (\ulcorner \phi \urcorner, \ulcorner \psi \urcorner) \mapsto \ulcorner \phi \vee \psi \urcorner & \dot{\exists} : (\ulcorner v \urcorner, \ulcorner \phi \urcorner) \mapsto \ulcorner \exists v \phi \urcorner \\ \neg : \ulcorner \phi \urcorner \mapsto \ulcorner \neg \phi \urcorner & \doteq : (\ulcorner s \urcorner, \ulcorner t \urcorner) \mapsto \ulcorner s = t \urcorner \\ \multicolumn{2}{l}{sub : (\ulcorner \phi \urcorner, \ulcorner v \urcorner, \ulcorner s \urcorner) \mapsto \ulcorner \phi(s/v) \urcorner} \end{array}$$

are provably total in *EA*.[24] Let $\bar{x}$ denote the xth numeral: $\bar{0} = 0$ and $\overline{x+1} = s(\bar{x})$. For particular applications of the substitution function, it is appropriate to introduce abbreviations: $\dot{T}(x)$ and $\dot{F}(x)$ denote the terms $sub(\ulcorner T(v) \urcorner, \ulcorner v \urcorner, x)$ and $sub(\ulcorner F(v) \urcorner, \ulcorner v \urcorner, x)$ respectively; $subn(\ulcorner \phi \urcorner, \ulcorner v \urcorner, x) = \ulcorner \phi(\bar{x}/v) \urcorner$; $\ulcorner \phi(\dot{v}) \urcorner = subn(\ulcorner \phi \urcorner, \ulcorner v, v \urcorner)$; and for $n > 1$,

$$\ulcorner \phi(\dot{v}_1, \ldots, \dot{v}_n) \urcorner := subn(\ulcorner \phi(\dot{v}_1, \ldots, \dot{v}_{n-1}) \urcorner, \ulcorner v_n \urcorner, v_n)$$

With the above operations, it is straightforward to define predicates $Form_{\mathcal{L}}(x)$, $Sent_{\mathcal{L}}(x)$ and $Term_{\mathcal{L}}(x)$ that represent (in *EA*) the property of being a code of a formula, sentence and term of $\mathcal{L}$. Finally, we also require a valuation function $^\circ : \ulcorner s \urcorner \mapsto s$ that, given the code of a term, returns the value of the term. This function is not provably total in *EA*, but is so in *PA*, which is where we make use of it.

With our formalization of syntax now fixed, we can precisely formulate the uniform reflection principle for a theory S and iterations thereof.

Definition 2. *Let S be a theory in the language $\mathcal{L}^+$ with an elementary decidable set of axioms, and let $Prov_S(x)$ be a standard provability predicate for S formalized within EA.* RFN_S *is the collection of*

[24] We adopt the usual convention of identifying the closed term $\ulcorner e \urcorner$ with its value in the standard model. This should not be confused with the valuation function $^\circ$ introduced below, which, given a number m, determines the term s such that $m = \ulcorner s \urcorner$ and returns the value of s.

formulae

$$\forall x_1 \ldots \forall x_n(Prov_S(\ulcorner \phi(\dot{x}_1, \ldots, \dot{x}_n)\urcorner) \rightarrow \phi)$$

for ϕ, *a formula of* $\mathcal{L}^+$ *with free variables among* $x_1, \ldots, x_n$. Let $S_1 = EA + \mathrm{RFN}_S$ and $S_{n+1} = EA + \mathrm{RFN}_{S_n}$.

The following result, originally due to Kreisel and Levy (1968), neatly demonstrates the connection between induction and reflection.[25]

Theorem 8. $EA_1 = PA + \mathrm{Ind}(\mathcal{L}^+)$

As the results in this paper make extensive use of the techniques used to prove theorem 8, we outline the proof.

Proof. That *PA* is a sub-theory of EA_1 requires proving that each instance of $\mathcal{L}^+$-induction is derivable from an instance of the reflection principle. Let $\phi(x)$ be a formula of $\mathcal{L}^+$ and let $\psi(x) = \phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(x + 1)) \rightarrow \phi(x)$. Within *EA*, it is possible to prove that for every n, a derivation in *EA* of $\psi(\bar{n})$ can be transformed into a derivation of $\psi(\overline{n+1})$. By Δ_0-induction, we deduce $EA \vdash \forall x Prov_S(\ulcorner \psi(\dot{x})\urcorner)$, and so $EA_1 \vdash \forall x \psi$. As $\forall x \psi$ is provably (in *EA*) equivalent to the axiom of induction for ϕ, we are done.

