Fix, Express, Quantify
Disquotation after its logic

Carlo Nicolai*

Truth-theoretic deflationism holds that truth is simple, and yet that it can fulfil many useful logico-linguistic roles. In this respect it is a simple but ambitious theory. Deflationism focuses on *axioms* for truth: There is no reduction of the notion of truth to more fundamental ones such as sets or higher-order quantifiers. This feature of the theory led to a proliferation of technical studies broadly motivated by the deflationary view of truth.[1] In this paper I argue that the fundamental properties of reasonable, primitive truth predicates are incompatible with the core tenets of classical truth-theoretic deflationism. The label 'deflationism' can certainly be employed to characterize a cluster of formal and philosophical approaches that take truth to be primitive. However, this has little to do with the original aims of the deflationary *theory* of truth.

I will focus in particular on the following theses of classical deflationism:

> FIX: the meaning of 'is true' is fixed by the Tarski-biconditionals " 'A' is true if and only if A".
>
> EXPRESS: the purpose of truth is to express – in virtue of FIX – infinite conjunctions and disjunctions.
>
> QUANTIFY: the truth predicate is fundamentally a device to perform sentential quantification over pronominal variables.

There are clear links between the theses just introduced and the *loci classici* of truth-theoretic deflationism. FIX can be traced back to Frege (1918) and (Quine, 1970, §1). In its propositional version, it is also present in Ramsey (1927). Horwich (1998) is certainly the recent main reference for it.[2] EXPRESS and QUANTIFY are deeply intertwined, and I shall treat them in this way below.

*I would like to thank: the participants to the Salzburg's Workshop 'New perspectives on truth and deflationism', the participants to the King's Staff Seminar, Alex Franklin, Johannes Stern, Volker Halbach. Special thanks go to Thomas Schindler for detailed comments on a previous draft.

[1] The monographs Halbach (2014), Horsten (2012), and Cieśliński (2017) contain detailed overviews of such studies and results.

[2] Horwich's position is somewhat more sophisticated: the meaning of truth is fixed by *our commitment* to (T). Since my discussion of FIX will not rely on such fine distinctions, I will not consider them in what follows.

The role of the truth predicate as a device to express infinite conjunctions (and disjunctions) – by allowing quantification on nominalized sentences – has been forcefully proposed by (Quine, 1990, §33); a formal rendering of Quine's claim has been put forth in Halbach (1999) – see also Heck (2005) for a discussion.

The main aim of the paper is to establish that these key theses of truth-theoretic deflationism are untenable. In particular, I will argue:

(i) that FIX, in one of its most plausible readings, leads to the adoption of dialetheism, and that deflationism shouldn't be tied to such nonclassical option;

(ii) that the combination of EXPRESS and QUANTIFY leads to the claim that an infinite conjunction and the assertion 'all conjuncts are true' should be equivalent in a strong sense. But they cannot be;[3]

(iii) that even if one considers QUANTIFY in isolation, the claim that the truth predicate fulfils the theoretical role of higher-order quantification in a first-order setting is highly dubious.

Before moving on to the details of my critique, let me emphasize that there is a further key deflationary thesis that I will not discuss in the paper:

> EXPLAIN: truth does not play a substantial role in philosophical and scientific explanations.

EXPLAIN is perhaps the most discussed deflationary tenet in the literature. This is mostly due to its translation into a precise formal claim (Shapiro, 1998; Ketland, 1999; Cieśliński, 2017): the deflationist's truth predicate, when added to a base theory B, should not be able to establish non-semantic facts about B that aren't already available in B itself. These non-semantic facts can be understood as theorems in the language of the base theory B. On this understanding, deflationary theories of truth ought to be conservative extensions of their base theories.

In this paper I will not discuss EXPLAIN, although I occasionally appeal to it; I have already extensively discussed elsewhere its nature and scope. I believe that its understanding in terms of the conservativeness of the theory of truth over the base theory is both too narrow, and also not required by deflationism. This latter claim is not at all original; similar points are made for instance by Halbach (2001), Horsten (2012), and more recently by Picollo and Schindler (2019). The former claim, by contrast, is developed for instance in Nicolai (2015); there it is argued that a theory of truth should be understood as a package containing

[3] As it shall be clear later on, this conclusion was already reached by Gupta (1993). I provide new formal results to corroborate Gupta's diagnosis.

a formal syntax and truth axioms, which can be added to *any* base theory, regardless of their strength or capability of formalizing syntax. In that setting, the conservativeness of the theory of truth becomes a trivial property of any theory of truth. However, the theory of truth *may, or may not* be reducible – in the precise sense of a relative interpretation – to the base theory. Actually, in many natural cases, the theory of truth will at least interpret the consistency assertion for the base theory, even in the case in which the truth predicate is not allowed in the non-logical schemata of the base theory (Nicolai, 2016). And the fact that a theory of truth may not be interpretable over a base theory seems entirely compatible with the kind of meta-theoretic explanations that no deflationist would want to rule out.

1 Fix

According to FIX, the meaning of 'is true' is fixed by the T-sentences

(T) 'A' is true if and only if A

where A is a sentence of English. Therefore, FIX is first and foremost an attribution of meaning to certain expressions containing the predicate 'is true'. In particular, it's a thesis about the meaning of ascriptions of truth to certain linguistic objects, sentence types in particular. Moreover, its core principle (T) has the logical structure of a *biconditional*. It is this latter feature of FIX that will be the central theme of this section. However, I will briefly consider the former as well to provide an adequate context to the discussion.

Of course (T) cannot possibly be right due to the Liar paradox. Horwich proposes to consider only non-problematic instances of it (Horwich, 1998, pp. 40-42). This isn't a trivial task: McGee (1992) has shown that there are uncountably many incompatible and maximally consistent sets of instances of (T). Alternatively, one could understand the 'if and only if' in (T) as a non-classical biconditional. As I shall argue more extensively later, I do not think that deflationism is (or should be) committed to a revision of logic. Therefore, in what follows I will stick with theories formulated in classical logic and consider a rather drastic restriction of (T) that is nonetheless sufficiently plausible to be compatible with different flavours of deflationism.

By assuming that the objects of truth are sentences of English, I deliberately depart from Horwich's minimalism.[4] However, my hope is that everything I am going to say will apply even if a suitable theory of (structured) propositions is employed as a theory of the objects of truth. Sentence types have an obvious

[4]Although minimalism is not entirely dependent on the choice of propositions as opposed to certain classes of sentences (Horwich, 1998, §2.1).

advantage for my purposes. I am going to rely on formal results on axiomatizations of the truth predicate, and while there are precise and widely accepted axiomatizations of sentence types – possibly obtained via coding –, the same cannot be said about propositions.

FIX is a thesis about the meaning of 'is true'. It is clear that disquotationalism cannot be at ease with classical truth-conditional semantics, in which the meaning of an expression is given primarily in terms of its truth-conditions as individuated by *that*-clauses.[5] Alternative approaches include verificationist theories, use theories, conceptual-role theories and variants thereof. For my purposes it is not necessary to settle precisely for one of these views. What will be important is that, whatever account of meaning one goes for, it is part of this account that sentences with the same meaning display a form of equivalence in their inferential behaviour. A more precise form that this equivalence might take will be considered shortly.

I now turn to the main topic of this section, namely the way in which the biconditional in (T) should be understood. My first claim is that (T) cannot be taken to express a simple material equivalence. To argue for it, I need to fix some notation. I work over a base language $\mathcal{L}$ that is capable of talking about its syntax. The language of Peano Arithmetic $\mathcal{L}_{\mathbb{N}}$ is an obvious choice,[6] but also the language of set theory or a theory of expressions would work. In practice, I work over an axiomatization of Peano Arithmetic. The language $\mathcal{L}_{\mathtt{Tr}}$ is simply $\mathcal{L}_{\mathbb{N}}$ expanded with a truth predicate $\mathtt{Tr}$.

Since the deflationist needs to focus on non-problematic instances of (T), I restrict my attention to a specific set of biconditionals. Due to complexity considerations, I do not want to appeal to a semantic classification of 'pathological' sentences.[7] Therefore I consider a syntactic restriction on the sentences appearing in the biconditionals: since the role of negation (or equivalent logical tools) is fundamental in the Liar paradox, I restrict my attention of *positive* sentences of $\mathcal{L}_{\mathtt{Tr}}$, that is sentences in which the truth predicate appears only in the scope of an even number of negations (Halbach, 2009). The rationale behind this choice is the following: by accepting a restriction of (T), one accepts that not for all sentences of $\mathcal{L}_{\mathtt{Tr}}$ the schema holds. It is only this asymmetry that is essential to the argumentation below, and not the nature or details of this restriction. Therefore, by choosing a particularly simple but comprehensive set of

[5] However, there are some deflationists who argue for the compatibility of deflationism and truth-conditional semantics (Williams, 1999).

[6] By the language of Peano arithmetic I mean the language with signature $\{0, 1, +, \times\}$ plus finitely many symbols for primitive recursive functions to render the formalization of syntax easier. On occasion I will take $\mathcal{L}_{\mathbb{N}}$ to be formulated in a relational signature.

[7] I take that a recursively enumerable set of axioms is the standard for disquotational theories based on simple axioms for truth. Virtually any semantically-based classification of safe v paradoxical sentences results in sets that are of complexity Π^1_1 or beyond (see e.g. Kripke (1975)).

