Is deflationism compatible with compositional and Tarskian truth theories?

Lavinia Picollo* and Thomas Schindler†

1 Introduction

For a number of reasons, deflationists about truth favour a formal treatment of the notion. What requirements deflationary formal theories of truth must satisfy is, thus, an important issue for deflationism. It is widely believed that compositional and Tarskian theories convey substantial concepts of truth or are otherwise unacceptable for the deflationist. Call this claim the 'incompatibility thesis'. Since compositional and Tarskian theories are often seen as superior to purely disquotational theories, the incompatibility thesis, if true, would provide support for substantial theories of truth over their deflationary rivals. Assessing whether the arguments for the incompatibility thesis are correct is therefore of great philosophical importance.

Here is the plan of the paper. After some preliminaries (§2), we will rehearse six arguments for the incompatibility thesis from the literature (§3). We contend that most of these arguments issue from an overly narrow understanding of what role formal theories of truth are supposed to play. In §4, we introduce an important but often overlooked distinction between theories that are intended for a *descriptive* purpose (roughly, a theory that provides a faithful account of the basic usage of 'true') and those that are intended for a *logical* purpose (roughly, a theory that characterises the correctness of inferences involving 'true').

The notion of a logical purpose raises the question what the role of 'true' exactly consists in, and what truth principles are needed to carry it out. Drawing on earlier work [27], we suggest (§5) that this role is best understood as enabling us to mimic sentential and predicate quantification within a first-order framework, and extract a criterion of functionality from that. However, not any theory that allows the truth predicate to fulfil its function might be acceptable to a deflationist: among other things, such a theory must not convey a substantial notion of truth. What this is supposed to mean is of course a controversial issue. We will not be able to provide an absolute criterion of substantiality, though we will propose (§6) a relative one: under certain circumstances, adding

*University College London
†University of Bristol

certain truth-theoretic principles to a deflationary theory will not inflate the notion of truth. In §7 we will defend this criterion against a popular objection.

In concluding this paper (§8), we will survey a variety of formal truth theories and assess them in light of our criteria. It will be seen that a number of compositional and Tarskian truth theories are, plausibly, acceptable from a deflationary point of view and, therefore, do not encapsulate a substantial notion of truth. We conclude that the incompatibility thesis is false. Interestingly, our account also suggests that some popular compositional truth theories on the market are in fact not acceptable from a deflationary point of view. As we will argue, this does not constitute an embarrassment for deflationism, as there are good reasons to reject these theories on independent grounds.

2 Deflationism and the orthodoxy

The variant of deflationism that will be the focus of this paper consists of two fundamental claims. Some of its proponents are Field, Horsten, and Horwich, although their views might differ from each other in other, more satellite aspects.

The first core thesis of deflationism is that 'true', as it is deployed in theoretical contexts, is a primitive term governed by some form of equivalence between each truth ascription and the sentence or proposition itself to which truth is attributed to, i.e. by a so-called *transparency principle*. We will refer to this as the 'equivalence thesis'. This thesis is taken to suggest that there is no need or possibility of further conceptual analysis, no point in the search for an explicit definition of truth in terms of simpler, fundamentally more basic concepts – i.e. a *substantial account*. This is what distinguishes (this version of) deflationism from robust or substantive approaches to truth, such as the correspondence and the coherence theories, according to which there is a hidden nature of truth to uncover by means of an explicit definition, in which truth is analysed in terms of simpler concepts. Thus, deflationists sometimes claim that truth is not an ordinary or substantive property.

The second fundamental thesis of deflationism is that the sole reason for having a truth predicate in natural language is that it plays an indispensable logico-linguistic role. We will refer to this claim as the 'logico-linguistic function thesis'. For instance, the truth predicate allows us to endorse a single statement without explicitly articulating it, as in 'Goldbach's conjecture is true', or several, even infinitely many statements at once, as in 'All theorems of arithmetic are true'. It is this second thesis that distinguishes modern deflationism from its predecessor, the redundancy theory of truth. While in sentences such as "Snow is white' is true' the truth predicate is easily eliminable and, therefore, dispensable, this is not so in the case of the two examples given above. For we may not know what Goldbach's conjecture is or what the theorems of arithmetic are. And in the latter case, even if we did, there are too many of them to assert them one by one.

There are several reasons why the deflationary account of truth motivates a formal treatment of the notion. Some authors outright assert that truth is a primitive undefinable notion that must be axiomatised (cf. Halbach & Horsten [16]). Moreover, the so-called transparency principle that, according to the equivalence thesis, governs the truth predicate is simple, schematic, and reminiscent of those governing logical vocabulary. In addition, despite its simplicity, transparency is riddled with paradoxes when unrestricted and formulated over sufficiently strong logics and base theories. To successfully avoid contradictions, precise formulations are needed. Finally, and perhaps most importantly, the study of the logico-linguistic function which deflationists – and many non-deflationists as well – attribute to the truth predicate, based on the inferential behaviour of truth, obviously demands a formal treatment.

Indeed, recent years have seen a proliferation of formal truth theories, both in connection with and independently of deflationism. There has been much subsequent discussion about which formal properties a deflationary truth theory can and should have. Most agree that deflationists should opt for axiomatic systems, which thus will be the focus of this paper. Although we don't directly address semantic theories, some of the arguments below can be applied equally to them.

Axiomatic truth theories consist of axioms for truth formulated over a base theory that contains a sufficient amount of syntax to provide the specific objects we will ascribe truth to, the truth bearers, which, as is customary, we take to be sentences – or numbers that code sentences.[1] Let $\mathscr{L}$ be a first-order language, the language of the base theory, and let $\mathscr{L}_{\mathrm{T}}$ extend $\mathscr{L}$ with a monadic predicate T, for truth. We assume $\mathscr{L}$ contains enough vocabulary to express certain syntactic properties, relations, and functions of expressions of $\mathscr{L}_{\mathrm{T}}$ to be specified, and a quote name $\ulcorner\varphi\urcorner$ for each formula φ of $\mathscr{L}_{\mathrm{T}}$. Let Σ, the base theory, be a recursively axiomatised system formulated in $\mathscr{L}_{\mathrm{T}}$ containing a syntax theory for $\mathscr{L}_{\mathrm{T}}$ itself, which we assume is strong enough to relatively interpret first-order Peano arithmetic. For simplicity, we assume $\mathscr{L}$ has a term for every object in the domain of its intended interpretation and that Σ proves this, although all of our claims can be easily generalised if a satisfaction predicate is adopted instead. We also assume that only logical or syntactic principles containing T are derivable in Σ, but no truth principles. An axiomatic truth system Γ is then a recursive extension of Σ with axioms governing T.

What formal theories of truth can and should deflationists endorse? What truth axioms can and should a theory Γ consist of? The orthodoxy dictates that Γ should extend the base theory Σ only with a transparency principle. These are given by instances of so-called principles of (local) disquotation, that is, either the following schema:

(T-schema) $\qquad\qquad \mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$

[1] Should propositions be preferable to sentences, one could understand our truth predicate as applying not directly to the sentences but to what these sentences express.

or the inference rules

(T-Intro) $\varphi \vdash \mathrm{T}\ulcorner\varphi\urcorner$

(T-Elim) $\mathrm{T}\ulcorner\varphi\urcorner \vdash \varphi$

possibly restricted to a class Δ of sentences of $\mathscr{L}_{\mathrm{T}}$, which may but need not necessarily coincide with $\mathscr{L}_{\mathrm{T}}$.

Some philosophers have claimed that the deflationist's truth axioms should consist of *all* instances of disquotation for sentences of $\mathscr{L}_{\mathrm{T}}$, including those that contain the truth predicate. Due to the semantic paradoxes, this would preclude the use of classical logic and force the adoption of weaker systems instead, adding yet another entry to the long list of restrictions imposed on deflationary theories. In [26] we have given some reasons for believing that this restriction cuts too deep. We will not rehearse these arguments here, but simply assume that deflationists can adhere to classical logic. Horwich, one of the most vocal deflationists, clearly shares our view on this matter.