The converse argument requires more delicate proof-theoretic analysis. *EA* may not be finitely axiomatizable, but $I\Sigma_1 = EA + \mathrm{Ind}(\Sigma^0_1)$, the extension of *EA* by induction for Σ^0_1 formulae (from $\mathcal{L}_{PA}$), can be axiomatized as a single Σ^0_3 formula by coding the induction schema as a single instance using a Σ^0_1 partial truth predicate (see, for example, Hájek and Pudlák 1998, theorem 2.52). Hence, for every theorem $\phi \in \Sigma_n$ of $I\Sigma_1$ (say $n \geq 3$) there is a derivation of ϕ in a sequent calculus formulation of $I\Sigma_1$ with a cut rule, in which axioms involve only Σ^0_3 formulae. Standard cut elimination for first-order logic yields a derivation of ϕ with cuts-only axioms, and so the sub-formula property implies that the whole derivation consists solely of Σ^0_n formulae. This partial cut elimination argument is formalizable within $I\Sigma_1$. Within *PA* therefore, a partial truth predicate $Tr^*(x)$ can be defined such that $PA \vdash Tr^*(\ulcorner \phi(\dot{x})\urcorner) \leftrightarrow \phi(x)$ for each Σ^0_n formula ϕ, and

$$PA + \mathrm{Ind}(\mathcal{L}^+) \vdash \forall x(Sent_{\Sigma_n}(x) \wedge Prov_{EA}(x) \rightarrow Tr^*(x))$$

we deduce $PA + \mathrm{Ind}(\mathcal{L}^+) \vdash \mathrm{RFN}_{EA}$. □

[25] For a thorough discussion of this result and its philosophical significance, see Dean (2015).

9.1 Typed truth

Definition 3. *The theories TB_0 and UTB_0 are the extensions of EA by, respectively, the local and uniform truth-biconditionals for $\mathcal{L}_{PA}$. That is, TB_0 contains the axiom $\phi \leftrightarrow T(\ulcorner \phi \urcorner)$ for each closed $\mathcal{L}_{PA}$-formula ϕ, and UTB_0 contains the axiom $\phi(x_1, \ldots, x_n) \leftrightarrow T(\ulcorner \phi(\dot{x}_1, \ldots, \dot{x}_n) \urcorner)$ for each $n \geq 0$ and each formula ϕ with at most $x_1, \ldots, x_n$ free. By TB and UTB we denote the respective theory with induction extended to all formulae of the language.*[26]

Theorem 9. *UTB is a sub-theory of TB_1.*

Proof. For each closed formula $\forall x \phi(x)$ of $\mathcal{L}_{PA}$ we have, provably in *EA*,

$$TB_0 \vdash T(\ulcorner \phi(\bar{n}) \urcorner) \leftrightarrow \phi(\bar{n})$$

for every n, whence reflection implies $TB_1 \vdash \forall x(T\ulcorner \phi(\dot{x}) \urcorner \leftrightarrow \phi(x))$. □

Definition 4. *CT is the $\mathcal{L}_T$ theory extending PA by induction for all $\mathcal{L}_T$ formulae and the following axioms*:

(1) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (T(s_1 \doteq s_2) \leftrightarrow s_1^\circ = s_2^\circ)]$

(2) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_T}(\alpha_1 \dot{\wedge} \alpha_2) \rightarrow (T(\alpha_1 \dot{\wedge} \alpha_2) \leftrightarrow T\alpha_1 \wedge T\alpha_2)]$

(3) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_T}(\alpha_1 \dot{\wedge} \alpha_2) \rightarrow (T(\alpha_1 \dot{\vee} \alpha_2) \leftrightarrow T\alpha_1 \vee T\alpha_2)]$

(4) $\forall \alpha[Sent_{\mathcal{L}_T}(\alpha) \rightarrow (T(\dot{\neg}\alpha) \leftrightarrow \neg T\alpha)]$

(5) $\forall \alpha[Sent_{\mathcal{L}_T}(sub_x(\alpha, \bar{0})) \rightarrow (T(\dot{\forall} x\alpha) \leftrightarrow \forall x\, T(sub_x(\alpha, x))]$

(6) $\forall \alpha[Sent_{\mathcal{L}_T}(sub_x(\alpha, \bar{0})) \rightarrow (T(\dot{\exists} x\alpha) \leftrightarrow \exists x\, T(sub_x(\alpha, x)))]$

Halbach (2001, lemma 4.2) established:

Theorem 10. *CT and UTB_1 are identical theories.*[27]

Proof. We outline Halbach's argument, as it demonstrates how compositional axioms arise through reflection.