Tr-biconditionals, I aim to show that *any* plausible restriction strategy may lead to problems. To be clear, I am not advocating the set (PT) of Tr-biconditionals as *sufficient* for deflationism; I only claim that it may be plausibly considered as a component of any adequate version that involves a self-referential truth predicate.

For my purposes, it is useful to employ a slightly different definition of the positive fragment of $\mathcal{L}_{\mathtt{Tr}}$. Following Horsten and Leigh (2017), I consider a *negation-free* language $\mathcal{L}^+$ – whose logical primitives are $\vee, \wedge, \exists, \forall$ – in which for every atomic predicate $P \in \mathcal{L}$ there is a dual predicate $\bar{P}$. The dual of Tr is then denoted with F. The duality of atomic predicates transfers to connectives and quantifiers: the dual of $\wedge$ is $\vee$, the dual of $\forall$ is $\exists$, and vice versa. The restricted sets of Tr-biconditionals I focus on is then, for any $A \in \mathcal{L}^+$:

$$\text{(PT)} \qquad \mathtt{Tr}\ulcorner A \urcorner \leftrightarrow A, \qquad\qquad \mathtt{F}\ulcorner \bar{A} \urcorner \leftrightarrow A,$$

where $\bar{A}$ is obtained from A by replacing every predicate, connective, and quantifier with its dual.

It is well-known that the disquotation schema (PT) – and virtually any other schematic presentation of the Tr-biconditionals – is not able to establish desirable general claims. Quine famously wrote:

> The harder sort of generalization is illustrated by generalization on the clause "time flies" in "If time flies then time flies". We want to say that this compound continues true when the clause is supplanted by any other; and we can do no better than to say just that in so many words, including the word 'true'. We say "All sentences of the form 'If p then p' are true." (Quine, 1990, p. 80)

Besides generalizations involving logical laws such as the ones considered by Quine, crucial general facts about truth are the so-called *compositional* principles such as

$$(\wedge) \qquad \forall \varphi, \forall \psi \in \mathcal{L}_{\mathbb{N}} \left(\mathtt{Tr}(\varphi \wedge \psi) \leftrightarrow \mathtt{Tr}\varphi \wedge \mathtt{Tr}\psi\right).$$

($\wedge$) generalizes only over sentences not containing the truth predicate – the quantified claim $\forall \varphi A(\varphi)$ is intended to abbreviate $\forall x(\mathrm{Sent}_{\mathcal{L}_{\mathbb{N}}}(x) \rightarrow A(x))$, similarly for the existential quantifier – and yet neither it nor *any* plausible truth-theoretic generalization on logical laws can be handled by the schema (PT) (Halbach, 2009, lemma 6.1): every truth theoretic generalizations that can be established by means of (PT) is bounded by a finite natural number n and it is therefore only a *finite* generalization.

(Horwich, 1998, Postscript, §5) tried to react to similar observations by

noting that since he's concerned with propositions and not with sentences, a non-logical principle of the form 'if some property P holds of *each* proposition, then it holds of all propositions' may be more justified. This is clearly a principle that cannot be derived from simple disquotation when sentences are at stake. This directly follows by Halbach's observation. For instance, (PT) entails

$$\mathtt{Tr}(A \wedge B) \leftrightarrow \mathtt{Tr}\,A \wedge \mathtt{Tr}\,B$$

for any $A, B \in \mathcal{L}$, but this is not enough to obtain the universally quantified claim ($\wedge$). It is important to add that this fact is purely structural: it concerns logical the properties of quantifiers in first-order theories, which cannot fix a standard interpretation – in the schemata, $A, B, \ldots$ range over 'standard' sentences, φ, ψ over possibly non-standard ones. Therefore, especially if propositions must share some structural features with sentences, the principle of generalization advocated by Horwich seems unjustified from the basis of disquotation alone.

Another – and perhaps more plausible – reaction to the deductive weakness of disquotation may be to give up any hope of establishing general laws from schematic principles, and require a more subtle form of equivalence. This is sensible: after all it's clear that a universal claim can never be logically equivalent to its instances. One might therefore require that schematic truth principles and truth-theoretic generalizations to be conceptually equivalent, or better equivalent *for all theoretical purposes*. We will see in §2.2 that this is also out of reach for the deflationist.

The schematic nature of the $\mathtt{Tr}$-biconditionals reveals the deductive weakness of material readings of FIX.[8] This would be a problem if deflationism did not admit the possibility of any additional logical tool. There is, however, a natural way of extending (PT) that is in fact based on a constitutive feature of any deflationary view.

1.1 Modality

The equivalence between A and 'A is true' is not material. However, a correct reading of FIX is certainly compatible with assigning to the $\mathtt{Tr}$-biconditionals a certain intensional status. This idea is not new. Several authors have articulated it in slightly different ways. Hartry Field proposes to understand the equivalence involved in the $\mathtt{Tr}$-biconditionals as a form of cognitive equivalence that, as such, naturally calls for a modal strengthening (Field, 1994, §6). Similarly, (Horwich,

[8]There is also another related argument in support of the claim that the equivalence of A and $\mathtt{Tr}\,A$ should not be material. It is due to Gupta (1993), and it is based on the idea that to be able to perform its quantificational role, the bi-conditional should express a form of synonymy. I will consider this option in detail in §2.

1998, §3, fn. 5) suggests a natural extension of the schema (T) to accommodate blind ascriptions of modal nature and the interaction between disquotational truth and alethic modalities. Another, and more precisely spelled out form of modal disquotationalism is defended in Halbach (2003); Halbach considers a 'typed' version of the schema (T), that is a version where the sentence A cannot contain the truth predicate, and shows how a modal strengthening of this schema – based on the notion of *truth-analyticity* – yields compositional laws of the form ($\wedge$).

For my purposes it is sufficient to give a general structural account of the kind of intensionality involved, without committing myself to a specific choice such as conceptual necessity or analyticity. What I only require is that the modality in question satisfies a few basic properties. The sort of modality that I consider will be formalized as a predicate applying to names of sentences, and not as a sentential operator. This is mainly because it is natural to think that the kind of necessity involved in the analysis of FIX should be used, if required, to state obvious general claims such as 'what is necessary is also true', also for sentences that we do not remember or cannot (presently) name. I express this modality via a unary predicate $\Box(x)$, where x stands for a suitable name of a sentence in the language of truth. I call $\mathcal{L}_{\mathtt{T}}^{\Box}$ the language $\mathcal{L}_{\mathtt{Tr}} \cup \{\Box\}$. An attractive structural account of the modal status of disquotation arises naturally from a generalization of the theory presented in Halbach (2003). I will now turn to describe the core principles of the *generalized Halbach's account*.[9]

The first condition on $\Box$ is that it should be closed under predicate (classical) logic. I take this as a harmless requirement that is shared by any plausible alethic modality. This requirement is spelled out more precisely by splitting the condition in two. On the one hand one requires that all theorems of predicate logic in the full language are boxed:[10]

(log1) $$\forall \varphi \in \mathcal{L}_{\mathtt{T}}^{\Box}(\mathtt{Prov}_{\mathtt{fol}}(\varphi) \rightarrow \Box\varphi).$$

Here $\mathtt{Prov}_{\mathtt{fol}}(\cdot)$ is a canonical provability predicate for first-order logic in $\mathcal{L}_{\mathtt{T}}^{\Box}$, that can be standardly constructed in Peano Arithmetic. On the other, one requires a closure condition that is reminiscent of the modal axiom K:

(log2) $$\forall \varphi, \psi \in \mathcal{L}_{\mathtt{T}}^{\Box}(\Box(\varphi \rightarrow \psi) \wedge \Box\varphi \rightarrow \Box\psi).$$

[9] Here I use the qualification 'generalized' because Halbach mostly deals with *typed* principles for truth and necessity. The theory I present is the natural generalization to type-free concepts of Halbach's account.

[10] On a reading of the box in terms of truth-analyticity, the closure of the box under classical logic may seem unnecessarily strong. However, the inconsistency considered below can equally arise if only sentences of $\mathcal{L}_{\mathtt{Tr}}$ are considered to be closed under classical logic.

The core principle of the generalized Halbach account states that all instances of the schema (PT), that is $\mathtt{Tr}$-biconditionals for positive formulas, are modalized:

(mpt) $$\forall \varphi \in \mathcal{L}^{+} \Box(\mathtt{Tr}\dot{\varphi} \leftrightarrow \varphi)$$

In (mpt), the dot represents the function that sends a formula to its name – or, since our discussion is framed in arithmetic, a function that sends a number to the code of its numeral.[11]

Finally, we require that proofs from our axioms are boxed and that our modality is factive. This therefore translates into a necessitation rule and a factivity condition:

(nec) if A is a consequence of the theory in $\mathcal{L}_{\mathtt{Tr}}$, then also $\Box A$ is;

(fact) $\forall x(\Box A(x) \rightarrow A(x))$, for all formulas $A(v)$ of $\mathcal{L}_{\mathtt{Tr}}$.