As anticipated in the introduction, it is usually maintained that compositional truth theories should be excluded from the deflationary picture. These theories get their name from their axioms, some of which are not instances of disquotation but compositional principles, such as

$(\mathrm{T}\wedge\restriction\Delta)\quad \forall x\forall y\,(\mathrm{Sent}_\Delta(x)\wedge\mathrm{Sent}_\Delta(y)\wedge\mathrm{Sent}_\Delta(x\dot{\wedge}y)\rightarrow(\mathrm{T}x\dot{\wedge}y\leftrightarrow\mathrm{T}x\wedge\mathrm{T}y))$

where $\mathrm{Sent}_\Delta(x)$ is a predicate that holds only of sentences in Δ and $\dot{\wedge}$ is a symbol for the function that maps every pair of formulae of $\mathscr{L}_{\mathrm{T}}$ to their conjunction (and similarly for the other logical connectives). $(\mathrm{T}\wedge\restriction\Delta)$ states that if x, y, and their conjunction belong to Δ, then the conjunction is true just in case both conjuncts are. Similar principles can be given for the other logical connectives and the quantifiers.

The orthodox view also maintains that no Tarskian truth theory shall be endorsed by a deflationist. These theories extend Σ with an axiom of the form

$(\mathrm{T}\restriction\Delta)\qquad \forall x\,(\mathrm{T}x\leftrightarrow\Phi(x))$

where $\Phi(x)$ holds only of sentences in Δ and T occurs in $\Phi(x)$ only applied to expressions of less complexity than x – sometimes considered to be a recursive (or explicit, if T doesn't occur in Φ at all) definition of T. We will occasionally refer to principles of the form $\mathrm{T}\restriction\Delta$ as 'Tarskian principles', or, if intended as definitions, as 'Tarskian definitions'.

Next we will consider and discuss a series or arguments in favour of the orthodoxy, and show that, at best, they have limited reach.

3 Arguments for the incompatibility thesis

In this section we will rehearse six arguments that have been given in favour of the incompatibility thesis.

Argument 1. A reason often given for restricting the deflationist's truth axioms to locally disquotational principles – of which Horwich [20] is perhaps the most vocal promoter but many others have echoed him – stems from the equivalence thesis. According to the latter, the only basic facts about truth from a deflationist viewpoint are instances of transparency; they are "the whole truth about truth" (Stoljar & Damnjanovic [37]). Thus, many have concluded, the axioms of a deflationary formal truth theory should consist exclusively of these *basic* principles, and every other fact about truth should be explained by – i.e. follow from – them. In support of this conclusion, consider the following remark by Horwich [21, p. 76]: "the minimalist thesis is that the *basic* facts (i.e. the axioms of the theory that explains *every* other fact about truth) will all be instances of the [equivalence] schema".

Argument 2. A related argument often wielded against compositional and Tarskian truth theories *qua definitions* is also based on the equivalence thesis, which suggests that a definition of 'true' is *neither necessary nor possible*:

> For "true" is a primitive term; so the only interesting account that can be given of its meaning is one that identifies which underlying property of the word (i.e. which aspect of our use of it) is responsible for its possessing that meaning. In particular, our truth predicate means what it does [...] in virtue of our underived commitment to the equivalence schema. (Horwich [21, pp. 75-76])

Thus, even if extensionally adequate, compositional and Tarskian truth theories cannot provide *real* definitions of truth.

Argument 3. Another argument commonly offered against the compatibility between deflationism and Tarskian truth theories stems from the logico-linguistic function thesis. If our truth predicate could be given by a Tarskian definition, then the language would already have the resources to formulate a predicate satisfying the relevant transparency principles. In this case, truth would be eliminable via the definiens, and thus the truth predicate would be *dispensable*. But according to the logico-linguistic function thesis, 'true' plays an indispensable role in (theoretically informed) natural language. Thus, Halbach & Horsten [16, p. 204] write: "definable notions of truth are not of primary interest to the deflationist because they are always just notions of truth for at best a part of our 'real' language."

Argument 4. Yet another reason given against compositional and Tarskian truth theories is that they *only work for simple, formal languages*. While Tarskian theories may be able to explain how the truth conditions of sentences of certain formal languages depend on the referents of their parts, it is not clear how they could deal with sentences of a natural language: "nobody has been able to show, for sentences involving 'that'-clauses, probabilistic locutions, attributive adjectives, or mass terms, how their truth could be explained by as a consequence of the referents of their parts." (Horwich [21, p. 77])

Argument 5. Following the line of thought of Argument 1, it has been argued

that compositional and Tarskian truth theories are not available to the deflationist because they cannot be *derived* from what the deflationist considers to be the basic facts about truth. A particularly forceful objection of this kind is due to Gupta [11], and has generated much discussion in the literature. Of course, it was essentially for this reason that Tarski [38, p. 257] rejected an axiomatisation of truth based purely on instances of T-schema.

Argument 6. The sixth argument for the incompatibility thesis is an argument from substantiality. It has been claimed that compositional and Tarskian truth theories encapsulate substantial conceptions of truth because such theories are often *non-conservative* over their base theory, i.e. they allow us to prove claims in the language of the base theory that are not already provable in the latter. In other words, Tarskian and compositional truth theories often allow us to gain more knowledge about the objects the base theory is about; thus, their truth predicate must be playing an explanatory role and, therefore, they must convey a substantial notion of truth.

At first glance, these arguments look convincing. At any rate, it appears that they have been accepted by many opponents of deflationism. Indeed, in light of the previous quotes by Horwich, one would think that (some) deflationists themselves have accepted the incompatibility thesis. However, there is a tension: there is a considerable amount of textual evidence that deflationists do in fact reject the incompatibility thesis. Field's work is a clear example, as he systematically advocates truth theories that validate compositional principles,[2] as does Horsten.[3] Moreover, Field [4] himself has offered a forceful response to Argument 6, defending compositional truth theories against the charge of substantiality.

In addition, Horwich explicitly notes that the notions of truth, reference, and satisfaction actually do interact in the way indicated by Tarski, at least for certain fragments of English. He does not object to Tarski's theory on the ground that its axioms are *incorrect*. Rather, he claims that these axioms "should not be treated as explanatorily basic, but should be explained in terms of simple, separate, minimal theories of truth, reference, and satisfaction." [20, pp. 111-112]. Similarly, Horwich explicitly endorses several compositional principles of truth, such as that a conjunction is true if and only if both conjuncts are true. Again, the reason that they do not feature among the axioms of his theory of truth is simply that he doesn't consider them as explanatorily basic.

In order to dissolve this tension and to show that the incompatibility thesis is incorrect, it will be helpful to have a closer look at the different purposes formal theories of truth can serve.

[2] See, for instance, Field [5, 6].
[3] See, e.g. Horsten [18].

4 What is a formal truth theory good for?

As many have pointed out, formal truth theories can serve various purposes. Soames [36], for instance, distinguishes between three things a truth theory can do. First, it can serve as a faithful account of the behaviour of our natural language truth predicate. Call this a 'descriptive purpose'. As Soames points out, not many philosophers have attempted to provide a truth theory suited for descriptive purposes; rather, this is seen as the proper domain of linguistics. Philosophers, instead, have been mostly concerned with truth theories that put forward a new, precise, and consistent truth-like predicate intended as a replacement for our (possibly defective) natural language truth predicate. Soames gives the example of Tarskian truth theories as an illustration of theories of this kind. The third purpose he discusses involves cases where a notion of truth, taken to be antecedently understood, is deployed to explicate other related concepts such as meaning or knowledge or some general metaphysical view. A prominent example here is the use to which Davidson attempted to put Tarski-style truth theories in giving an account of natural language semantics.

In addition to these three purposes that a truth theory can serve, we would like to propose a fourth, which should be very close to the deflationist's heart. According to the deflationist's logico-linguistic function thesis, the truth predicate serves a role akin to that of the logical connectives. If deflationism is right, the truth predicate – roughly like conjunction, the conditional, the universal quantifier, etc. – plays an important expressive or inferential role. Thus, for deflationists and other philosophers who believe that truth plays such a role, it is only reasonable to want a formal truth theory capable of characterising the validity or correctness of inferences involving the notion of truth. As an analogy, it is helpful to compare the way in which, for instance, calculi for first-order logic play the role of characterising the validity or correctness of inferences involving negation, conjunction, quantifiers, etc. When a theory of truth plays this role, let us say that it serves a 'logical purpose'.[4]

The language of a formal truth theory intended to serve a logical purpose should be extensive or extensible enough that we can formalise (most of) our arguments involving the truth predicate (and other logical terms), just as first-order languages do for their logical terms. The aim, then, is to provide a theory that diagnoses an argument (suitably formalised and regimented) as valid just in case its premises entail the conclusion against the background of the truth theory.