Arguing informally within TB_1, we observe that for all $\mathcal{L}$-sentences ϕ and ψ,

$$TB_0 \vdash T(\ulcorner \phi \wedge \psi \urcorner) \leftrightarrow T(\ulcorner \phi \urcorner) \wedge T(\ulcorner \psi \urcorner)$$

[26] In general, we append a subscript 'o' to theories to emphasize that induction is *not* extended; truth theories without subscripts will feature induction for the whole language.

[27] The theories *CT* and UTB_1 are denoted by $PA(S)$ and *AT* respectively in Halbach (2001).

An application of reflection therefore yields the compositional axiom for conjunction. A similar argument implies the compositional axiom for disjunction and atomic predicates. The quantifier axioms follow by an analogous argument though the theory UTB_1 is needed in place of TB_1. Thus all axioms of CT are derivable in UTB_1.

To deduce that the axioms of UTB_1 are derivable in CT, one uses a slight generalization of Kreisel and Lévy's argument as described in Halbach (2001). It is clear that UTB_0 is a sub-theory of CT. Crucially, there is a finite set of axioms of CT that suffice to derive the axioms of UTB_0, namely, the extension of $I\Sigma_1$ by the six axioms in definition 4. Utilizing partial cut elimination for this theory as well as extensions to the partial truth predicates that are available in CT, it follows that the schema RFN_{UTB_0} is derivable in CT. □

Combining this result with theorem 9 we conclude:

Theorem 11. *CT is a proper sub-theory of TB_2.*

Proof. That CT is a sub-theory of TB_2 follows from the previous two theorems. To conclude that the reflection principle for TB_1 is not derivable in CT we resort to known facts from the proof-theoretic analysis of CT.

Since induction for the whole language is already derivable in TB_1, it follows that within TB_2 transfinite induction through the ordinal ε_0 is derivable for every $\mathcal{L}_T$ formulae. It is well known (see, for example, Feferman 1991) that within CT, transfinite induction for $\mathcal{L}_T$ formulae is derivable only for ordinals strictly below ε_0.[28] Thus CT must be a proper sub-theory of TB_2. □

9.2 Type-free truth

Recall the language $\mathcal{L}_P$ that contains all terms and relations of $\mathcal{L}_T$, and

(1) for each atomic predicate A of $\mathcal{L}_T$, a fresh predicate symbol $\bar{A}$ in $\mathcal{L}_P$ of the same arity as A; each of A and $\bar{A}$ is referred to as the *dual* of the other,

(2) the connectives $\wedge$ and $\vee = \bar{\wedge}$, and quantifiers $\forall$ and $\exists = \bar{\forall}$. Similarly we call $\wedge$ and $\vee$ dual as well as $\forall$ and $\exists$.

We denote the dual of T by F. In this way $\mathcal{L}_P$ can be considered as the negation-free sub-language of $\mathcal{L}_{T,F}$ (the extension of $\mathcal{L}$ by two unary

[28] In contrast to PA_T, however, CT does prove transfinite induction to ordinals below ε_0 for formulae of $\mathcal{L}_{PA}$.

predicate symbols, T and F) in which F is identified with $\neg T$. Each formula ϕ of $\mathcal{L}_P$ has a dual $\bar{\phi}$ that results from recursively replacing each connective, quantifier and predicate symbol in ϕ by its dual.

Definition 5. *TFB_0 is the $\mathcal{L}_{T,F}$-theory extending EA by the truth and falsity biconditionals for all formulae in $\mathcal{L}_P$: the collection of axioms*

$$T(\ulcorner\phi\urcorner) \leftrightarrow \phi F(\ulcorner\bar{\phi}\urcorner) \leftrightarrow \phi$$

for sentences ϕ of $\mathcal{L}_P$.
$UTFB_0$ extends TFB_0 by the equivalences

$$T(\ulcorner\phi(\dot{v}_1, \ldots, \dot{v}_n)\urcorner) \leftrightarrow \phi F(\ulcorner\overline{\phi(\dot{v}_1, \ldots, \dot{v}_n)}\urcorner) \leftrightarrow \phi$$

for each formula ϕ, with at most $v_1, \ldots, v_n$ free. KF is the $\mathcal{L}_{T,F}$-theory extending PA by induction for all $\mathcal{L}_{T,F}$ formulae and the following twelve axioms.