Also principles (mpt), (nec), (fact) seem straightforwardly compatible with an understanding of $\Box$ as truth-analyticity (Halbach, 2003), or conceptual necessity (Field, 1994, §3). The fundamental intuition behind them is that truths about the syntactic structure of the language – modulo isomorphism with the structure of natural numbers – are necessary, and so are the disquotation principles. Of course if one extended the theory with contingent vocabulary (and axioms), then the necessitation principle (nec) would have to be restricted to the 'rigid' part of the language. Since for our purposes it is sufficient to isolate necessary conditions for the modal status of deflationary principles, I will not discuss such extensions. I call MPT the $\mathcal{L}^{\Box}_{\mathtt{Tr}}$-theory PA+(log1)-(fact).

The restrictions to the applicability of the schemata are in place to avoid paradox. Some of them are not necessary: by carefully defining the sublanguages of $\mathcal{L}^{\Box}_{\mathtt{Tr}}$ – even the codified versions of them – one might formulate slightly more general principles. Since we aim to a negative result, it will suffice to consider the present version of MPT.

FACT 1 (Halbach, 2003, Thm. 4). MPT *is consistent.*

The intensional equivalence between A and ' "A" is true' articulated by Halbach's generalized theory MPT overcomes the weakness of the material reading

[11] To be precise (mpt) is an intensional formulation of

(1) $$\forall x(\mathtt{Sent}_{\mathcal{L}^{+}}(x) \rightarrow \Box(\mathtt{eq}(\mathtt{sub}(\ulcorner \mathtt{Tr}\, v \urcorner, \ulcorner v \urcorner, \mathtt{num}(x)), x)))$$

where $\mathtt{sub}$ is the arithmetical substitution function such that $\mathtt{sub}(\ulcorner \varphi(v) \urcorner, \ulcorner v \urcorner, \ulcorner t \urcorner) = \ulcorner \varphi(t) \urcorner$, $\mathtt{num}$ is the numeral function such that $\mathtt{num}(n) = \ulcorner \mathtt{S} \underbrace{\ldots}_{n\text{-times}} \mathtt{S0} \urcorner$, and $\mathtt{eq}(x, y)$ sends the codes of two formulas to the code of their biconditional. Similar considerations apply to (fact) below.

of the $\mathtt{Tr}$-biconditional. Many general claims that were not available before are now consequences of the theory. For instance, the principle $(\wedge)$ for $\mathcal{L}^+$ now follows from the provability of[12]

$$\forall \varphi, \psi \in \mathcal{L}^+ \, \Box\big(\mathtt{Tr}(\varphi \mathbin{\dot{\wedge}} \psi) \leftrightarrow \mathtt{Tr}\dot{\varphi} \wedge \mathtt{Tr}\dot{\psi}\big) \tag{5}$$

and one application of (`fact`). A similar argument enables one to obtain in MPT the useful principles for truth ascriptions:[13]

$$\forall t \, (\mathtt{Tr}(\mathtt{Tr}t) \leftrightarrow \mathtt{Tr}t) \tag{6}$$

$$\forall t \, (\mathtt{Tr}(\mathtt{F}t) \leftrightarrow \mathtt{F}t) \tag{7}$$

$$\forall t \, (\mathtt{F}(\mathtt{Tr}t) \leftrightarrow \mathtt{F}t) \tag{8}$$

$$\forall t \, (\mathtt{F}(\mathtt{F}t) \leftrightarrow \mathtt{Tr}t) \tag{9}$$

I will not provide the proofs of these claims since the techniques are well-known (Halbach, 2001; Horsten and Leigh, 2017).

I conclude this section with an important disclaimer. By endorsing the modal status of disquotation, I do not want to claim that this is a *sufficient* formal analysis of FIX. In fact, in the following we will consider good reasons for requiring an even stronger form of equivalence between the two sides of the $\mathtt{Tr}$-biconditionals. For the sake of my argument it is sufficient to hold that the modal status of the $\mathtt{Tr}$-biconditional as articulated in MPT is necessary to an adequate analysis of FIX. This does not rule out that the principles of MPT may follow from a set of stronger principles articulating a stricter analysis of FIX.

1.2 An inconsistency

In this section I will argue for two main claims: the first is that MPT does not adequately capture the interplay between truth and modality that one would expect in such an expressive framework. The second claim is that, once this natural interplay is allowed, an inconsistency arises.

[12] In particular, (5) is obtained by the following theorems of MPT:

$$\forall \varphi, \psi \in \mathcal{L}^+ \, \Box\big(\mathtt{Tr}(\varphi \mathbin{\dot{\wedge}} \psi) \leftrightarrow \varphi \wedge \psi\big), \tag{2}$$

$$\forall \varphi \in \mathcal{L}^+ \, \Box(\mathtt{Tr}\dot{\varphi} \leftrightarrow \varphi), \tag{3}$$

$$\forall \psi \in \mathcal{L}^+ \, \Box(\mathtt{Tr}\dot{\psi} \leftrightarrow \psi). \tag{4}$$

[13] The principles that follow are in fact abbreviations. For instance, (6) is an abbreviation of the longer:

$$\forall x(\mathrm{Cterm}_{\mathcal{L}_\mathbb{N}}(x) \rightarrow (\mathtt{Tr}\dot{\mathtt{Tr}}x \leftrightarrow \mathtt{Tr}\,\mathtt{val}(x))$$

where $\mathrm{Cterm}_{\mathcal{L}_\mathbb{N}}(x)$ is the predicate representing in $\mathcal{L}_\mathbb{N}$ the set of its closed terms, $\dot{\mathtt{Tr}}$ the function representing the syntactic operation $(\ulcorner \mathtt{Tr} \urcorner, \ulcorner t \urcorner) \mapsto \ulcorner \mathtt{Tr}t \urcorner$, and $\mathtt{val}(x)$ the arithmetical evaluation function.

The factivity principle (fact) can be straightforwardly paraphrased as 'if a sentence is necessary/analytic to truth, then it's true'. Its schematic formulation of factivity principles is the standard choice in modal or epistemic logic when the modality in question is given in the form of a sentential operator. But a schematic formulation is also required in the case of an arithmetical language when one reads $\Box$ as provability and formulates proof-theoretic reflection principles. In such cases, $\Box$ is not a primitive anymore, but provability becomes definable and its fundamental properties provable in Peano Arithmetic or in any reasonable base theory. Yet, by the arithmetical undefinability of truth, one is forced towards a schematic formulation.

However, when a truth predicate is around, it is natural to formulate factivity principles as object-linguistic statements. And this is especially so for the deflationist. EXPRESS and QUANTIFY dictate that the truth predicate is there precisely to offer finitary versions of infinite lots of sentences such as the one described by (fact). An obvious advantage of a finite formulation is to avoid quantification in the metalanguage proper of schematic formulations; this translates in the possibility of analyzing in our language the negation of such general principles.[14]

As a consequence, one should formulate MPT by replacing (fact) with the single $\mathcal{L}_{\mathrm{T}}^{\Box}$-sentence

(tfact) $$\forall \varphi \in \mathcal{L}_{\mathtt{Tr}} (\Box \varphi \rightarrow \mathtt{Tr} \varphi).$$

Unfortunately, this is still not completely satisfactory. The truth predicate of MPT can only deal with sentences of the language $\mathcal{L}^{+}$. This is crucially required to retain consistency via a sound – although arguably incomplete – restriction of the schema (T). But the necessitation principle (nec) introduces under the scope of the $\Box$ sentences of a different language: for instance, since MPT is a classical theory in $\mathcal{L}_{\mathtt{Tr}}$, the sentence $\neg \mathtt{Tr} l \vee \mathtt{Tr} l$ will be an $\mathcal{L}_{\mathtt{Tr}}$-theorem of MPT – where $\neg \mathtt{Tr} l$ is a liar sentence with l a closed term such that, by the diagonal lemma, $l = \ulcorner \neg \mathtt{Tr} l \urcorner$. Therefore, so will be $\Box(\neg \mathtt{Tr} l \vee \mathtt{Tr} l)$. But $\neg \mathtt{Tr} l$ is a *negative* sentence, and the truth predicate of MPT has nothing to say about these sentences: they simply belong to a different language. Similarly, and perhaps more importantly, negative sentences such as $\neg \mathtt{Tr} \ulcorner 0 \neq 0 \urcorner$ are consequences of MPT. And so is $\Box(\neg \mathtt{Tr} \ulcorner 0 \neq 0 \urcorner)$, by (nec). By adopting (tfact), therefore, one would be able to conclude $\mathtt{Tr} \ulcorner \neg \mathtt{Tr} \ulcorner 0 \neq 0 \urcorner \urcorner$. However, the truth predicate of

[14] Somewhat different arguments to prefer a truth-theoretic formulation of (fact) are given by Leon Horsten and Johannes Stern (Stern, 2016). Horsten considers the notion of truth to be both the primary source of paradox; Stern adds to it that it is the main tool to quote and disquote. On this view, any principle involving disquotation – e.g. (fact) – should be formulated by means of a truth predicate.

MPT can only manipulate sentences of $\mathcal{L}^+$. Even with respect to of innocuous sentences such as $\neg\mathtt{Tr}\ulcorner 0 \neq 0\urcorner$, it is powerless. This is not ideal.