In the remainder of this section we will argue that, while the first four arguments considered in the previous section have significant force when applied to formal truth theories intended to play a *descriptive* purpose, their force considerably diminishes when applied to formal theories intended to play a *logical*

[4] We use this terminology for lack of a better alternative: in particular, in using it we do *not* wish to suggest that truth is a distinctively logical notion; rather, we use it to emphasise the aim of laying down general principles governing the validity or correctness of inferences involving truth.

purpose instead.

If one is working with a broadly descriptive goal in mind, it is natural to impose certain constraints on one's formal truth theory Γ. Most significantly, it will be required that the truth-theoretic component of Γ closely reflects the actual usage or meaning of 'true'. Of course, it is almost inevitable in practice that even a truth theory offered in a descriptive spirit will be idealised in various ways; but the point is that the main criterion of success is fidelity to established usage. Similarly, the theory should not lapse into oversimplification. As Argument 4 suggests, we should plausibly expect that a descriptive theory Γ satisfactorily describes not only the behaviour of the truth predicate taken alone, but also within complex environments, e.g. within 'that'-clauses, probabilistic locutions, environments containing attributive adjectives, mass terms, etc., as these constructions are prevalent in (even theoretically informed) natural language.

Plausibly, given the emphasis that deflationism places on the equivalence and the logico-linguistic function thesis, one's deflationist commitments will impose additional constraints on theories put forward to serve the descriptive project. One is that no real definition of truth is possible, as Argument 2 suggests. Another is that no descriptively adequate truth theory will put forward an even nominally *definable* and, therefore, *eliminable* truth predicate, as prescribed by Argument 3. Finally, given the basic and exhaustive role the equivalence thesis ascribes to formal transparency principles in governing our usage of 'true', deflationists are committed to the claim that a broadly descriptively adequate truth theory will consist of instances (perhaps within a restricted class of sentences) of local disquotation. In particular, as Argument 1 suggests, all other truth-theoretic principles must then be derived from those instances; there seems to be no room for compositional or Tarskian axiomatisations within the descriptive project, at least as carried out by deflationists.

It is this line of reasoning, we believe, that lends a spurious plausibility to Arguments 1-4 in the previous section. To the extent that deflationists are attempting to offer an axiomatic truth theory capable of playing a *descriptive* role, these arguments can be endorsed and the constraints they propose can be taken as genuine ones. However, we believe that the plausibility of these arguments diminishes substantially when applied to a truth theory intended to serve a logical purpose, as we will now explain.

Assume we are attempting to formulate a formal theory of truth capable of serving a logical purpose. Naturally, the first thing we require is that the theory contains or entails principles governing the truth predicate sufficient to allow it to serve its logico-linguistic role.

At first, it may seem as if a truth theory of this kind must satisfy the same conditions that deflationists impose on their *descriptive* truth theories. After all, if a formal truth theory adequate for logical purposes was not also descriptively adequate, it is not clear how we could be able to properly formalise natural

language arguments involving the truth predicate. Moreover, if deflationism is right about the role of the (theoretically informed) natural language truth predicate, it seems that the truth predicate of a theory that is faithful to our usage should also be capable of serving that role.

However compelling these points may seem, we will argue that they do not adequately take into account the fact that simplification and idealisation are considerably more admissible in a theory intended for logical purposes than in a purportedly descriptive account. Moreover, for theories serving a logical purpose, it is less important to be faithful to the precise way that the meaning of the truth predicate is fixed in English. To make this point clear, we turn once again to the analogy with first-order languages and calculi.

Note first that first-order languages do not admit indexicals, 'that'-clauses, probabilistic locutions, attributive adjectives, mass terms, etc.; they work, as it were, with eternal or context-independent sentences only. For instance, if one wishes to formalise an English argument in a first-order language, one first needs to replace all indexicals with names referring to their referents (in the context of utterance) and make corresponding amendments. Arguably, this is a small price to pay for perspicuity and elegance, which is evidenced by how widely first-order logic is implicitly deployed in the analysis of the validity of arguments. It would seem equally reasonable for us to pay this price in the case of a formal truth theory we wish to adopt for logical purposes. *Pace* Argument 4, a theory of truth (be it a deflationary one or not) should not be expected to account for the interaction between truth and indexicals, 'that'-clauses, and other natural language oddities. Simplification and idealisation are permissible to a larger degree if one's purpose is not fundamentally descriptive.

Second, note that for logical purposes, whether the axioms of our theory coincide with the most basic principles governing our usage of the truth predicate in natural language does not matter. Adverting again to the analogy with logical constants, note that natural language usage is often ignored, just as e.g. Hilbert-style or sequent calculi for first-order logic make no pretence of capturing the most psychologically basic patterns of inference in their basic rules. All that matters is that, taken together, they provide us with an adequate account of validity for arguments involving truth, modulo simplification and idealisation. Thus, whilst Argument 1 and 2 might be compelling when considering formal truth theories intended for descriptive purposes, they aren't so when we want our theories for logical purposes instead.

Finally, if we are interested in making inferences involving only certain expressions, it would seem permissible to restrict our logical or truth-theoretic principles in such a way that our logical terms or truth predicate interact exclusively with the relevant class of expressions. Of course, as an account of validity for a more encompassing class of expressions, the resulting theories will not be satisfactory; their use would be limited. In the case of truth, this could amount, for instance, to restricting the sound instances of disquotation (whichever these are) to a proper subclass. As a result, the truth predicate of the theory could

turn out to be nominally definable. However, this does not conflict with the fundamental tenets of deflationism, as Argument 3 suggests, provided that the truth predicate of certain *extended* truth theories is not definable. It is compatible with the indefinability of our natural language truth predicate that when transparency is restricted to a subclass of expressions in our formal theories, the resulting predicate admits a nominal definition.

Despite the fact that Arguments 1-4 of the previous section fail to apply to truth theories intended for a logical purpose, theories of this kind should nevertheless still be expected to satisfy certain other conditions. We would like to mention two fundamental requirements: the *functionality* criterion and the *insubstantiality* criterion.[5]

The functionality criterion, indicated a few paragraphs above, demands that the axioms of the theory allow the truth predicate to perform the logico-linguistic function of truth – at least to a reasonable extent. Of course, which axioms are sufficient to achieve this goal depends entirely on the precise nature of the logico-linguistic function of truth, so an account of the latter is needed. We have developed such an account in [27]. In Section 5 we briefly review it and extract a precise criterion of functionality from it.

The insubstantiality criterion demands that truth theories do not convey a non-deflationary notion of truth, i.e. they should not entail that truth is a substantial property. What counts as a substantial truth property is naturally a very controversial issue which we are not able to resolve here. However, in Section 6 we will argue that if one starts with a truth theory that is taken to be insubstantial, then adding certain principles which, in a sense to be explained, follow from the axioms of the theory does not render the relevant notion of truth substantial.

Taken together, these criteria will allow us to recognise certain compositional and Tarskian theories as being deflationary, and hence to refute the thesis that compositional and Tarskian theories are necessarily committed to a substantial notion of truth (cf. §8).

5 The functionality criterion

Our first criterion on an axiomatisation of truth intended for logical purposes is that such a theory must enable the truth predicate to fulfil its logico-linguistic role. In the present section, we will provide a precise formulation of this criterion.

It is striking that the purported logico-linguistic function the truth predicate is supposed to play has rarely been the subject of study in the literature, even

[5]There have been several attempts to formulate general requirements on axiomatic theories of truth, e.g. Leitgeb [23] and Sheard [35]; a list of desiderata specifically designed for deflationists has been proposed by Halbach & Horsten [16]. Although reasons of space prevent a direct comparison, it should be emphasised that our desiderata differ decidedly from theirs.

by logicians or deflationists. One of the few articles providing a positive and formally precise account of the function is Halbach's [12]. In [26] we discussed this account and others that are hinted at in the literature, and argued they are unsuccessful; in [27] we put forward our own positive account. Inspired by a tradition that originated in Ramsey, and to which Quine, Grover, and Azzouni among others also belong,[6] we argued that the function deflationism ascribes to the truth predicate is best understood as enabling us to simulate sentential and predicate quantification within a first-order framework. In other words, the truth predicate lets us quantify into sentence and predicate position in an indirect way, i.e. without introducing sentential or predicate quantifiers. In still other words, the truth predicate and sentential and predicate quantifiers serve the same purpose.