(1) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (T(s_1 \doteq s_2) \leftrightarrow s_1^\circ = s_2^\circ)]$

(2) $\forall s_1, s_2[Term_{\mathcal{L}}(s_1) \wedge Term_{\mathcal{L}}(s_2) \rightarrow (F(s_1 \doteq s_2) \leftrightarrow s_1^\circ \neq s_2^\circ)]$

(3) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (T(\alpha_1 \wedge \alpha_2) \leftrightarrow T\alpha_1 \wedge T\alpha_2)]$

(4) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (T(\alpha_1 \vee \alpha_2) \leftrightarrow T\alpha_1 \vee T\alpha_2)]$

(5) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (F(\alpha_1 \vee \alpha_2) \leftrightarrow F\alpha_1 \wedge F\alpha_2)]$

(6) $\forall \alpha_1, \alpha_2[Sent_{\mathcal{L}_P}(\alpha_1 \wedge \alpha_2) \rightarrow (F(\alpha_1 \wedge \alpha_2) \leftrightarrow F\alpha_1 \vee F\alpha_2)]$

(7) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (T(\forall x\alpha) \leftrightarrow \forall x\, T(sub_x(\alpha, x)))]$

(8) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (T(\exists x\alpha) \leftrightarrow \exists x\, T(sub_x(\alpha, x)))]$

(9) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (F(\forall x\alpha) \leftrightarrow \exists x\, F(sub_x(\alpha, x)))]$

(10) $\forall \alpha[Sent_{\mathcal{L}_P}(sub_x(\alpha, \bar{0})) \rightarrow (F(\exists x\alpha) \leftrightarrow \forall x\, F(sub_x(\alpha, x)))]$

(11) $\forall s[Term_{\mathcal{L}}(s) \rightarrow (T(\dot{T}\, s) \leftrightarrow T(s^\circ)) \wedge (T(\dot{F}\, s) \leftrightarrow F(s^\circ))$

(12) $\forall s[Term_{\mathcal{L}}(s) \rightarrow (F(\dot{T}\, s) \leftrightarrow F(s^\circ)) \wedge (F(\dot{F}\, s) \leftrightarrow T(s^\circ))$

The reader may be worried by the fact that the sentential quantifiers in axioms 3–10 of *KF* are restricted to the language $\mathcal{L}_P$, which does not contain the negation symbol, and not to either the languages $\mathcal{L}_{T,F}$ or $\mathcal{L}_T$, as is most common in the literature. The reason is that it is only when viewed as a theory of truth for the language $\mathcal{L}_P$ that the axioms of *KF* can be described as compositional with respect to *all* the

connectives—*KF* (in any of its incarnations) is inconsistent with the compositional axiom for negation, $Sent_{\mathcal{L}}(\alpha) \rightarrow T(\dot{\neg}\alpha) \leftrightarrow \neg T(\alpha)$ (where $\mathcal{L}$ is any one of $\mathcal{L}_T$, $\mathcal{L}_{T,F}$ or $\mathcal{L}_P$). In the languages $\mathcal{L}_T$ and $\mathcal{L}_{T,F}$, occurrences of the negation symbol under the truth predicate are afforded a different interpretation from those outside, being either unspecificied (as in Halbach 2001, for instance) or (as in Feferman 1991) interpreted as behaving as our duality operator, exchanging the roles of 'true' and 'false', etc.

We begin by noting that, taken in isolation, the truth- and falsity-biconditionals are model-theoretically weak.

Theorem 12. *$UTFB_0$, and hence also TFB_0, is semantically conservative over EA.*

Proof. We present here the proof that every model of *EA* can be extended to a model of TFB_0 via a variant of Kripke's fixed-point construction; a proof of the same result for $UTFB_0$ follows the same lines and is left to the reader.

For an $\mathcal{L}$-structure $\mathfrak{M}$, R, S subsets of the domain of $\mathfrak{M}$ (i.e. $R, S \subseteq |\mathfrak{M}|$), and χ a formula of $\mathcal{L}_{T,F}$, let $\langle \mathfrak{M}, R, S\rangle \models \chi$ be defined according to the usual Tarskian satisfaction rules for classical logic with R interpreting the extension of the truth predicate T and S interpreting F. Notice that for χ in $\mathcal{L}_P$, $R_0 \subseteq S_0 \subseteq |\mathfrak{M}|$ and $R_1 \subseteq S_1 \subseteq |\mathfrak{M}|$ we have

$$\langle \mathfrak{M}, R_0, R_1\rangle \models \chi \text{ implies } \langle \mathfrak{M}, S_0, S_1\rangle \models \chi$$