A slight modification of (tfact) will however deliver a more palatable principle. It will be based on the idea that there is a natural mapping of the language $\mathcal{L}_{\mathtt{Tr}}$ into the language $\mathcal{L}^+$ that essentially replaces negative occurrences of truth predicates with with the falsity predicate F of $\mathcal{L}^+$ – the details of the translation are given in Appendix A. I denote with φ^* the $\mathcal{L}^+$-sentence resulting from the translation of the $\mathcal{L}_{\mathtt{Tr}}$-sentence φ. The required modification of (tfact) amounts to

(tfact*) $\qquad\qquad \forall\varphi \in \mathcal{L}_{\mathtt{Tr}}\,(\Box\varphi \rightarrow \mathtt{Tr}\varphi^*)$

I call MPT* the variant of MPT in which (fact) is replaced with (tfact*).[15] It is clear that (tfact*) deals satisfactorily with negative theorems of MPT such as $\neg\mathtt{Tr}\ulcorner 0 \neq 0\urcorner$. For instance, we can easily unravel its chain of truth-ascriptions and go from $(\neg\mathtt{Tr}\ulcorner 0 \neq 0\urcorner)^*$ – i.e. $\mathtt{F}\ulcorner 0 \neq 0\urcorner$ – to its non-semantic basis $0 = 0$. I take MPT* to be a more satisfactory modal strengthening of (PT) than MPT, and also more faithful to the deflationist's tenets FIX, EXPRESS, and QUANTIFY.

The generalized Halbach's approach, once modified in the way prescribed, reveals a fundamental problem. The proof of the following result in given in Appendix A:

PROPOSITION 1. *The truth predicate of* MPT* *is inconsistent.*

MPT* is both natural and fruitful, in that it overcomes the deductive weakness of material renderings of FIX. However, its truth predicate becomes *inconsistent*; there are sentences of $\mathcal{L}^+$ that the deflationist's truth predicate deems both true and false. This, I take, is against the spirit of deflationism. First of all, several proponents of FIX, also in its modal rendering, explicitly reject dialetheism as a way to deal with the paradoxicality of the schema (T) – see again (Field, 1994; Halbach, 2003). Moreover, it is a core feature of a disquotational truth predicate that its fundamental expressive role should be *neutral* with respect to the underlying logical principles. Horwich, for instance, writes about his minimalism:

> . . . a central tenet of the point of view advance here is that the theory of truth and the theory of logic have nothing to do with one another. (Horwich, 1998, §24)

[15] It's important to notice that the translation $(\cdot)^*$ is natural and uncontroversial. It is for instance the translation that one would employ to translate the truth predicate of standard, type-free theories of truth formulated in $\mathcal{L}_{\mathtt{Tr}}$ with a partial interpretation of the truth predicate, such as the well-known Kripke-Feferman theory KF, into its equivalent version with truth and falsity predicates.

Horwich's basic idea seems to be that the theory of truth should deliver universally quantified versions of the logical rules that one accepts, or at least as many of them as possible. We have already seen that, in order to avoid paradox, some of the classical laws involving negation may need to be dropped at the level of the internal logic of the truth predicate. What seems reasonable to take from Horwich's passage – and surrounding discussion – is that, whereas it might be acceptable that certain general claims of logical character such as

'every sentence is either true or false'

may not be provable in the theory of truth in their unrestricted form (although $A \vee \neg A$ is a theorem), one should not obtain truth-theoretic consequences of logical nature that violate the basic underlying principles. But this, as I have shown, is exactly what happens in the case the proponent of modal deflationism.[16]

A final word about the scope of the argument given. I have focuses on what I believe is the best available framework articulating the modal status of FIX: what I called the generalized Halbach approach. Of course this does rule out the possibility of finding a better modal rendering of FIX that is not leading to inconsistencies. While this is of course possible, the strategy outlined above is likely to generalize to any framework that introduces a discrepancy between the logical principles assumed in the formulation of the theory, and the logical principles that are valid under the scope of the truth predicate. In addition, there is a more general source of discontent with the modal analysis of FIX. Although it seems to be necessary to disquotational truth, it also does not seem to be sufficient. It is this aspect of the deflationary picture that will be analyzed in the following sections.

2 Express, quantify

The links between EXPRESS and QUANTIFY are clear if one looks at the canonical examples appearing in the deflationist canon. In Quine's passage on p. 5, the

[16] It may be objected, of course, that the reduction strategy I am offering may instead simply be additional evidence for the inconsistency of the concept of truth associated with the truth predicate. And inconsistency in this context may be understood in two ways: as a dialetheist account of truth, or as an inconsistency view of the *ordinary concept of truth*. The remarks above are directed to the former option. As to the latter option, this is simply not what I am after: I explicitly started with a *consistent* set of bi-conditionals, and show that even starting with such a restricted set one ends up with an inconsistency. So this is different from accepting *all* instances of the schema (T) and learning to live up with their inconsistency (Azzouni, 2006, ch. 5).

infinite conjunction

(10) (snow is white → snow is white) ∧
(grass is green → grass is green) ∧ ...

can be taken to be equivalent, via FIX, to

(11) ($\mathtt{Tr}\ulcorner$snow is white → snow is white$\urcorner$) ∧
($\mathtt{Tr}\ulcorner$grass is green → grass is green$\urcorner$) ∧ ...

The sentences between corners are now terms standing for names of sentences. The conjunction of EXPRESS and QUANTIFY then should enable one to conclude:

(12) for all sentences σ: $\sigma \rightarrow \sigma$ is true.

There are different options to understand the relationship between (10) and (12), or more generally between an infinite lot of sentences and the truth theoretic general claim that is intended to 'express' it. One hasty thought to make this idea precise would be to compare sets of sentences definable in $\mathcal{L}_{\mathtt{Tr}}$ and the ones definable in suitable infinitary languages, such as fragments of $\mathcal{L}_{\omega_1\omega}$ that allow only for certain infinite conjunctions and disjunctions.[17] Such approach would then explicate a core thesis of disquotationalism via the crucial contribution of semantic truth, so it is safe to say that it is a non-starter.

An approach more in line with the disquotationalist's assumptions is given by Halbach (1999); he shows that any adequate base theory T such as Peano Arithmetic extended with the 'infinite conjunction' $\{A(\ulcorner B\urcorner) \rightarrow B \mid B \text{ a sentence of } \mathcal{L}_T\}$ proves the same theorems without the truth predicate as T plus (i) the set of biconditionals $\mathtt{Tr}\ulcorner B\urcorner \leftrightarrow B$, for B truth-free, and (ii) the single sentence $\forall x(A(x) \rightarrow \mathtt{Tr}(x))$. A similar result holds for disjunctions. The claim is then that this results captures faithfully the interplay of EXPRESS and QUANTIFY.[18]

This proposal has certainly the merit of articulating a clear and precise rendering of EXPRESS. However, it is also incomplete in many respects. In the

[17] For instance, if one considers an acceptable structure $\mathcal{M}$ and defines $\mathcal{L}^*_{\omega_1\omega}$ as the language that contains only disjunctions or conjunctions that are definable in $\mathcal{M}$, one obtains an unsurprising correspondence of the sets of $\mathcal{L}$-sentences definable in $\mathcal{L}_{\mathtt{Tr}}$ and $\mathcal{L}^*_{\omega_1\omega}$ – relative to $\mathcal{M}$.

[18] As Halbach writes:

> I take the result to be an exact formulation of the disquotationalist claim that infinite conjunctions can be expressed in a language containing a truth predicate which is characterized by the Tarskian equivalences. (Halbach, 1999, p. 14)

first place, it focuses on the mere material equivalence of A and $\texttt{Tr}\ulcorner A\urcorner$, which we have already seen to be insufficient in the previous section. Secondly, as shown in Heck (2005), the result breaks down when one considers the *joint* addition of infinite conjunctions and disjunctions, which may result in a *non*-conservative extension of T. Moreover, even if one only takes into account conjunctions or disjunctions separately, Halbach's result(s) can already be obtained from the principle $\texttt{Tr}\ulcorner A\urcorner \rightarrow A$, for A an *arbitrary* sentence of $\mathcal{L}_{\texttt{Tr}}$, a principle that is often called `T-out`.[19] Now any T augmented with `T-out` is consistent: one can for instance read `Tr` as 'is provable in T'. This seems to clearly indicate that the criterion overgenerates: there's nothing special about disquotational truth that enables one to perform what's required by EXPRESS.

There are, however, independent reasons to require a much stricter reading of EXPRESS and QUANTIFY. As already argued by Anil Gupta, EXPRESS should be understood as a thesis about the *sameness of meaning* of infinite conjunctions, such as (10), and their corresponding universally quantified sentence, (12) in the example (Gupta, 1993). The fundamental reason for this is that there are certain tasks that the disquotationalist's truth predicate is bound to perform, or features that it should possess, which simply cannot obtain if EXPRESS is read in a weaker way, such as material or necessary equivalence for instance.

One example of this is the status of certain law-like generalizations such as

(13) true beliefs tend to facilitate success.

To affirm their explanatory role without violating EXPLAIN, disquotationalists have argued that general claims such as (13) *only* express an infinite conjunction of simple facts such as (Horwich, 1998, pp. 22-23):

(14) Subject S wants X; S believes that by doing Y she will achieve X.