For instance, to assert all theorems of first-order Peano arithmetic one could work with a monadic operator $\Box$ expressing provability in this theory plus sentential quantifiers, and write $\forall\alpha\,(\Box\alpha \to \alpha)$. Alternatively, one could use a provability predicate $\mathrm{Prov}(x)$ and a truth predicate, and assert $\forall x\,(\mathrm{Prov}(x) \to \mathrm{T}x)$. Similarly, one can generalise on $\varphi(t) \vee \neg\varphi(t)$ using second-order quantifiers, as in $\forall X\,(Xt \vee \neg Xt)$, or one can turn to the truth predicate and say

$$\forall x\,(\mathrm{Form}_1(x) \to (\mathrm{T}x(\ulcorner t \urcorner) \vee \neg \mathrm{T}x(\ulcorner t \urcorner))$$

where $\mathrm{Form}_1(x)$ expresses the property of being a formula with only one free variable and $x(y)$ is the result of substituting in x the free variable with the term denoted by y.

In order to explain and substantiate our claims, in [27] we have offered a series of formal results that establish that every theory in a language with sentential or predicate quantifiers – second- or higher-order, predicative or impredicative – can be 'naturally reformulated' in a language containing a purely disquotational truth predicate instead.[7] We first indicated how to translate every formula of a higher-order language into a first-order language with a truth predicate in a natural and effective way, i.e. along the lines of our examples in the previous paragraph. Instances of sentential comprehension – be they predicative or impredicative – translate into instances of local disquotation, provided the comprehension instances contain no free sentential variables, and into instances of *uniform* disquotation otherwise. On the other hand, instances of predicate comprehension always require uniform disquotation. The latter is a principle that generalises local disquotation to formulae with free variables. For instance, the following does so for formulae with one free variable:

(Uniform T-schema) $\qquad\qquad \forall t\,(\mathrm{T}\ulcorner\varphi(\dot{t})\urcorner \leftrightarrow \varphi(t^\circ))$

[6] See, for instance, Ramsey [30, p. 158], Quine [28], Grover [9], and Grover et al. [10], and Azzouni [2].

[7] This result relies on the assumption, mentioned in §2, that every object in the domain has a name. Again, that restriction can be lifted if we work with a satisfaction rather than a truth predicate.

Here, $\forall t\, \psi$ abbreviates $\forall v\, (\mathrm{ClTerm}(v) \rightarrow \psi)$ for a suitable variable v, where $\mathrm{ClTerm}(v)$ expresses the property of being a closed term; $\ulcorner \varphi(\dot{t}) \urcorner$ denotes the result of substituting t for the free variable in φ, and t° denotes the value of the term t.

We then proved that the proposed translation is a relative interpretation of the higher-order calculus into a – classical and consistent – disquotational truth theory, as it maps every higher-order derivation into a derivation in the truth theory that extends first-order classical logic with a suitable syntax theory and all instances of the (Uniform) T-schema for formulae in the range of our translation. This result is novel in so far as it establishes that (uniform) disquotation can even interpret *full impredicative predicate comprehension*, i.e. principles of the form

$$\exists X\, \forall v\, (Xv \leftrightarrow \varphi)$$

where φ itself may contain bound predicate variables. It shows that the proof-theoretic power of truth is much greater than previously thought. Moreover, we also showed that all *inferences* between translations that can be carried out in this truth theory are derivable in the calculus for higher-order logic.

Sentential and predicate quantifiers allow us to directly generalise over all sentences and formulae in the higher-order language. The truth predicate, we concluded, can bring about the same logical power: if (uniform) disquotation for translations of higher-order formulae is available, we can simulate quantification over the latter using their translations as proxies. More generally, one can use a truth predicate to simulate sentential and predicate quantification over a given class of expressions as long as the instances of disquotation for the expressions in this class – or their translations – are available. Our account of the function of truth confirms the common but rarely substantiated claim that (uniform) disquotation is both sufficient and necessary for the truth predicate to fulfil its role.

As a consequence, local disquotation for the class of sentences we wish to generalise over is desirable in our truth systems, and uniform disquotation even more so, especially if we wish to generalise into predicate position. In general, we would like to put forward the following adequacy criterion for formal truth theories intended for logical purposes:

Functionality A formal theory of truth intended for logical purposes should entail all instances of (uniform) disquotation for the class of expressions one wishes to generalise over.

Note that this criterion implies that generalising over the whole class of expressions of the language of the theory itself is not possible if classical logic is assumed in the background. For that would require that all instances of disquotation for sentences containing the truth predicate are derived, and triviality would follow. If one wishes to generalise unrestrictedly over all expressions of the language, one should probably look into non-classical truth theories instead. However, this might turn out to be not as straightforward as it seems. It

is not entirely clear to us what inferences the truth predicate should validate in that case, as our results only establish the relative interpretability of *classical* higher-order theories in a disquotational truth theory. On the one hand, classical higher-order theories seem to be too strong to be relatively interpretable in a non-classical truth theory. On the other hand, very little is known about non-classical systems of higher-order quantification. In any case, the general lesson of our discussion should be clear: one first needs to determine what axioms and rules for truth are needed in a particular logic for it to fulfil its logico-linguistic function, and then derive a criterion of functionality from that.

6 The insubstantiality criterion

In the previous section we have formulated a criterion of functionality that any formal truth theory intended for logical purposes ought to satisfy. However, not any such theory will do – for example, inconsistent or trivial theories are excluded, as they would obviously fail to adequately characterise the validity or correctness of inferences involving the notion of truth. Moreover, as we anticipated towards the end of Section 4, truth theories that can be legitimately endorsed by deflationists for logical purposes should also satisfy the insubstantiality criterion: they must not encapsulate a substantial notion of truth.

What encapsulating a substantial notion of truth amounts to is of course a matter of controversy, and we will not engage with the general metaphysical question of what a substantial property is. Instead, we will show that if one starts with an insubstantial theory, then the addition of certain compositional and Tarskian principles will not inflate that notion of truth. This will be the case, for instance, if the latter generalise on a schematic consequence of the starting theory. Again, we will not say much about what constitutes an insubstantial theory of truth. However, if deflationism is correct, then there must be at least one such theory – e.g. the theory consisting of all correct instances of disquotation. The purpose of this section is to show that if one starts from such an insubstantial and restricted truth theory, then adding certain compositional or Tarskian principles will not lead to an inflated notion of truth.

Despite its aspirations for generality, the T-schema cannot be stated by a single, universally quantified claim of the form $\forall x\,(\mathrm{T}x \leftrightarrow \dots)$, as each instance has a sentence φ occurring inside quotes on the left-hand side and outside them on the right-hand side. The fact that each φ is both used and mentioned in its corresponding instance of the T-schema precludes a straightforward generalisation of the disquotational principle.

However, there are salient schematic principles that follow from the T-schema together with background syntactic assumptions, and which can easily be generalised. A simple warm-up example is given by the Uniform T-schema, which we already discussed in the previous section. Let Δ be the class of sentences for which an instance of local disquotation is available. Assume that, for

some formula $\varphi(x)$, the sentence $\varphi(t)$ is in Δ for every term t:

$$\mathrm{T}\ulcorner\varphi(t)\urcorner \leftrightarrow \varphi(t)$$

Then it is easily seen that the relevant instance of uniform disquotation, i.e.

$$\forall t\,(\mathrm{T}\ulcorner\varphi(\dot{t})\urcorner \leftrightarrow \varphi(t^\circ))$$

is a straightforward generalisation of the schematic principle.

Let us look at another example. Consider the set of sentences φ such that both it and its negation are in Δ. Then, for each such sentence φ we can prove:

$$\mathrm{T}\dot{\neg}\ulcorner\varphi\urcorner \leftrightarrow \neg\mathrm{T}\ulcorner\varphi\urcorner$$

Since $\ulcorner\varphi\urcorner$ is a singular term, we can generalise on this principle as follows:

$$(\mathrm{T}\neg\restriction\Delta) \qquad \forall x\,(\mathrm{Sent}_\Delta(x) \wedge \mathrm{Sent}_\Delta(\dot{\neg}x) \rightarrow (\mathrm{T}\dot{\neg}x \leftrightarrow \neg\mathrm{T}x))$$

Analogously, all the instances of the following principle (where φ, ψ, and $\varphi \wedge \psi$ are in Δ) follow from the T-schema as well:

$$\mathrm{T}\ulcorner\varphi\urcorner\dot{\wedge}\ulcorner\psi\urcorner \leftrightarrow \mathrm{T}\ulcorner\varphi\urcorner \wedge \mathrm{T}\ulcorner\psi\urcorner$$