Let $\bar{R} = \{\ulcorner \bar{\chi} \urcorner | \chi \in R\}$. We now define a sequence of sets R_α indexed by countable ordinals:

$$R_0 = \varnothing$$
$$R_{\alpha+1} = \{\ulcorner \psi \urcorner | \psi \text{ is a sentence of } \mathcal{L}_P \text{ and } \langle \mathfrak{M}, R_\alpha, \bar{R}_\alpha\rangle \models \psi\}$$
$$R_\lambda = \bigcup_{\alpha < \lambda} R_\alpha \text{ (for limit } \lambda)$$

Since $R_\alpha \subseteq R_\beta$ for every $\alpha \leq \beta$, by cardinality considerations there exists κ for which $R_\kappa = R_{\kappa+1}$. Then for a sentence ψ of $\mathcal{L}_P$,

$$\begin{aligned} \langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa\rangle \models \psi &\Leftrightarrow \psi \in R_{\kappa+1} \\ &\Leftrightarrow \langle \mathfrak{M}, R_{\kappa+1}, \bar{R}_{\kappa+1}\rangle \models T(\ulcorner \psi \urcorner) \\ &\Leftrightarrow \langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa\rangle \models T(\ulcorner \psi \urcorner) \end{aligned}$$

and similarly for the falsity predicate. Therefore $\langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa \rangle \models TFB_0$.

Cieśliński establishes that induction for the extended language does not increase the proof-theoretic strength of the local positive truth-biconditionals, that is, *PTB* conservatively extends *PA* (Cieśliński 2011). The proof naturally extends to the case of a falsity predicate:

Theorem 13. *$TFB = TFB_0 + \mathrm{Ind}(\mathcal{L}_{T,F})$, and hence also PTB, conservatively extends PA.*

We present a simplified version of Cieśliński's proof below.[29] Before we proceed, observe that the proof of the theorem 12 does not suffice for this result, as the constructed structure $\langle \mathfrak{M}, R_\kappa, \bar{R}_\kappa \rangle$ need not satisfy the extended induction schema $\mathrm{Ind}(\mathcal{L}_{T,F})$. Instead we prove that for each finite set S of $\mathcal{L}_P$ sentences there exists an $\mathcal{L}_{PA}$-definable subset R of $|\mathfrak{M}|$ *and* $\langle \mathfrak{M}, R, \bar{R} \rangle$ satisfies the truth and falsity biconditionals for all formulae in S. The first requirement ensures that the structure $\langle \mathfrak{M}, R, \bar{R} \rangle$ satisfies the induction schema for formulae of $\mathcal{L}_{T,F}$ whenever $\mathfrak{M}$ is a model of $\mathcal{L}_{PA}$-induction.

Proof. Let $\mathfrak{M}$ be an arbitrary model of *PA*. We prove that if ϕ is derivable in $TFB_0 + \mathrm{Ind}(\mathcal{L}_{T,F})$ and ϕ is a formula of $\mathcal{L}_{PA}$, then $\mathfrak{M} \models \phi$.

Suppose ϕ in the language $\mathcal{L}_{T,F}$ is derivable in *TFB*. Let S denote the (finitely many) $\mathcal{L}_P$-formulae such that ϕ is derivable from $EA + \mathrm{Ind}(\mathcal{L}_{T,F})$ extended by the axioms

$$T(\ulcorner \psi \urcorner) \leftrightarrow \psi \qquad\qquad F(\ulcorner \psi \urcorner) \leftrightarrow \psi$$

for each $\psi \in S$. We inductively define

$$R_0 := \varnothing \; R_{n+1} := \{\ulcorner \psi \urcorner \mid \psi \in S \wedge \langle \mathfrak{M}, R_n, \bar{R}_n \rangle \models \psi\}$$

Since S contains only formulae of $\mathcal{L}_P$ we have $R_n \subseteq R_{n+1}$ for every n. As S is finite, there is some k (indeed $k = |S|$ suffices) such that $R_k = R_{k+1}$. Thus we deduce that $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle$ satisfies $\psi \leftrightarrow T(\ulcorner \psi \urcorner)$ and $\psi \leftrightarrow F(\ulcorner \psi \urcorner)$ for each $\psi \in S$, as in the proof of the previous theorem. We claim also $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \mathrm{Ind}(\mathcal{L}_{T,F})$.