Now the `Tr`-biconditionals, suitably formulated, would explain why the truth of the belief that by doing Y, S achieves X, makes it more likely for S to achieve X. Therefore it would also explain this infinite conjunction. But, since EXPRESS prescribes that (14) is only a less concise way of affirming (13), the latter is also explained by the `Tr`-biconditionals. The obvious conclusion is that EXPRESS requires a form of equivalence between infinite conjunctions (disjunctions) and their truth-theoretic counterparts that preserves their status as explananda. This certainly fails for materially equivalent or necessary equivalent claims.[20]

[19] That Halbach's result only needs `T-out` is observed also in print, by Picollo and Schindler (2017).

[20] The example employed by Gupta is, for instance: 'Cicero is Tully' is necessarily true, at least according to a widespread position, and so is 'No chemical reaction will produce caustic soda from saltpeter and sulfuric acid'. However, this does not entail that by explaining one

An equally strong connection is envisaged by deflationists for logico-linguistic laws (Field, 1994, pp. 258-9). In the light of the conceptual/cognitive equivalence of the two sides of the Tr-biconditionals, the equivalence of truth-functional laws such as

(15) $A \vee B$ if and only if A or B

(16) $A \vee B$ is true if and only if A is true or B is true

(17) for all φ and ψ, $\varphi \vee \psi$ is true if and only if φ is true or ψ is true

should be a matter of conceptual necessity or analyticity. In addition, EXPRESS and EXPLAIN clearly rule out the possibility that the explanatory status of (17) can substantially differ from the explanatory status of (15). Of course, in both the examples considered, a severely deflated sense of explanation may dictate that a single sentence can be more explanatory than a schema, but I take that this is not what's at stake in EXPLAIN, especially if one combines it with EXPRESS and QUANTIFY. And it's also clear from the discussion above that there is no way of establishing the inter-derivability of schemata such as (15) and corresponing truth-theoretic generalizations, one has to look elsewhere to understand this relationship.

I propose to understand the purported equivalence of (10) and (12) – or of (15) and (17) – in terms of formal criteria of *conceptual equivalence* for theoretical concepts. In particular, in the next section I consider notions of theoretical equivalence widely employed to compare scientific concepts. I take this analysis to be a vindication of the deflationist's point that a infinite conjunction and the corresponding truth-theoretic generalization should be faithful to the congnitive or conceptual equivalence involved in the Tr-biconditionals, but also they should be on a par in their explanatory and theoretical status. If it turned out that, for instance, (10) and (12) are indeed equivalent in this sense, one would immediately obtain also conclusive evidence of their equivalent explanatory status, in the same way as the theoretical equivalence of two scientific concepts would entail their equivalent explanatory status.

However, I will show that once such notions of equivalence are in place, there is a clear verdict on the plausibility of the deflationist's claim. I will show that an infinite lot of sentences and the corresponding truth-theoretic generalization can *never* be theoretically equivalent in the required sense.

one also explains the other.

2.1 Sameness of meaning and theoretical equivalence

The notions of theoretical equivalence that I will employ are well-known. I will mainly focus on *biinterpretability and definitional equivalence (or synonymy)*, and on some variants of them. *Definitional equivalence* is certainly considered to be one of the best candidates to capture theoretical equivalence (Halvorson, 2012).[21] Bi-interpretability is a weaker notion than definitional equivalence, and will be mainly employed to strengthen some negative results.

Both notions can be defined in terms of relative interpretations.[22] I give precise definitions in Appendix B and keep here the discussion at a semi-formal level. Given first-order theories U and V, we say that they are *definitionally equivalent* if there are (relative) interpretations $K\colon U \to V$ and $L\colon V \to U$ that are *provably inverse* for all primitive concepts of the two theories: that is, such that U proves that $\forall\vec{x}(P_i(\vec{x})^{L\circ K} \leftrightarrow P_i(\vec{x}))$, that V proves $\forall\vec{x}(P_j(\vec{x})^{K\circ L} \leftrightarrow P_j(\vec{x}))$, for i ranging over U-primitives, and j over V-primitives, and in addition that the two domains – employed to relativize quantifiers – are provably equivalent in both theories.[23]

Biinterpretability can be defined in an analogous way; however, instead of requiring the material equivalence of the identity interpretation and the compositions of the two interpretations involved, one requires them to be *provably isomorphic*. More precisely, let U, V, K, and L be as before: we say that U and V are biinterpretable if there is a U-isomorphism I between $L\circ K$ and the identity interpretation on U, and a V-isomorphism J between $K\circ L$ and the identity interpretation on V (cf. Appendix B for the definition of the isomorphisms I and J).

It is useful to consider slight variants of the two notions in which only the relation of interpretation is changed. If U and V share part of their signature, we will occasionally require (i) that the interpretations between the two behave like the identity interpretation for the common vocabulary, and (ii) that such interpretations do not relativize quantifiers. If this common signature is Θ, we call such an interpretation, a Θ-interpretation. For K a Θ-interpretation of U in V, we write $K\colon U \to_\Theta V$. The definitions of Θ-synonymy and Θ-biinterpretability are then given in the obvious way by replacing interpretations with Θ-interpretations.

[21] See also Glymour (1970) and, for a more recent exposition relating definitional equivalence to Quine (1975), also Barrett and Halvorson (2016).

[22] I employ the definition of definitional equivalence given by Visser (2006). This definition may not coincide with the definition of Barrett and Halvorson (2016) in full generality, but it does in the specific cases I will consider.

[23] Although this notion of definitional equivalence, called *synonymy* in Visser (2006), is not in general equivalent to other notions of definitional equivalence defined in terms of common definitional extensions (Barrett and Halvorson, 2016), *it is* equivalent to it in the present setting.

2.2 *Expressing* general claims

I now turn to the main claims of this section. Let us work with a sufficiently expressive base theory B containing the usual machinery for formal syntax including self-reference via diagonalization – Kalmar's elementary arithmetic can be safely considered to be a safe lower bound (Nicolai, 2017). Given an infinite set of sentences S that is *definable* in B – I write φ_S for the formula defining it – I now compare, by means of theoretical equivalence, on the one hand the result of extending B with the 'infinite conjunction' of all instances of S, and on the other an extension of B in $\mathcal{L}_{\mathtt{Tr}}$ with principles entailing suitable $\mathtt{Tr}$-biconditionals *and* the single sentence $\forall x(\varphi_S(x) \rightarrow \mathtt{Tr}\, x)$.

It is important to notice that I am not investigating whether a universal generalization and its instances are deductively equivalent, nor whether they are only mutually interpretable over B. Both relations are in fact not adequate. In the first case, logic already tells us that the universal claim is stronger than its instances. The two are clearly not identical.[24] Similarly, mutual interpretability alone is clearly not sufficient to preserve virtually any form of theoretical status; B is for instance mutually interpretable with $B+$'B is inconsistent', and in general mutual interpretability does not even preserve the possibility of maintaining the 'standard' or 'intended' interpretation of the language under scrutiny.

I am instead concerned with whether there is a further, plausible sense in which the two logico-linguistic tools at hand – infinite lots of sentences, and truth-theoretic generalizations – can be *theoretically*, and therefore *explanatorily* the same. The first result is as follows:

Proposition 2. *Let B be finitely axiomatizable and let $B^{\mathtt{Tr}}$ a finite extension of B that proves all $\mathtt{Tr}$-sentences for $\mathcal{L}_B$. Furthermore, let S be a set of sentences of $\mathcal{L}_B$ that cannot be finitely axiomatized over B. Then the theory $B^{\mathtt{Tr}}+\forall x(\varphi_S(x) \rightarrow \mathtt{Tr}\, x)$ cannot be biinterpretable with $B+\{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in S\}$.*

The proof is a straightforward application of lemma 1 of Appendix B: if the two theories were biinterpretable, then S would be finitely axiomatizable over B; but this contradicts the assumption. As an immediate corollary, the infinite conjunction of members of S and the generalization $\forall x(\varphi_S(x) \rightarrow \mathtt{Tr}\, x)$ cannot be definitionally equivalent.

A few remarks on the plausibility of the assumptions of proposition 2; first, on the assumption of the non-finite axiomatizability of S over B. This seems to be perfectly in line with the deflationist's justification of the role of truth : if S were already finitely axiomatizable over B, we simply would not need the truth

[24] In addition, in the particular setting under consideration, Halbach (1999) has already shown that B with all $\mathtt{Tr}$-sentences for $\mathcal{L}$ cannot prove $\forall x\big(\varphi_S(x) \rightarrow \mathtt{Tr}\, x\big)$.

predicate to assert all of its elements. Second, the result relies on the finiteness of $B^{\mathtt{Tr}}$, both in the support theory B and in the truth-theoretic component of $B^{\mathtt{Tr}}$. The first of these assumptions can be relaxed, as I will now show. The second will be considered right after.

We have already seen that the notions of $\mathcal{L}_B$-definitional equivalence and $\mathcal{L}_B$-biinterpretability are defined in the same way as definitional equivalence and biinterpretability, except that one replaces the interpretations involved with $\mathcal{L}_B$-interpretations. Such notions appear natural in the present scenario: the infinite set of sentences S and its truth theoretic version that we are comparing are added to a common background theory providing the standard syntactic tools needed to formulate truth ascriptions and the instances of S. It looks entirely plausible to keep the meaning of such syntactic/structural machinery fixed in investigating the conceptual/theoretical equivalence of infinite lots of sentences and truth principles. Starting with $\mathcal{L}_B$-interpretations, one can obtain the following generalization of proposition 2 for arbitrary B, which follows immediately from proposition 5 in Appendix B.