Replacing all occurrences of $\ulcorner\varphi\urcorner$ with x and those of $\ulcorner\psi\urcorner$ with y we can generalise on this schema by the following:

$$(\mathrm{T}\wedge\restriction\Delta) \quad \forall x\forall y\,(\mathrm{Sent}_\Delta(x) \wedge \mathrm{Sent}_\Delta(y) \wedge \mathrm{Sent}_\Delta(x\dot{\wedge}y) \rightarrow (\mathrm{T}x\dot{\wedge}y \leftrightarrow \mathrm{T}x \wedge \mathrm{T}y))$$

Analogous principles for the other propositional connectives can be obtained likewise. Similarly, compositional principles for the quantifiers can be seen as generalisations of schematic consequences of local disquotation.[8]

Provided that Δ is a T-free *sublanguage* of $\mathscr{L}_\mathrm{T}$ (i.e. Δ is closed under logical predicates and operators) containing finitely many predicate symbols, one can also generalise on the T-schema by means of a so-called Tarskian definition, $\mathrm{T}\restriction\Delta$. For instance, if all closed terms of $\mathscr{L}_\mathrm{T}$ occur in formulae in Δ, identity is the only predicate symbol, and $\neg$, $\wedge$, and $\forall$ are the only logical operators occurring in formulae in Δ, the following principle just 'puts together' the instance of uniform disquotation for the identity predicate and the compositional principles for the logical terms:

$$\begin{aligned}\forall x\,(\mathrm{T}x \;\leftrightarrow\; \mathrm{Sent}_\Delta(x) \;\wedge\; &(\exists s\exists t\,(x = s\dot{=}t \wedge s^\circ = t^\circ) \vee\\ &\exists y\,(x = \dot{\neg}y \wedge \neg\mathrm{T}y) \vee\\ &\exists y\exists z\,(x = y\dot{\wedge}z \wedge \mathrm{T}y \wedge \mathrm{T}z) \vee\\ &\exists y\exists z\,(x = \dot{\forall}y\, z \wedge \forall t\,(\mathrm{T}z(\dot{t}))))\end{aligned}$$

[8] Recall that we assumed that we can prove in the base theory that for every object there is a term denoting this object. Again, if one wants to lift that restriction, we need to work with a satisfaction predicate instead.

We have seen how uniform disquotation, as well as compositional and Tarskian principles can be "extracted" from the T-schema by generalising on certain schematic principles that follow from it. They are general principles all of whose instances are already entailed by the latter. Arguably, these principles just provide more general ways of presenting the T-schema itself or some of its schematic consequences. If this is on the right track, then it is hard to see why they should be more substantial than the principles we started with. In this context, it is interesting to note that the way in which uniform disquotation generalises on local disquotation is not too different from the way the compositional principles do, so it is surprising that nobody has disputed the suitability of uniform disquotation as a deflationary truth principle. As we see things, if uniform disquotation is acceptable, then so are compositional principles.

There is one obvious worry regarding our reasoning above. Compositional, Tarskian, and uniform disquotation principles don't follow *logically* from their corresponding instances, but are usually (proof-theoretically) stronger than them, due to the compactness of the logical consequence relation. The possibility that this additional content inflates the notion hasn't been completely ruled out, despite the fact that these principles merely generalise schematic consequences of local disquotation.

Horwich's method for dealing with this objection is well known. Non-basic facts about truth need to be explained in terms of transparency principles together with *further* explanatory factors, i.e. principles that have nothing specifically to do with the truth predicate (cf. Horwich [20, p. 24], [21]). We believe this strategy is essentially sound. Horwich himself appears to appeal to some form of ω-rule as an additional principle, which has provoked some criticism due to its infinitary character (cf. Raatikainen [29]). Fortunately, there are other suitable principles. We will first describe what these principles are, and then discuss whether they are available to the deflationist.

We can bridge the gap between generalisations such as compositional, Tarskian, and uniform disquotation principles and their instances by informing the truth theory we are working with that, whenever it schematically proves all instances of a certain formula, the inference to the general claim that all instances of this formula hold is permissible (see Halbach [13] and Horsten & Leigh [19] for some formal results). Let $\mathrm{Prov}_\Gamma(x)$ express in $\mathscr{L}$ that x is a theorem of the formal theory Γ. Our gap-bridging principles take then the following form:

$$(\mathrm{GBP}(\Gamma)) \qquad \forall t\, \mathrm{Prov}_\Gamma(\ulcorner \varphi(\underset{\cdot}{t}) \urcorner) \rightarrow \forall t\, \varphi(t^\circ)$$

Principles of this kind – not provable in Γ for familiar Gödelian reasons – allow us to formalise the "extraction" of a general claim from its instances into a proper derivation. For example, let Γ extend the base theory Σ with all instances of local disquotation for sentences in Δ. Since Γ schematically derives all instances of

$$\mathrm{Sent}_\Delta(t) \wedge \mathrm{Sent}_\Delta(\underset{\cdot}{\neg} t) \rightarrow (\mathrm{T}\underset{\cdot}{\neg} t \leftrightarrow \neg \mathrm{T}t)$$

for every closed term t, adding GBP(Γ) to Γ delivers

$$\forall t\,(\mathrm{Sent}_\Delta(t^\circ) \wedge \mathrm{Sent}_\Delta(\dot{\neg}t^\circ) \rightarrow (\mathrm{T}\dot{\neg}t^\circ \leftrightarrow \neg \mathrm{T}t^\circ)$$

which, together with the fact – provable in Σ – that each sentence in Δ is denoted by a term in the language, entails the compositional principle $\mathrm{T}\neg{\restriction}\Delta$. Applying a similar reasoning, we can derive the Uniform T-schema and compositional principles restricted to Δ for the other propositional connectives in Γ extended with GBP(Γ). Finally, compositional principles for the quantifiers can be derived in Γ+GBP(Γ'), where $\Gamma' = \Gamma$+GBP(Γ).

A similar argument can be given in the case of so-called Tarskian definitions, $\mathrm{T}{\restriction}\,\Delta$. Let Γ be as before. If, additionally, Δ is a T-free sublanguage of $\mathscr{L}_\mathrm{T}$ as before, then $\mathrm{T}{\restriction}\,\Delta$ follows in Γ from uniform disquotation and the compositional principles for all logical terms, both restricted to Δ, plus the following:

$$(\beta{\restriction}\Delta) \qquad\qquad \forall x\,(\mathrm{T}x \rightarrow \mathrm{Sent}_\Delta(x))$$

Thus, $\mathrm{T}{\restriction}\,\Delta$ follows in Γ from (iterated applications of) GBP(Γ) together with $\beta{\restriction}\Delta$, which states that only sentences in Δ can be true.

Does GBP(Γ) qualify as a suitable additional principle that the deflationist can employ in explaining certain facts about truth? Suppose we employ classical logic, as Horwich does, and assume for a moment that we firmly endorse the deflationary acceptable truth theory Γ: when I learn that some sentence is provable in Γ, I have good reasons to believe it. Now let $\varphi(x)$ be a formula of $\mathscr{L}_\mathrm{T}$ and consider the following instance of excluded middle:

$$(\forall t\,\mathrm{Prov}_\Gamma(\ulcorner\varphi(\dot{t})\urcorner) \rightarrow \forall t\,\varphi(t^\circ)) \;\vee\; \neg\,(\forall t\,\mathrm{Prov}_\Gamma(\ulcorner\varphi(\dot{t})\urcorner) \rightarrow \forall t\,\varphi(t^\circ))$$

Which of the two disjuncts should we endorse? Consider the second disjunct. Accepting it commits us to the claim that although $\varphi(t)$ is provable in Γ for every closed term t, nonetheless there is a closed term t such that $\neg\varphi(t)$. This entails that we should not accept some consequences of Γ! Since we firmly endorse Γ, we should reject the second disjunct, and therefore accept the first disjunct. But the latter is just an instance of GBP(Γ).

Let us clarify one point, before dealing with some objections. Given an instance of excluded middle, one can in general remain agnostic about which disjunct obtains. For example, a classical set theorist is committed to the claim that either the continuum hypothesis or its negation holds, but she may remain agnostic about which disjunct holds barring new evidence. However, we maintain that the present case is different. The second disjunct entails that some consequences of Γ don't hold. Thus, if you firmly endorse Γ, you ought to reject it and accept the first disjunct, even if the statement is independent of Γ. Anything else would be incoherent. But now, once you have accepted GBP(Γ) as an additional (non-truth-theoretic) principle, other truth-theoretic principles follow.

We cannot see any good reason why the truth-theoretic principles that follow from adding GBP(Γ) to our truth theory Γ should inflate the notion of truth.