Define formulae $T_i(x)$ and $F_i(x)$ as

$$T_0(x) \leftrightarrow \bot \qquad\qquad F_i(x) \leftrightarrow \bot$$

[29] It is worth remarking that unlike the proof presented here, Cieslinski's own argument implies a more general result than stated in theorem 13, namely, that every recursively saturated model of PA can be extended to a model of PTP (and indeed also PTB).

$$T_{i+1}(x) \leftrightarrow \bigvee_{\psi \in S}(x = \ulcorner \psi \urcorner \wedge \psi^i) \quad F_{i+1}(x) \leftrightarrow \bigvee_{\psi \in S}(x = \ulcorner \psi \urcorner \wedge \psi^i)$$

where ψ^i denotes the result of replacing in ψ the predicates T and F by T_i and F_i respectively.

$$R_k = R_{k+1} = \{\ulcorner \chi \urcorner | \chi \in S \wedge \mathfrak{M} \models \chi^k\}$$

and $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \mathrm{Ind}(\mathcal{L}_{T,F})$. Thus $\langle \mathfrak{M}, R_k, \bar{R}_k \rangle \models \phi$ and so $\mathfrak{M} \models \phi^k$. Since $\phi^k = \phi$ if ϕ is in the language of $\mathcal{L}_{PA}$, we are done. □

In contrast to the case of theorem 12, in the above theorem *TFB* cannot be replaced by *UTFB*. Since *PUTB* is interpretable in *UTFB*, it follows that *KF* is also interpretable in *UTFB* (see below). The proof above breaks down in this case, because choosing a finite collection of *uniform* biconditionals does not ensure a fixed point is reached within a finite number of steps. Consider, for example, a formula $\phi(x)$ with only x free (constructed via diagonalization) such that

$$\phi(x) \leftrightarrow x = 0 \vee (x > 0 \wedge T(\ulcorner \phi(\dot{x} - 1) \urcorner))$$

If we fix the case that $\mathfrak{M} = N$ is the standard model of arithmetic, it is natural to choose $S = \{\ulcorner \phi(\bar{n}) \urcorner < \omega\}$ and define the sets $R_0 = \varnothing$, R_1, $R_2, \ldots$ as before. In this case, $\phi(\bar{n}) \in R_{n+1} \backslash R_n$ for every n, so the first structure in the hierarchy satisfying the biconditional $\phi(x) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner)$ will be the limit structure $\langle \mathfrak{M}, \bigcup_{k<\omega} R_k, \bigcup_{k<\omega} \bar{R}_k \rangle$. In general, we cannot expect the fixed point to be definable.

We now turn our attention to extensions of the basic type-free theories of truth by reflection principles.

Theorem 14 (Halbach 2009, theorem 5.1). *KF is directly interpretable in PUTB.*

The interpretation is a trick of diagonalization: one constructs, via the diagonalization lemma, a formula $\phi(x)$ such that

$$\phi(x \vee y) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner) \vee T(\ulcorner \phi(\dot{y}) \urcorner)$$
$$\phi(\forall v x) \leftrightarrow \forall y\, T(\ulcorner \phi(subn(\dot{x}, \dot{v}, \dot{y})) \urcorner)$$
$$\phi(\underset{\cdot}{T}(x)) \leftrightarrow T(\ulcorner \phi(\dot{x}^{\circ}) \urcorner)$$
$$\ldots$$

Observe that ϕ is positive in T, so one has $PUTB \vdash \phi(x) \leftrightarrow T(\ulcorner \phi(\dot{x}) \urcorner)$, and hence the formula ϕ may serve as the interpretation of the *KF*-truth predicate.

Further reflection on the positive truth-biconditionals does, in a sense, yield the *KF* axioms:

Theorem 15 (Halbach 2001, theorem 5.2). *There are recursive functions f_T and f_F such that the translation that maps $T(s)$ to $T(f_T(s))$ and $F(s)$ to $F(f_F(s))$ is a direct interpretation of KF in $PUTB_1$.*

As a corollary, the same interpretation yields an embedding of *KF* into PTB_2. Since *KF* is formulated in the language $\mathcal{L}_{T,F}$ (both here and in Halbach 2001), it is not a sub-theory of PTB_n for any n. The missing axioms for the falsity predicate can be obtained by starting instead from *TFB*.

Theorem 16. *$KF = UTFB_1$ and KF is a sub-theory of TFB_2.*

Proof. We begin by showing that $KF = UTFB_1$. Both directions follow similar arguments as for theorem 10. As before, all the basic compositional axioms of *KF* are derivable from particular instances of reflection over $UTFB_0$. To deduce the final two axioms we observe that $T(\ulcorner Ts \urcorner) \leftrightarrow T(s)$, and hence,

$$(2) \quad T(\dot{T}(\ulcorner s \urcorner)) \leftrightarrow T(\ulcorner s \urcorner^{\circ})$$

is derivable in *TB* for each term s.[30] But then $UTB_1 \vdash \forall x(Term_{\mathcal{L}}(x) \rightarrow (T(\dot{T}\, x) \leftrightarrow T(x^{\circ})))$ as required, and similarly for the other variations. The proof that $UTFB_1$ is a sub-theory of *KF* is analogous to theorems 8 and 11, and we refer the reader interested in the differences to Halbach (2001, lemma 6.5).[31] As in the typed case, $UTFB_1$ is a sub-theory of TFB_2, whence the second part is established.