Proposition 3. *Let X be a finite set of $\mathcal{L}_{\mathtt{Tr}}$-sentences that entails all the $\mathtt{Tr}$-sentences for $\mathcal{L}_B$ over B, and S a set of $\mathcal{L}_B$-sentences that is not finitely axiomatizable over B. Then the theory $B + X + \forall x(\varphi_S(x) \rightarrow \mathtt{Tr}\, x)$ cannot be $\mathcal{L}_B$-biinterpretable with $B + \{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in S\}$.*

Again, the fact that the two theories are not $\mathcal{L}_B$-definitionally equivalent immediately follows.

I now turn to the requirement that the cluster of truth principles are finitely presented. I take this to be a desideratum of a deflationist theory of truth. A finite formulation of the minimal theory is already explicitly endorsed in (Horwich, 1998, §5), although there it is also claimed that there are difficulties to achieve it. Similarly, more general deflationists positions, such as the one articulated in Horsten (2012), clearly resort to a finite presentation of the truth principles. For the structure of my argument, it is sufficient that the disquotationalist regards a finite presentation as the ideal – although perhaps not yet realized – for the presentation of their theory. Proposition 3 then tells us that an infinite lot of sentences and its corresponding truth-theoretic generalization cannot be equivalent in the strong sense required.

It is perhaps worth highlighting the parameter-free nature of the results above. On the one hand, propositions 2 and 3 do not rely on the choice of a *typed* version of the disquotational theory. I have formulated the results in terms of typed $\mathtt{Tr}$-biconditionals because of their simple and uncontroversial nature. As mentioned in the previous section, the choice of a consistent, type-free set of $\mathtt{Tr}$-biconditionals involves a certain degree of arbitrariness. However, the result

would still hold if for instance we replaced the Tarskian, typed bi-conditionals with (a finite theory entailing) biconditionals of the form (PT). On the other hand, both propositions still hold if we require only that the theory of truth only entails one direction of the disquotation schema. Picollo and Schindler (2017), for instance, convincingly show that for most of intended applications of EXPRESS, the sole `Tr-out` is in fact enough.[25] The results above apply equally to possible, weaker understandings of EXPRESS based on a restriction of the disquotation schema.

2.3 The quantificational role of truth

Even if the truth predicate of the disquotationalist cannot serve the expressive purposes one had hoped, there may still be room for the position that truth essentially serves the purpose of formulating, in natural or regimented languages suitable for philosophical theorizing, forms of quantification that do not standardly belong to it. Obvious targets are first-order versions of propositional and second-order quantifiers (Field, 1994). Examples such as (12) substantiate the role of truth, prescribed by QUANTIFY, in mimicking sentential quantification. Similarly, a disquotational truth predicate can be used to mimic quantification in *predicate* position, as in 'Maria is strong', which entails, via disquotation, 'there is a predicate P that is true of Maria'.

The results of the previous section established that the infinite conjunction of a set S of sentences and the claim 'all sentences in S are true' cannot be equivalent in the strong sense required by deflationism. However, this does not settle the question whether a deflationst truth predicate might *just be* a form of higher-order or propositional quantification, regardless of how powerful it may be in capturing infinite conjunctions and disjunctions. A view in which truth *is just* a tool to replicate higher-order quantification in a first-order setting is in fact weaker than the combination of EXPRESS and QUANTIFY considered above. For instance, one can consistently maintain that the truth predicate is not needed to express infinite lots of sentences – for instance, because only one direction of the `Tr`-schema is sufficient for the task – and yet claim that it is its pure quantificational role that calls for a disquotational truth (Picollo and Schindler, 2019).

In this section I discuss to what extent this residual role can be successfully carried out by the truth predicate. I focus on the case of second-order quantification, but I expect that similar considerations will apply also to the case of propositional quantification. I will establish a negative result: the theoretical equivalence between disquotational truth and second-order quantification breaks

[25] Although they do not claim that this is an adequate reading of EXPRESS.

down even at the most simple level. This strongly suggests that there is no hope to establish the equivalence at the more general level. Notice that, again, simple mutually interpretability or even mutual $\mathcal{L}_{\mathbb{N}}$-interpretability would not be enough for establishing the theoretical equivalence of truth and higher-order quantification. For instance, in the latter case, there are theories of compositional truth such as Kripke-Feferman truth (KF) that are mutually $\mathcal{L}_{\mathbb{N}}$-definable with expressively poor comprehension principles for positive elementary operators for $\mathcal{L}_{\mathbb{N}}$ (Cantini, 1989, §3). In these results, the truth/satisfaction predicate is not translated as the (second-order) predication relation of the second-order theory but by some predicate obtained by a diagonalization trick. This should not be allowed when one requires a natural correspondence between truth and quantification as in the reading of QUANTIFY under consideration.

The paradigmatic case of reduction between a disquotational truth predicate and second-order quantification concerns a minimal set of $\mathtt{Tr}$-sentences that only involve sentences not containing $\mathtt{Tr}$, on the one hand, and a form of *predicative*, second-order comprehension, on the other.[26] For simplicity, I here take B to be Peano arithmetic (PA), expand its language $\mathcal{L}_{\mathbb{N}}$ – given by a relational signature with finitely many relation symbols – with the truth predicate $\mathtt{Tr}$, and add to its axioms the schema

$$\forall x \left(\mathtt{Tr}^{\ulcorner} A(\dot{x})^{\urcorner} \leftrightarrow A(x)\right) \text{ for all } A(v) \text{ of } \mathcal{L}_{\mathbb{N}}. \tag{18}$$

(18) is a simple generalization of the $\mathtt{Tr}$-sentences, in which also parameters are allowed. We call the resulting theory, following Halbach (2014), UTB. The theory of predicative comprehension that we consider is also an extension of PA. After enriching $\mathcal{L}_{\mathbb{N}}$ with second-order quantifiers – governed by the usual rules of inference –, one adds to PA the schema

$$\exists Y \forall x \left(x \in Y \leftrightarrow A(x)\right) \tag{19}$$

where $A(x)$ does not contain second-order quantifiers or free set parameters. I call the theory ACA^{-}.[27] The folklore result that links the two theories is a strong form of interpretability that holds between them (Halbach, 2014; Nicolai, 2017): in particular, via translations that preserve the vocabulary of $\mathcal{L}_{\mathbb{N}}$ and translate only the truth via predication and vice versa. By employing the terminology introduced above, the two theories are mutually $\mathcal{L}_{\mathbb{N}}$-interpretable.

[26] Second-order should be here intended in the proof-theoretic sense. Semantically, this corresponds to choosing Henkin or many-sorted semantics. This choice is obvious; given our presentation of deflationist theories, the proof-theoretic presentation of second-order logic is the only one for which there may be hopes of interreducibility with deflationist truth.

[27] In particular, as it is natural to require, the induction schemata of ACA^{-} and UTB are extended to second-order formulas and the truth predicate respectively.

The notions of theoretical equivalence introduced earlier enable us to shed light on the relationships between UTB and ACA^-.

PROPOSITION 4. ACA^- *and* UTB *are* not *biinterpretable, and therefore not definitionally equivalent.*

The proof of proposition 4 is given in Appendix C. In essence the proof indicates that the truth predicate, when seen as a quantifier, can only be provably applied to first-order definable sets. But second-order quantification is much richer: it does not rule out classes that are not immediately definable by first-order quantifiers.[28]

The upshot of this result for the present discussion is clear. One might hope to re-calibrate the role and purpose of the truth predicate by tuning down the role of EXPRESS and by focusing only on QUANTIFY. The role of the truth predicate, on this view, would then consist in providing a first-order *reformulation* of higher-order quantification – in our example, predicative second-order quantification. This correspondence between the two devices needs to preserve the theoretical status of the claims involving them, including their explanatory status. For this reason, only a notion of theoretical equivalence suits the deflationist's need. And this is what is excluded by proposition 4.

3 Conclusion

Recent developments of truth-theoretic deflationism have focused on a formal analysis of principles of truth and their logical properties. In this extended sense, any position that considers truth as a primitive concept, and characterizes it *only* by means of a simple set of axioms – or rules of inference –, would count as deflationary (Horsten, 2012). In this paper I have attempted to reconcile these formal approaches to the classical tenets of truth-theoretic deflationism: FIX, EXPRESS, QUANTIFY, and EXPLAIN.

The upshot of the analysis seems clear. The best approach available to make explicit the modal status of disquotation in FIX leads to inconsistency. And this is likely to generalize to structurally similar accounts of FIX. The combination of EXPRESS and QUANTIFY requires that infinite lots of sentences and their corresponding truth theoretic generalizations stand in a close theoretical relationship that preserves their explanatory status. In the most natural way of understanding this relationship, that is via formal notions of conceptual or theoretical equivalence, such relationship simply cannot exist. Finally, even if one considers QUANTIFY in isolation, by claiming that the principles of the deflationist's truth predicate are there *just* to mimic higher-order quantification

[28] For many more results linking predicative comprehension and typed truth theories, I refer to Nicolai (2017).

in a first-order setting, one requires the theoretical equivalence between truth and quantification. In the natural sense of definitional equivalence or biinterpretability, this equivalence cannot hold.