We have assumed that the truth-theoretic principles of Γ are insubstantial. In arguing for GBP(Γ), we have not appealed to the notion of truth, let alone a substantial notion of truth. Moreover, GBP(Γ) itself isn't formulated in terms of truth. In what follows, we anticipate three possible objections.

Objection 1. The argument assumes that $\mathrm{Prov}_\Gamma(x)$ "expresses" the property of being provable in Γ. The standard explanation of why it does so involves the notion of truth in the standard model of the base theory Σ – e.g. the standard model of arithmetic. However, the latter is not admissible to a deflationist, because on their account truth is characterised through transparency.

Our reply to this objection is essentially identical to that given by Cieslinski [3, p. 153]. Very roughly, $\mathrm{Prov}_\Gamma(x)$ "expresses" the property of being provable in Γ because the way the predicate is defined structurally resembles the way how 'provable in Γ' is defined in the metalanguage of Γ. We find this response especially plausible in this context because deflationists usually rely on a use theory of meaning – rather than on truth-conditional semantics – according to which the meaning of 'provable in Γ' must be given through some rules for using that expression.

Objection 2. GBP(Γ) is a schematic principle, and according to deflationists, the sole purpose of the truth predicate is to generalise sentence places in our language. Thus, we ought to formulate GBP(Γ) as a single statement deploying the truth predicate. But then it becomes apparent that our additional principle is of a truth-theoretic nature after all.

We do not find this objection very convincing. First, it is not generally the case that whenever we generalise a schema using the truth predicate, the resulting statement is a truth-theoretic statement. The claim that everything the Pope said is true or that all theorems of arithmetic are true are not truth-theoretic statements, although they involve the notion of truth. According to the logico-linguistic function thesis (the second core tenet of deflationism), such generalisations do little more than express all papal assertions or all theorems of arithmetic in a compact way.

Second, even if the truth predicate allows us to express the schema in a single statement, we are certainly not obliged to do so. At any rate, it is hard to see how the fact that we can derive compositional principles of truth using GBP(Γ) – which is not stated in terms of truth – could be undermined by the fact that we can generalise GBP(Γ) using the notion of truth.

Third, we know that due to the paradoxes it is not possible to generalise over all sentence places in our language (at least as long as we adhere to classical logic). We can only do so for a restricted class of sentences. But GBP(Γ) is a schema that ranges over all sentences. Thus it is not even clear that we can generalise GBP(Γ) using the notion of truth. (It might be thought that all this shows is that the deflationary account of truth is incompatible with the use of classical logic. We have argued in [26] that this is not the case.)

Objection 3. GBP(Γ) is unacceptable because it is inconsistent with certain

unrestricted compositional axioms for truth.

We are not particularly worried by this objection either. On our view, deflationists ought to reject unrestricted compositional axioms for truth on quite independent grounds already, so their inconsistency with GBP(Γ) cannot cast doubt on the latter. Very roughly, the reason why deflationists ought to reject unrestricted compositional axioms for truth is *precisely* because not all of their instances are generally entailed by restricted disquotational principles of truth. If only instances of disquotation for a given class of expressions Δ are available, it is hard to see how compositional or Tarskian principles whose instances go beyond Δ can be justified on the basis of the original theory, even if additional non-truth-theoretic principles are invoked.[9] We return to this point in Section 8, at the end of the paper.

Our preceding argument for GBP(Γ) relies on the law of excluded middle and so might not be available to all deflationists. It is difficult to say something in general here, as the matter will depend on the details of the non-classical system. At any rate, since our goal is merely to show that deflationism is compatible with compositional and Tarskian truth theories, it is sufficient if we can make our point in the case where the deflationist account is based on classical logic.

To sum up, we maintain that the addition of certain compositional, Tarskian, and uniform disquotation principles does not thicken the notion of truth conveyed by a deflationary adequate truth theory. First, we pointed out that certain compositional principles and Tarskian definitions are mere generalisations of schematic consequences of a class of instances of local disquotation, so it is hard to see how they could possibly inflate the notion of truth. We then pointed at the existence, under certain given conditions, of derivations of the more general principles from local disquotation plus other non-truth theoretic claims deflationists may reasonably endorse. (Of course, if such proofs are not available – which will largely depend on the restrictions imposed on local disquotation and the background logic – there is no guarantee of the legitimacy of the general principles.) This motivates the following criterion:

Relative Insubstantiality The (truth-theoretic) axioms of a formal truth theory are insubstantial if they are derivable in an insubstantial locally disquotational theory of truth together with additional non-truth-theoretic principles a deflationist may reasonably endorse.

The qualification 'derivable in an *insubstantial* locally disquotational theory etc.' is important: not every class of instances of local disquotation is necessarily insubstantial. For instance, the class that comprises all the instances, being inconsistent, entails every truth principle whatsoever, even those one would readily call inflationary, e.g. that truth is correspondence with fact (if expressible in the language). Other consistent subsets of this class will also be inadmissible for similar reasons, for although they will not entail every sentence of the language, some of them will entail substantial claims about truth, as will

[9] A similar point was made by Armour-Garb and Beall [1, section 5.1].

be seen in Section 8. As we have said before, we won't offer a definition of what constitutes an insubstantial disquotational theory of truth, but if deflationism is correct, such theories do exist – the theory consisting of all correct instances of disquotation being one of them.

7 The argument from conservativeness

There is one objection that one could mount against our criterion of insubstantiality. This is the argument from conservativeness, mentioned in Section 3. We will now deal with this objection.

The equivalence thesis commits deflationism to the idea that any attempt to uncover the nature of truth beyond disquotation, the quest for a real definition of truth in terms of simpler notions is futile. According to deflationism, truth cannot be defined or further analysed; it is a *sui generis* property, if a property at all. This is often expressed by saying that truth has no nature, is metaphysically thin, or is otherwise insubstantial, but of course these are just metaphors. Many, however, have taken them to be a – and even the – defining feature of deflationism. Moreover, some understand the insubstantiality of truth to entail that truth cannot have any explanatory power. Shapiro [34, 497], for instance, claims that "If truth/satisfaction is not substantial – as the deflationist contends – then we should not need to invoke truth in order to establish any results not involving truth explicitly". Formally, this translates in a natural way into what is known as the 'conservativeness requirement': deflationary truth theories should be conservative over their respective base theories – which should contain some amount of syntax (cf. Halbach [14]) – i.e. the addition of truth principles to a base theory should not allow us to prove new theorems in the language without the truth predicate. This requirement has been argued for by e.g. Horsten [17], Shapiro [34], and Ketland [22].

Another – related – road to conservativeness draws from the function deflationism assigns to truth. Its only purpose, as stated by the logico-linguistic function thesis, is a logico-linguistic one. Thus, it has been argued, there is no room for an explanatory role of truth within deflationism. In Horwich's [20, p. 52] words:

> A deflationist attitude toward truth is inconsistent with the usual view of it as a deep and vital element of philosophical theory. Consequently the many philosophers who are inclined to give the notion of truth a central role in their reflections in metaphysical, epistemological, and semantic problems must reject the minimalist account of its function. Conversely, those who sympathize with deflationary ideas about truth will not wish to place much theoretical weight on it. They will maintain that philosophy may employ the notion only in its minimalist capacity – that is, as something enabling the formulation of certain generalizations – and that theoretical problems must be resolved without it.

Again, if deflationary truth must not play a role in the resolution of theoretical issues, then the conservativeness requirement follows (or so it argued).

Most compositional theories of truth on the market are, however, not conservative over their respective base theory (cf. Halbach [15], Horsten [18]). Thus, if the conservativeness requirement is right, these theories are not deflationary. But also many *untyped disquotational* theories aren't conservative over their base theory either, some of which seem fairly attractive from a deflationary perspective, as the restriction they impose on the instances of disquotation can be justified from a philosophical point of view (cf. Picollo [25], for instance).

On our view, however, the conservativeness requirement not only does not follow from the core theses of deflationism outlined in the introduction of this paper but also is not a reasonable requirement to be imposed on deflationary truth theories. Indeed, we claim that the conservativeness requirement is the result of (a) inferring too much from the metaphor of insubstantiality and (b) failing to see what the function of truth really amounts to. The analysis of this function, briefly sketched in Section 5, actually points (in many cases) in the opposite direction.