The inclusion $KF \subseteq TFB_2$ turns out to be strict:

Theorem 17. *KF is a proper sub-theory of TFB_2.*

[30] Since $^{\circ}$ is not provably total in *EA*, the formula represented in (2) is not, strictly speaking, derivable in *TB*. To be fully precise (and to suffice for the present argument), there exists a formula ϕ such that $PA \vdash \phi(\ulcorner s \urcorner, s)$ for every term s and $PA \vdash \forall x \forall y(Term_{\mathcal{L}}(x) \wedge \phi(x, y) \rightarrow (T(\dot{T}\, x) \leftrightarrow T(y)))$. As $^{\circ}$ is not, strictly speaking, a term in *PA*, this will be the form of *KF* axioms anyway.

[31] Halbach's argument concerns the theory $PUTB_1$ (denoted therein AT^{+}) in place of $UTFB_1$ and an axiomatization of *KF* with only a single truth predicate, but the argument is identical.

Proof. The proof of theorem 14 establishes an $\mathcal{L}$-conservative interpretation of *KF* into the theory *PUTB*, proving that the two theories have the same truth-free consequences. As this latter theory is a sub-theory of TFB_1, it follows that *KF* is $\mathcal{L}$-conservatively interpretable in TFB_1. Since the consistency statement for TFB_1 is derivable in TFB_2, the result follows.[32]

References

Baker, Alan 2005: 'Are There Genuine Mathematical Explanations of Physical Phenomena?' *Mind*, 114, pp. 223–38.

Burge, Tyler 1993: 'Content Preservation'. *Philosophical Review*, 102, pp. 457–88.

—— 1998: 'Computer Proof, Apriori Knowledge, and Other Minds'. *Philosophical Perspectives, 12: Language, Mind, and Ontology*, pp. 1–37.

—— 2003: 'Perceptual Entitlement'. *Philosophy and Phenomenological Research*, 67, pp. 503–48.

—— 2007: 'Self and Self-Understanding'. In Burge 2013, pp. 187–226.

—— 2013: *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection. Philosophical Essays*, Volume 3. Oxford: Oxford University Press.

Burgess, John P. 1986: 'The Truth is Never Simple'. *Journal of Symbolic Logic*, 51, pp. 663–81.

Cieśliński, Cezary 2010: 'Truth, Conservativeness, and Provability'. *Mind*, 119, pp. 409–22.

—— 2011: 'T-equivalences for Positive Sentences'. *Review of Symbolic Logic*, 4(2), pp. 319–25.

Davidson, Donald 1967: 'Truth and Meaning'. *In his Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 1984.

Dean, Walter 2015: 'Arithmetical Reflection and the Provability of Soundness'. *Philosophia Mathematica*, 23(1), pp. 31–64.

Feferman, Solomon 1962: 'Transfinite Recursive Progressions of Axiomatic Theories'. *Journal of Symbolic Logic*, 27(3), pp. 259–316.

—— 1988: 'Turing in the Land of O(z)'. In Rolf Herken (ed.), *The Universal Turing Machine: A Half-Century Survey*, pp. 113–47. Hamburg: Kammerer & Univerzagt.

[32] The authors wish to thank Volker Halbach, Jeff Ketland, Philip Welch, Marianna Antonutti, Walter Dean and Carlo Nicolai for their valuable comments and suggestions for improvement. The second author was supported by Arts and Humanities Research Council UK grant no. AH/H039791/1.