What has been said, of course, does not impact on truth theoretic deflationism intended in a loose sense, as the formal and philosophical study of principles of truth. If, however, the deflationary approach to truth is charaterized – as it is commonly done – by FIX, EXPRESS, QUANTIFY, and EXPLAIN, its chances of success are slim.

References

Azzouni, J. (2006). *Tracking Reason: Proof, Consequence, and Truth*. Oxford University Press USA.

Barrett, T. W. and Halvorson, H. (2016). Glymour and Quine on theoretical equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

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

Cieśliński, C. (2017). *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge University Press.

Field, H. (1994). Deflationist views of meaning and content. *Mind*, 103(411):249–285.

Frege, G. (1918). Thoughts. In Frege, G., editor, *Logical Investigations*. Blackwell.

Glymour, C. (1970). Theoretical realism and theoretical equivalence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1970:275–288.

Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, 21(1):57–81.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, 108(429):1–22.

Halbach, V. (2001). Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66(4):1959–1973.

Halbach, V. (2003). Modalized disquotationalism. In Horsten, L. and Halbach, V., editors, *Principles of Truth*, pages 75–102. De Gruyter.

Halbach, V. (2009). Reducing compositional to disquotational truth. *Review of Symbolic Logic*, 2(4):786–798.

Halbach, V. (2014). *Axiomatic theories of truth. Revised edition*. Cambridge University Press.

Halvorson, H. (2012). What scientific theories could not be. *Philosophy of Science*, 79(2):183–206.

Heck, R. (2005). Truth and disquotation. *Synthese*, 142(3):317–352.

Horsten, L. (2012). *The Tarskian Turn*. MIT University Press, Oxford.

Horsten, L. and Leigh, G. E. (2017). Truth is simple. *Mind*, 126(501):195–232.

Horwich, P. (1998). *Truth*. Clarendon Press.

Ketland, J. (1999). Deflationism and tarski's paradise. *Mind*, 108(429):69–94.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72:690–712.

McGee, V. (1992). Maximal consistent sets of instances of tarski's schema (t). *Journal of Philosophical Logic*, 21(3):235–241.

Nicolai, C. (2015). Deflationary truth and the ontology of expressions. *Synthese*, 192(12):4031–4055.

Nicolai, C. (2016). A note on typed truth and consistency assertions. *Journal of Philosophical Logic*, 45(1):89–119.

Nicolai, C. (2017). Equivalences for truth predicates. *Review of Symbolic Logic*, 10(2):322–356.

Picollo, L. and Schindler, T. (2017). Disquotation and infinite conjunctions. *Erkenntnis*, 83:899–928.

Picollo, L. and Schindler, T. (2019). Deflationism and the function of truth. *Philosophical Perspective*. forthcoming.

Quine, W. V. (1970). *Philosophy of Logic*. Harvard University Press.

Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis*, 9(3):313–28.

Quine, W. V. (1990). *Pursuit of Truth*. Harvard University Press.

Ramsey, F. P. (1927). Facts and propositions. *Proceedings of the Aristotelian Society*, 7(1):153–170.

Shapiro, S. (1998). Proof and truth. *Journal of Philosophy*, 95(10):493–521.

Stern, J. (2016). *Toward Predicate Approaches to Modality*. Switzerland: Springer.

Visser, A. (2006). Categories of theories and interpretations. In Enayat, A., Kalantari, I., and Moniri, M., editors, *Logic in Tehran*, Vol. 26. Lecture Notes in Logic. La Jolla, CA.

Williams, M. (1999). Meaning and deflationary truth. *Journal of Philosophy*, 96(11):545–564.

Appendix A: Modality

The *positive complexity* $\pi(\cdot)$ of a formula φ of $\mathcal{L}_{\mathtt{Tr}}$ is defined inductively:

$$\pi(\varphi) = \begin{cases} 0, & \text{if } \varphi \text{ is a atomic or negated atomic} \\ \pi(\psi) + 1, & \text{if } \varphi \text{ is } \neg\neg\psi,\ \forall v\psi,\ \neg\forall v\psi,\ \exists v\psi,\ \neg\exists v\psi \\ \max(\pi(\psi), \pi(\chi)), & \text{if } \varphi \text{ is } \psi \circ \chi \text{ or } \neg(\psi \circ \chi), \text{ with } \circ = \wedge, \vee. \end{cases}$$

The primitive recursive translation $*\colon \mathcal{L}_{\mathtt{Tr}} \to \mathcal{L}^{+}$ is defined by induction on the positive complexity of formulas of $\mathcal{L}_{\mathtt{Tr}}$ and it essentially employs Kleene's (second) recursion theorem (Halbach, 2014, Ch. 5):

$$\begin{array}{ll} (R(t_1, \ldots, t_n))^* := R(t_1, \ldots, t_n) & \text{with } R \text{ a relation of } \mathcal{L}_{\mathbb{N}} \\ (\neg R(t_1, \ldots, t_n))^* := \neg R(t_1, \ldots, t_n) & \text{with } R \text{ a relation of } \mathcal{L}_{\mathbb{N}} \\ (\mathtt{Tr}t)^* := \mathtt{Tr}t^* & (\neg\mathtt{Tr}t)^* := \mathtt{F}t^* \\ (\neg\neg\varphi)^* := \varphi^* & (\varphi \wedge \psi)^* := \varphi^* \wedge \psi^* \\ (\neg(\varphi \wedge \psi))^* := (\neg\varphi)^* \wedge (\neg\psi)^* & (\varphi \vee \psi)^* := \varphi^* \vee \psi^* \\ (\forall v\varphi)^* := \forall v\varphi^* & (\neg\forall v\varphi)^* := \exists v(\neg\varphi)^* \\ (\exists v\varphi)^* := \exists v\varphi^* & (\neg\exists v\varphi)^* := \forall v(\neg\varphi)^* \end{array}$$

Proof of Proposition 1. MPT^* is a classical theory in $\mathcal{L}_{\mathtt{Tr}}$. Therefore it will entail $\mathtt{Tr}l \vee \neg\mathtt{Tr}l$. By necessitation, MPT^* will also prove $\Box(\mathtt{Tr}l \vee \neg\mathtt{Tr}l)$; therefore $\mathtt{Tr}(\mathtt{Tr}l \vee \neg\mathtt{Tr}l)^*$ by ($\mathtt{tfact}^*$). By the nature of the mapping $(\cdot)^*$, therefore,

$$\mathtt{Tr}(\mathtt{Tr}l^* \vee \mathtt{F}l^*). \tag{20}$$

Since both $\mathtt{Tr}l^*$ and $\mathtt{F}l^*$ are $\mathcal{L}^{+}$-sentences, and MPT proves compositional principles for $\mathcal{L}^{+}$, we can distribute the truth predicate over the disjunction to obtain

$$\mathtt{Tr}(\mathtt{Tr}l^*) \vee \mathtt{Tr}(\mathtt{F}l^*). \tag{21}$$

Now we can reason by cases: if $\mathtt{Tr}(\mathtt{Tr}l^*)$, then also $\mathtt{Tr}l^*$ by (6). By the identity of l and $\ulcorner\neg\mathtt{Tr}l\urcorner$, also $\mathtt{TrF}l^*$; then, by (7), $\mathtt{F}l^*$. Thus MPT^* entails both $\mathtt{Tr}l^*$ and $\mathtt{F}l^*$. Similarly, if $\mathtt{Tr}(\mathtt{F}l^*)$, we immediately obtain $\mathtt{F}l^*$ by (7), but also $\mathtt{Tr}l^*$ by (9). Again, both $\mathtt{F}l^*$ and $\mathtt{Tr}l^*$. It follows that MPT^* itself proves the existence of sentences – in fact, by parametrizing, infinitely many – that are both true and false, as required. □

Appendix B: Theoretical Equivalence

Given first-order theories T and W, a *relative translation* of $\mathcal{L}_T$ into $\mathcal{L}_W$ – formulated in a relational signature – can be described as a pair (δ, F) where δ is a $\mathcal{L}_W$-formula with one free variable – the domain of the translation – and F is a (finite) mapping that takes n-ary relation symbols of $\mathcal{L}_T$ and gives back formulas of $\mathcal{L}_W$ with n free variables. The translation extends, modulo suitable renaming of bound variables, to the mapping τ:

- $(R(x_1, ..., x_n))^\tau :\leftrightarrow F(R)(x_1, ..., x_n)$;
- τ commutes with propositional connectives;
- $(\forall x\ A(x))^\tau \ :\leftrightarrow\ \forall x\ (\delta(x) \to A^\tau)$.

Definition 1. *An interpretation K is specified by a triple (T, τ, W), where τ is a translation of $\mathcal{L}_T$ in $\mathcal{L}_W$, such that for all formulas $\varphi(x_1, .., x_n)$ of $\mathcal{L}_T$ with the free variables displayed, we have:*

$$\textit{if } T \vdash \varphi(x_1, ..., x_n), \textit{ then } W \vdash \textstyle\bigwedge_{i=1}^{n} \delta_K(x_i) \to \varphi^\tau$$

I write $K : T \to W$ for 'K is an interpretation of T in W'. An interpretation is *direct* if it maps identity to identity and it does not relativize quantifiers. T and W are said to be *mutually interpretable* if there are interpretations $K\colon T \to W$ and $L\colon W \to T$.