Let us focus briefly on sentential and predicate quantifiers. While their role – whether logical or quasi-logical, we would not like to enter this dispute here – is merely expressive, their addition to a first-order base theory does not always yield a conservative extension. Now, in [27] we've shown that the logico-liguistic function deflationism ascribes to the truth predicate is best understood as enabling us to simulate sentential and predicate quantification in a first-order setting, as mentioned in Section 5. In other words, from a deflationist perspective, the truth predicate – together with the first-order quantifiers – has the *same function* as sentential and predicate quantifiers. As a consequence, we should not expect a formal truth theory well suited for functional purposes to conservatively extend its base theory either. On the contrary, non-conservativeness is just a feature of the truth predicate fulfilling its role. The conservativeness requirement cannot stem from the logico-linguistic function thesis; a 'mere' expressive role is compatible with the violation of conservativeness.

Can the equivalence thesis support an argument for conservativeness? If so, it would be devastating for deflationism: while one of its core theses would point to conservative theories, the other points in the opposite direction. Are the two fundamental theses of deflationism incompatible with each other? We believe this is not the case. If we look closely at the equivalence thesis, there is good reason to believe that the insubstantiality metaphor is just meant to indicate that the truth predicate, unlike other predicates, does not play a *descriptive* role in our language; truth ascriptions are not descriptions of the truth bearers involved. To quote Frege [8, p. 293], "nothing is added to the thought by my ascribing to it the property of truth", so the latter is not an ordinary or substantial property. As such, truth cannot play the explanatory role ordinary properties play, i.e. to highlight an aspect of the object of study that would explain some of the characteristics of this object. But this doesn't exclude

the possibility that the truth predicate plays an explanatory role of a different kind, i.e. in proofs. Indeed, sentential and predicate quantifiers can lead to new knowledge as well and therefore have explanatory value (assuming that proofs can have explanatory value), without being in any way descriptive. Their explanatory value derives solely from their role as a logico-linguistic device; the fact that they have explanatory value does not indicate in any way that they are 'substantial'. Since the truth predicate plays the same function as these quantifiers, similar considerations apply to it. Thus, we echo Field [4, p. 537] when he says that "any use of 'true' in explanations which derives solely from its role as a device of generalization should be perfectly acceptable".

We therefore conclude that the conservativeness requirement should be given up; it cannot be used as an argument against the admissibility of certain truth-theoretic axioms for deflationism.

8 Revisiting the incompatibility thesis

It is time to take stock. We have looked at a number of arguments for the incompatibility of deflationism, on the one hand, and compositional and Tarskian truth theories, on the other. We have pointed out that the majority of these arguments ostensibly presuppose a particular purpose, i.e. to *describe* the basic usage of the truth predicate in natural language. This is a legitimate enterprise and we do not necessarily disagree with some of the objections if judged against this purpose. However, we were quick to point out that the deflationist may want a formal theory of truth for a slightly different purpose, that is, to provide an account of the validity or correctness of arguments involving the truth predicate.

We have formulated two constraints that any formal truth theory intended to serve a logical purpose ought to satisfy: functionality and insubstantiality. A formal truth theory intended for logical purposes should entail all instances of (uniform) disquotation for the class of expressions one wishes to generalise over and, moreover, its axioms should be insubstantial. Although we did not provide a general criterion of insubstantiality, we argued that a formal truth theory is insubstantial if its axioms are derivable in a locally disquotational truth theory which is itself insubstantial together with additional non-truth-theoretic principles a deflationist may reasonably endorse. With these constraints at hand, let us now have a look at some of the classic formal truth theories one can find in the literature and see whether they can be endorsed by a deflationist.[10]

Let us start with what is probably the best-known and most simple formal truth theory: the theory that extends the base theory with all T-free instances of the T-schema – usually known as TB, for 'Tarski Biconditionals'. This theory satisfies our functionality criterion: if one merely wishes to quantify into sentence position over the class of T-free sentences, TB will do. Moreover, it

[10] For an overview of axiomatic truth theories, see Halbach [15] or Horsten [18].

is widely believed to convey an insubstantial notion of truth. Based on this, the uniform version of TB, UTB (for 'Uniform Tarski Biconditionals'), can also be seen to be deflationary because its axioms follow from TB together with additional non-truth-theoretic principles an advocate of TB may reasonably endorse, i.e. GBP(TB). Since UTB entails all instances of uniform disquotation for T-free predicates, it improves on TB, as it also allows us to quantify into predicate position over this class of formulae.

Similar considerations also apply to other locally disquotational theories: if the local theory is in good standing, so will be its uniform version. Note, however, that some locally disquotational theories might actually not be in good standing. This is obviously the case of the (classical) theory containing an instance of the T-schema for each sentence of $\mathscr{L}_{\mathrm{T}}$, as it is inconsistent. But there are other purely disquotational theories that are consistent and yet violate some of our criteria. Assume φ expresses a substantial truth principle – e.g. that truth is correspondence with the facts. Deploying a trick of McGee [24], we know there is an instance of local disquotation that is provably equivalent (in the base theory) to φ. Hence, any theory containing that instance will be substantial.

Let us now turn to compositional truth theories, i.e. systems in which instances of disquotation are only given for atomic expressions – or sometimes also negations of atomic expressions – whereas other truth axioms are compositional. As we have argued, the latter are admissible if they follow from Γ and suitable non-truth-theoretic principles. Such is the case of the axioms of CT, which extends the base theory with uniform disquotation for each *primitive* predicate in the T-free language and compositional principles for the connectives and quantifiers, also restricted to sentences without T. CT is acceptable because it follows from GBP(UTB) (cf. Halbach [13]), which we already have seen to be acceptable.

Still, one might wonder what the use of compositional theories like CT would be, given that they merely generalise on instances of disquotation, which are already sufficient for the function of truth. Since all these instances of disquotation are derivable in the compositional theory, there seems to be no reason not to endorse it. But are there any positive reasons?

There are at least two – intertwined – motives why compositional theories could be preferable to corresponding disquotational systems. First, compositional principles allow us to reason more generally about truth. This can, in turn, provide us with simpler and shorter proofs (cf. Fischer [7]). Second, compositional principles can be used to provide us with a finite or more concise theory. If the T-free fragment of the language contains finitely many primitive predicates, CT can be seen as a finite and more general way of "formulating" the truth-theoretic part of both TB and UTB. If there are infinitely many primitive predicates in the language instead, CT also contains infinitely many axioms, but is still more general and concise than its disquotational counterparts, as e.g. it doesn't contain one instance of disquotation for each negated expression but all negations are dealt with by a single axiom in a general manner, and similarly

for the other logical terms.

For analogous reasons, Tarskian truth theories can be deflationary admissible for logical purposes and even preferable to local or uniform disquotational theories for the same class of expressions: they are more general and concise than the latter. Furthermore, since they have the form of a (recursive) definition, we know they do not introduce any inconsistencies to the base theory, which is clearly a theoretical advantage.

This shows that the incompatibility thesis – i.e. that deflationism, on the one hand, and compositional and Tarskian theories, on the other, are not compatible – is mistaken after all. However, so far we have only given evidence of the admissibility of typed theories of truth. Let us therefore conclude the paper by briefly surveying some untyped theories.

Let us first consider the well-known system KF, Feferman's axiomatisation of Kripke's fixed-point theory of truth in classical logic. KF extends the base theory with uniform disquotation for atomic and negation of atomic formulae that don't contain T, plus "positive" compositional principles for *every* sentence of the language, including those containing T, and an axiom governing attributions of untruth. No axiom of the theory states that truth commutes with negation, but compositional axioms for double negations, conjunctions, disjunctions, universal claims, etc., and negated conjunctions, negated disjunctions, negated universal statements, etc. belong to KF. For instance, the compositional axiom for negated disjunctions is the following:

$$(\mathrm{T}\neg\vee) \qquad \forall x\forall y\,(\mathrm{Sent}_{\mathscr{L}_\mathrm{T}}(x) \wedge \mathrm{Sent}_{\mathscr{L}_\mathrm{T}}(y) \rightarrow (\mathrm{T}\dot{\neg}(x\dot{\vee}y) \leftrightarrow (\neg\mathrm{T}x \wedge \neg\mathrm{T}y)))$$

Let us now ask whether KF satisfies the criteria we set out. Is it functional? KF implies instances of (uniform) disquotation for a certain class of expressions Δ (including all T-free sentences), so if one's goal is to quantify over expressions in Δ, functionality is satisfied. Is it insubstantial? We have not provided an absolute criterion of insubstantiality, but one way to show it to be insubstantial would be to look for an insubstantial disquotational theory that implies the axioms of KF, given additional non-truth-theoretic principles a deflationist may reasonably endorse.