——— 1991: 'Reflecting on Incompleteness'. *Journal of Symbolic Logic*, 56, pp. 1–49.
Ferreira, Fernando, and Gilda Ferreira 2013: 'Interpretability in Robinson's Q'. *Bulletin of Symbolic Logic*, 19(3), pp. 289–317.
Field, Hartry 1999: 'Deflating the Conservativeness Argument'. *Journal of Philosophy*, 99, pp. 534–40.
——— 2006: 'Compositional Principles vs. Schematic Reasoning'. *Monist*, 89, pp. 9–27.
Franzén, Torkel 2004: *Inexhaustibility: A Non-Exhaustive Treatment*. Wellesley, MA: A. K. Peters.
Friedman, Harvey, and Michael Sheard 1987: 'An Axiomatic Approach to Self-Referential Truth'. *Annals of Pure and Applied Logic*, 33, pp. 1–21.
Hájek, Petr, and Pavel Pudlák 1998: *Metamathematics of First-Order Arithmetic*. Berlin: Springer-Verlag.
Halbach, Volker 2000: 'Disquotationalism Fortified'. In André Chapuis and Anil Gupta (eds.), *Circularity, Definition, and Truth*. New Delhi: Munshiram Manohararlal Publishers.
——— 2001: 'Disquotational Truth and Analyticity'. *Journal of Symbolic Logic*, 66, pp. 1959–73.
——— 2002: 'Modalized Disquotationalism'. In Volker Halbach and Leon Horsten (eds.), *Principles of Truth*, pp. 75–101. Frankfurt am Main: Dr. Hänsel-Hohenhausen.
——— 2009: 'Reducing Compositional to Disquotational Truth'. *Review of Symbolic Logic*, 2, pp. 786–98.
——— 2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.
Horsten, Leon 2011: *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.
Horwich, Paul 1998: *Truth*, 2nd edn. Oxford: Clarendon Press.
Kaplan, David, and Richard Montague, 1960. 'A Paradox Regained'. *Notre Dame Journal of Formal Logic*, 1, pp. 79–90.
Ketland, Jeffrey 1999: 'Deflationism and Tarski's Paradise'. *Mind*, 108, pp. 70–94.
——— 2005: 'Deflationism and the Gödel Phenomena: Reply to Tennant'. *Mind*, 114, pp. 76–88.
——— 2010: 'Truth, Conservativeness, and Provability: Reply to Cieśliński'. *Mind*, 119, pp. 423–36.
Kotlarski, H., S. Krajewski ,and A. H. Lachlan 1981: 'Construction of Satisfaction Classes for Nonstandard Models'. *Canadian Mathematical Bulletin*, 24, pp. 283–93.

Kreisel, Georg 1967: 'Informal Rigour and Completeness Proofs'. In Imre Lakatos (ed.), *Problems in the Philosophy of Mathematics*. pp. 138–86. Amsterdam: North-Holland.

——1970: 'Principles of Proof and Ordinals Implicit in Given Concepts'. In A. Kino, J. Myhill, and R. E. Vesley (eds.), *Intuitionism and Proof Theory*. pp. 489–516. Amsterdam: North-Holland.

Kreisel, Georg and A. Lévy 1968: 'Reflection Principles and their Use for Establishing the Complexity of Axiomatic Systems'. *Mathematical Logic Quarterly*, 14(7-12), pp. 97–142.

Lachlan, A. H. 1981: 'Full Satisfaction Classes and Recursive Saturation'. *Canadian Mathematical Bulletin*, 24, pp. 295–7.

Martin, D. A. 1998: 'Mathematical Evidence'. In H. G. Dales and G. Oliveri (eds.), *Truth in Mathematics*, pp. 215–31. Oxford: Oxford University Press.

McGee, Vann 1992: 'Maximal Consistent Sets of Tarski's Schema (T)'. *Journal of Philosophical Logic*, 21, pp. 235–41.

Quine, W. V. 1970: *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Soames, Scott 2003: *Philosophical Analysis in the Twentieth Century, vol. 2: The Age of Meaning*. Princeton, NJ: Princeton University Press.

Shapiro, Stewart 1998: 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy*, 95, pp. 493–521.

Strollo, Andrea 2013: 'Deflationism and the Invisible Power of Truth'. *Dialectica*, 67, pp. 521–43.

Tarski, Alfred 1935: 'The Concept of Truth in Formalized Languages'. In Tarski 1983, pp. 152–278.

——1983: *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* rev. edn. Translated by J. H. Woodger. Indianapolis: Hackett.

——1986: 'What are Logical Notions?' Edited with an introduction by John Corcoran in *History and Philosophy of Logic*, 7, pp. 143–54.

Tennant, Neil 2002: 'Deflationism and the Gödel Phenomena'. *Mind*, 111, pp. 551–82.

——2005: 'Deflationism and the Gödel Phenomena: Reply to Ketland'. *Mind*, 114, pp. 89–96.

——2010: 'Deflationism and the Gödel Phenomena: Reply to Cieśliński'. *Mind*, 119, pp. 438–50.

Williams, Michael 1988: 'Epistemological Realism and the Basis of Scepticism'. *Mind*, 97, pp. 415–39.