Given $\tau_0\colon \mathcal{L}_T \to \mathcal{L}_W$ and $\tau_1\colon \mathcal{L}_W \to \mathcal{L}_V$, the composite of $K = (T, \tau_0, W)$ and $L = (W, \tau_1, V)$ is the interpretation $L{\circ}K = (T, \tau_1{\circ}\tau_0, V)$, where $\delta_{L\circ K}(x) :\leftrightarrow \delta_K^L(x) \wedge \delta_L(x)$. Two interpretations $K_0, K_1\colon T \to W$ are *equal* if W, the target theory, proves this. In particular, one requires,

$$W \vdash \forall x\,(\delta_{K_0}(x) \leftrightarrow \delta_{K_1}(x))$$
$$W \vdash \forall \vec{x}\,(R_{K_0}(\vec{x}) \leftrightarrow R_{K_1}(\vec{x})) \qquad \text{for any relation symbol } R \text{ of } \mathcal{L}_T$$

A W- definable *morphism* between interpretations $K_0, K_1\colon T \to W$ is a triple (K_0, I, K_1), with I a $\mathcal{L}_W$-formula with two free variables, such that W proves:

(22) $\quad \forall x, y\,\big(I(x, y) \to (\delta_{K_0}(x) \wedge \delta_{K_1}(y))\big)$

(23) $\quad \forall x, y, u, v\big(x =_{K_0} y \wedge u =_{K_1} v \wedge I(y, u) \to I(x, v)\big)$

(24) $\quad \forall x\,\big(\delta_{K_0}(x) \to \exists y\,(\delta_{K_1}(y) \wedge I(x, y))\big)$

(25) $\quad \forall x, y, z\big(I(x, y) \wedge I(x, z) \to y =_{K_1} z\big)$

(26) $\quad \forall \vec{x} \forall \vec{y}\,\big(\bigwedge_{i=1}^{n} I(x_i, y_i) \wedge R_{K_0}(\vec{x}) \ \to\ R_{K_1}(\vec{y}))\big)$

for any n-ary relation $R \in \mathcal{L}_T$.

To obtain an *isomorphism* from K_0 to K_1 one needs to add the requirement that W proves:

$$\forall y \left(\delta_{K_1}(y) \to \exists x \left(\delta_{K_0}(x) \wedge I(x,y)\right)\right) \tag{27}$$

$$\forall x, y, z \left(I(x,y) \wedge I(z,y) \to x =_{K_0} z\right) \tag{28}$$

$$\forall \vec{x} \forall \vec{y} \Big(\bigwedge_{i=1}^{n} I(x_i, y_i) \wedge R_{K_1}(\vec{y}) \to R_{K_0}(\vec{x})\Big) \tag{29}$$

for any relation $R \in \mathcal{L}_T$.

We write $F\colon K_0 \cong K_1$ for 'F is an isomorphism from the interpretation K_0 to K_1'.

Definition 2 (SYNONYMY, DEFINITIONAL EQUIVALENCE). *U and V are synonymous if and only if there are interpretations $K\colon U \to V$ and $L\colon V \to U$ such that V proves that $K \circ L$ and id_V are equal and, symmetrically, U proves that $L \circ K$ is equal to id_U.*

Definition 3 (BI-INTERPRETABILITY). *Given a pair of interpretations $K\colon U \to V$ and $L\colon V \to U$, U and V are bi-interpretable if and only if (i) there is a $\mathcal{L}_V$-formula F_0 such that V proves F_0 to be an isomorphism between $K \circ L$ and id_V and (ii) there is an $\mathcal{L}_U$-formula F_1 such that U proves F_1 to be an isomorphism between $L \circ K$ and id_U.*

LEMMA 1 (Visser 2006). *Let U, V be theories in finite signatures. Assume that $K\colon U \to V$ and $L\colon V \to U$ are interpretations and that U defines an isomorphism F from $L \circ K$ to id_U. Assume further that V is finitely axiomatizable. Then U is finitely axiomatizable.*

Proof. Let V_0 be the conjunction of a finite axiomatization of V. A finite $U_0 \subseteq U$ is specified by the single sentences: (i) F is an isomorphism between $L \circ K$ and id_U; (ii) V_0^L. The theory U_0 is clearly a subtheory of U. For the converse direction, one verifies that if U proves the sentence A, then $U_0 \vdash A^{K^L}$ by (ii) and the definition of retract. Thus $U_0 \vdash A$ by (i). □

To lift the proposition 2 to arbitrary base theories B interpreting a modicum of formal syntax, one needs to resort to $\mathcal{L}_B$-interpretations. Then we can prove the following:

PROPOSITION 5. *Given a first-order base theory B, let T_0 be the theory $B + X$, where X is a set of sentences in a signature that finitely expands $\mathcal{L}_B$. Moreover, let T_1 be $B + Y$, where Y is finitely axiomatizable over B in a language again finitely expanding $\mathcal{L}_B$. If there is a T_0-isomorphism between $L \circ K$ and Id_{T_0} with $K\colon T_0 \to_{\mathcal{L}_B} T_1$ and $L\colon T_1 \to_{\mathcal{L}_B} T_0$, then X is finitely axiomatizable over B.*

Proof. Let A be a finite axiomatization of Y over B. We let

$$T_0^* := B + A^L + \text{`}I\colon L \circ K \cong \mathrm{Id}_{T_0}\text{'}$$

Clearly, T_0^* is a subtheory of T_0. For the converse direction: for an arbitrary $\mathcal{L}_{T_0}$-sentence φ, if $T_0 \vdash C$, then $B + A \vdash C^K$. But then also $T_0^* \vdash \varphi^{K^L}$, and therefore $T_0^* \vdash C$ by the existence of a $I\colon L \circ K \cong \mathrm{Id}_{T_0}$ in T_0^*. □

Appendix C: Predicative Comprehension

That ACA^- and UTB are mutually $\mathcal{L}_{\mathbb{N}}$-interpretable can be seen as follows. On the one hand, the truth predicate of UTB can be understood as a *partial truth class*, namely a definable class in ACA^-: essentially, one defines by (19) a class of all true $\mathcal{L}_{\mathbb{N}}$-formulas of a fixed complexity n. I call this interpretation K. This is possible because truth predicates for $\mathcal{L}_{\mathbb{N}}$-formulas of fixed complexity are already definable in PA. On the other, class membership of ACA^- can be defined as 'true of' in UTB: a class X is translated as a unary formula $A_X(v)$ of $\mathcal{L}_{\mathbb{N}}$, and predications of the form $n \in X$ are translated as $\mathtt{Tr}^{\ulcorner}A(\overline{n}/v)^{\urcorner}$. I call this interpretation L.

Proof of proposition 4 . This is a generalization of an argument due to unpublished work by Albert Visser and Ali Enayat. It is originally contained in Nicolai (2017). Seeking a contradiction, let's assume that ACA^- and UTB are biinterpretable.

Now let us consider the two-sorted structure $(\mathbb{N}, \mathrm{P}(\omega)) \vDash \mathsf{ACA}^-$. By assumption, in $(\mathbb{N}, \mathrm{P}(\omega))$ we can find an internal model $(\mathcal{M}, S) \vDash \mathsf{UTB}$ – $S \subset M$ being the extension of $\mathtt{Tr}$ – that, in turn, contains a model $(\mathcal{N}, \mathcal{R}) \vDash \mathsf{ACA}^-$ with the property that $(\mathcal{N}, \mathcal{R})$ is isomorphic to $(\mathbb{N}, \mathrm{P}(\omega))$ – verifiably in $(\mathbb{N}, \mathrm{P}(\omega))$. Since $(\mathcal{M}, S)$ interprets $(\mathcal{N}, \mathcal{R})$, the isomorphism of $(\mathcal{N}, \mathcal{R})$ and $(\mathbb{N}, \mathrm{P}(\omega))$ gives us an interpretation of $(\mathbb{N}, \mathrm{P}(\omega))$ in $(\mathcal{N}, \mathcal{R})$, and therefore of $(\mathbb{N}, \mathrm{P}(\omega))$ in $(\mathcal{M}, S)$ because interpretability is a transitive relation – in particular, this means that there are formulas δ_ω, and $\delta_{\mathrm{P}(\omega)}$ of $\mathcal{L}_{\mathtt{Tr}}$ and a surjection from the extension of these formulas in $\mathcal{M}$ to ω and $\mathrm{P}(\omega)$ which is well-behaved with respect to the arithmetical primitives. As a consequence $(\mathcal{M}, S)$ can define its standard natural numbers. But also, since $(\mathcal{M}, S)$ satisfies full induction with $\mathtt{Tr}$, we can $(\mathcal{M}, S)$-define an injection $f\colon \mathcal{M} \to \omega$ – see (Nicolai, 2017, Lemma 2.2). So $(\mathcal{M}, S)$ is countable. This contradicts the assumption that $(\mathcal{M}, S)$ interprets the uncountable model $(\mathbb{N}, \mathrm{P}(\omega))$. □