Note that KF's compositional axioms are unrestricted, that is, they govern the interaction of the truth predicate and the logical operators as they apply to every expression of the language. Thus, a natural theory of disquotation that implied them (given additional non-truth-theoretic principles) would be the theory containing all instances of the T-schema. But this class of sentences is obviously not in good standing, for it leads to triviality. Could some other disquotational theory do the job?

The short answer is yes, trivially. Recall that McGee's trick entails that every sentence of $\mathscr{L}_\mathrm{T}$ is provably equivalent to an instance of the T-schema in the base theory. Thus, for every truth theory, whether compositional, Tarskian, disquotational, or else, there is a purely disquotational theory that proves the

same theorems. A fortiori, there is a disquotational theory that has exactly the same consequences as KF (even without any gap bridging principles). However, since these theories are otherwise highly unmotivated, we have little reason to believe that they are themselves in good standing.

Perhaps more interestingly, as Horsten & Leigh [19] have shown, the axioms of KF can be derived by iterating GBP twice over the theory PTB, which extends the base theory with an instance of local disquotation for each sentence of $\mathscr{L}_{\mathrm{T}}$ in which the truth predicate occurs only positively – i.e. under the scope of an even number of negations. However, whether this theory is in good standing is rather doubtful. Restricting the T-schema to positive instances is quite *ad hoc*. It isn't based on any well-motivated criterion of what an acceptable instance is, but merely on the observation that the liar sentence and other paradoxical expressions aren't positive. Just like positive set theory, which avoids Russell's paradox by restricting comprehension to positive instances, this leads to a mathematically interesting theory, but to a rather strange picture of truth (sets).[11] Of course, one could justify PTB by pointing out that its axioms are derivable from KF, as Halbach [15] observes, but this is of little use in the present context. Overall, we have little reason to believe that KF qualifies as a deflationary theory of truth.

Similar considerations apply to FS, though in this case one can actually give positive reasons to reject it. FS is the classical theory extending Σ with uniform disquotation for T-free atomic expressions and compositional axioms for the connectives and the quantifiers just like CT's, except the restriction to T-free sentences is lifted. Additionally, FS contains two "meta"-rules of inference that allow us to attach the truth predicate to and remove it from every theorem of the theory. As is well known, FS is ω-inconsistent, i.e. it proves all instances $\varphi(t)$ of a formula $\varphi(x)$ but, at the same time, it also entails $\neg\forall x\,\varphi(x)$. Thus, FS is in a sense unsound, as is every disquotational theory that entails the axioms of FS (with or without additional non-truth-theoretic principles). So no such disquotational theory appears to be in good standing.

In general, we are suspicious that *classical* theories containing unrestricted compositional axioms – i.e. axioms applying to *all* sentences of the language, including those with the truth predicate – can be shown to follow from some insubstantial disquotational theory together with additional non-truth-theoretic principles. In most cases, the only disquotational theories that come to mind here are those obtained by McGee's trick, for which it is quite doubtful that they are in good standing. Thus, as far as classical type-free theories are concerned, it would seem to be more promising to search for systems that restrict disquotational or compositional principles to a proper subclass of sentences of the language of truth, such as e.g. the grounded ones.[12] It is no coincidence that the theories of truth proposed by the authors, e.g. Picollo [25] or Schindler [31], have gone in that direction. In this respect, non-classical theories might

[11] See Schindler [32, pp. 398-399] for further arguments against PTB.
[12] See Schindler [33, sec. 3-4] for further discussion.

be at an advantage, insofar as they might have all instances of disquotation at their disposal, though this requires some further investigation.

Acknowledgements. A first draft of this paper was written while Thomas Schindler received support from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 792202) within the project *The Logical Function of Property Talk* (LOFUPRO). The final version was written while Thomas Schindler received support from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 803684) within the project *Truth and Semantics* (TRUST). In addition, both authors acknowledge the support of the German Research Council (DFG, Deutsche Forschungsgemeinschaft) within the project "Reference patterns of paradox" (GZ: PI 1294/1-1).

References

[1] Armour-Garb, B., and Beall, J. C. Minimalism, epistemicism, and paradox. In *Deflationism and Paradox*, B. Armour-Garb and J. C. Beall, Eds. Oxford University Press, 2005, pp. 85–96.

[2] Azzouni, J. Truth via anaphorically unrestricted quantifiers. *Journal of Philosphical Logic 30* (2001), 329–354.

[3] Cieslinski, C. *The epistemic lightness of truth. Deflationism and its logic.* Cambridge University Press, Cambridge, 2017.

[4] Field, H. Deflating the Conservativeness Argument. *Journal of Philosophy 96* (1999), 533–540.

[5] Field, H. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic 32* (2003), 139–177.

[6] Field, H. *Saving truth from paradox.* Oxford University Press, New York, 2008.

[7] Fischer, M. Truth and Speed-up. *Review of Symbolic Logic 7* (2014), 319–40.

[8] Frege, G. The thought: a logical inquiry. *Mind*, 65 (1956), 289–311.

[9] Grover, D. L. Propositional Quantifiers. *Journal of Philosophical Logic 1* (1972), 111–136.

[10] Grover, D. L., Camp, J. L., and Belnap, N. D. A Prosentential Theory of Truth. *Philosophical Studies: An International Journal for Philosophy in the AnalyticTradition 27* (1975), 73–125.

[11] Gupta, A. Minimalism. *Philosophical Perspectives 7* (2000), 359–369.

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

[13] Halbach, V. Disquotational truth and analyticity. *Journal of Symbolic Logic 66* (2001), 1959–1973.

[14] Halbach, V. How innocent is deflationism? *Synthese 126* (2001), 167–194.

[15] Halbach, V. *Axiomatic Theories of Truth*, 2nd ed. Cambridge University Press, Cambridge, 2014.

[16] Halbach, V., and Horsten, L. The deflationist's axioms for truth. In *Deflationism and Paradox*, B. Armour-Garb and J. C. Beall, Eds. Oxford University Press, 2005.

[17] Horsten, L. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In *The many problems of realism*, P. Cortois, Ed., vol. 3 of *Studies in the general philosophy of science*. Tilburg University Press, 1995, pp. 173–187.

[18] Horsten, L. *The Tarskian Turn: Deflationism and Axiomatic Truth*. MIT Press, Cambridge, 2011.

[19] Horsten, L., and Leigh, G. Truth is simple. *Mind 126* (2017), 195–232.

[20] Horwich, P. *Truth*, second ed. Oxford University Press, 1998.

[21] Horwich, P. A minimalist critique of tarski on truth. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 75–84.

[22] Ketland, J. Deflationism and Tarski's paradise. *Mind 108* (1999), 69–94.

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

[24] McGee, V. Maximal consistent sets of instances of Tarski's schema. *Journal of Philosphical Logic 21* (1992), 235–241.

[25] Picollo, L. Reference and Truth. *Journal of Philosophical Logic* (2019), https://doi.org/10.1007/s10992–019–09525–9.

[26] Picollo, L., and Schindler, T. Disquotation and infinite conjunctions. *Erkenntnis 83* (2018), 899–928.

[27] Picollo, L., and Schindler, T. Deflationism and the function of truth. *Philosophical Perspectives 32* (2019), 326–351.

[28] Quine, W. V. O. *Philosophy of Logic*. Harvard University Press, 1970.

[29] Raatikainen, P. On Horwich's way out. *Analysis 65* (2005), 175–177.

[30] Ramsey, F. P. Facts and propositions. *Proceedings of the Aristotelian Society 7* (1927), 153–170.

[31] Schindler, T. Axioms for grounded truth. *Review of Symbolic Logic 7* (2014), 73–83.

[32] Schindler, T. A disquotational theory of truth as strong as Z_2^-. *Journal*

of Philosophical Logic, 44 (2015), 395–410.

[33] Schindler, T. A note on Horwich's notion of grounding. *Synthese 197* (2020), 2029–2038.

[34] Shapiro, S. Proof and Truth: Through Thick and Thin. *Journal of Philosophy 95* (1998), 493–521.

[35] Sheard, M. Truth, provability and naive criteria. In *Principles of Truth*, V. Halbach and L. Horsten, Eds. Hänsel-Hohenhausen, Frankfurt am Main, 2002, pp. 169–181.

[36] Soames, S. What is a theory of truth? *Journal of Philosophy 81* (1984), 411–429.

[37] Stoljar, D., and Damnjanovic, N. The deflationary theory of truth. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., fall 2014 ed. Metaphysics Research Lab, Stanford University, 2014.

[38] Tarski, A. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*. Clarendon Press, Oxford, 1935, pp. 152—278.