Fix, Express, Quantify

Disquotation after its logic

Carlo Nicolai*
King's College London

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 at odds 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 the truth predicate 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. EXPRESS

*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 and Bristol Philosophy Seminar, Alex Franklin, Johannes Stern, Volker Halbach, Leon Horsten, Albert Visser, and two anonymous referees for this journal. Special thanks go to Thomas Schindler for detailed comments on a previous draft. The initial stages of this research were supported by the VENI NWO Grant 275-20-057.

[1]The monographs Halbach (2014), Horsten (2012), and Cieśliński (2017) contain detailed overviews of such studies and results in the context of classical logic. Non-classical accounts are considered in Beall (2009), Field (2008).

and QUANTIFY are deeply intertwined, and I shall treat them in this way below. 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 (Quine, 1990, §33); a formal rendering of Quine's claim has been put forth in Halbach (1999).

In the paper I 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. And they cannot be;[2]

(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.

There is, however, 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 debated 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. It should be *conservative* over the base theory. In this paper I will not discuss EXPLAIN, although I occasionally appeal to it; I have already 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.[3]

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

[2] As it shall be clear later on, this conclusion was already defended by Gupta (1993) on philosophical grounds. One can see my approach as providing new formal results to corroborate Gupta's diagnosis.

[3] See Nicolai (2015, 2016) for a discussion of the former claim. The latter is not at all original and is discussed already in Halbach (2001), Horsten (2012), and more recently by Picollo and Schindler (2019).

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.[4] 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.

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.

My starting point is the observation that deflationists do not take the schema (T) to express simple material equivalence. Several theorists have articulated its status it in slightly different ways. Hartry Field proposes to understand the equivalence involved in the Tr-biconditionals as a form of cognitive equivalence that, as such, is empirically indefeasible (Field, 1994, §6). Similarly, Horwich suggests a natural extension of the schema (T) to accommodate blind ascriptions of modal nature and the interaction between disquotational truth and alethic modalities (Horwich, 1998, §3, fn. 5). A detailed form of *modal disquotationalism*, based on the notion of truth-analyticity, is defended in Halbach (2003). The thesis that FIX entails *at least* the necessary truth of the schema (T) will be discussed in detail shortly.[5]

That FIX cannot be understood in terms of material equivalence can also be inferred by the deductive weakness of a material reading of (T). To see this, and to lay down some groundwork for the sections to come, I work over a base language $\mathcal{L}$ that is capable of talking about the syntax of formal languages and theories in first-order (many-sorted) logic. The language of Peano Arithmetic

[4]By assuming that the objects of truth are sentences of English, I deliberately depart from Horwich's minimalism, although minimalism is not entirely dependent on the choice of propositions as opposed to certain classes of sentences (Horwich, 1998, §2.1). Moreover, 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. More palatable alternatives 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.

[5]There is also another, related argument in support of the claim that the equivalence of A and 'A is true' 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 biconditional should express a form of synonymy. I will consider this option in §2.

$\mathcal{L}_{\mathbb{N}}$ 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 $\mathcal{L}_{\mathrm{Tr}}$ is simply $\mathcal{L}_{\mathbb{N}}$ expanded with a truth predicate 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 appeal to a semantic classification of 'pathological' sentences, but 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 focus on *positive* sentences of $\mathcal{L}_{\mathrm{Tr}}$, that is sentences in which the truth predicate appears only in the scope of an even number of negations (Halbach, 2009). By accepting a restriction of (T), one accepts that the schema does not hold for all sentences of $\mathcal{L}_{\mathrm{Tr}}$. 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 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 its instances may be plausibly considered to be a component of any adequate version of it that involves a self-applicable truth predicate.

For my purposes, it is useful to employ a slightly different definition of the positive fragment of $\mathcal{L}_{\mathrm{Tr}}$. Following Horsten and Leigh (2017), I consider a *negation-free* language $\mathcal{L}^+$ with logical primitives $\vee, \wedge, \exists, \forall$, and in which every atomic predicate $P \in \mathcal{L}$ has a dual $\bar{P}$. The dual of Tr is 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 $A \in \mathcal{L}^+$:

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

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.[7] Such generalizations typically concern logical laws such as "All sentences of the form 'If p then p' are true" (Quine, 1990, p. 80), or

[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] See, for instance, (Quine, 1990, p. 80).

semantic principles of *compositionality* such as

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

$(\wedge)$ generalizes only over sentences not containing the truth predicate.[8] Yet, neither it nor *any* plausible truth-theoretic generalization can be handled by the schema (PT) (Halbach, 2009, lemma 6.1): every truth theoretic generalization 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 reacted to similar observations by proposing a non-logical principle of the form 'if some property P holds of *each* proposition, then it holds of all propositions' (Horwich, 1998, Postscript, §5). It directly follows from Halbach's result that such principle cannot be derived from simple disquotation when sentences are at stake. Analogously, it only suffices that propositions share some structural feature with sentences to conclude that the principle of generalization advocated by Horwich cannot follow from a material reading of the propositional disquotation schema.

It is nonetheless natural to wonder whether generalizing tools of similar kind may be independently justified. Truth theorists have proposed different strategies. Horsten and Leigh (2017) frame Feferman's classical theory of implicit commitment (Feferman, 1962, 1991) in terms of Burge's and Wright's accounts of entitlement (Wright, 2004; Burge, 2011), and show that reflection principles can be used to recover compositional principles from material disquotation. Inspired by Feferman (1991), Field (2006) focuses on the combination of a dynamic reading of schemata, such as the Tr-biconditionals, and suitable substitution rules. Consistently with the proposal above, Horwich argues that there is a sense in which one's disposition to accept all instances of a schematic principle suffices to ground one's acceptance of the general claim on the basis of principles of 'introspection' or 'awareness' (Horwich, 2010, p. 45ff). Cieśliński (2017) proposes an interesting development of Horwich's solution (Cieśliński, 2017, §13.4), based on the notion of 'believability': if one accepts, say, all instances of the Tr-biconditionals for $\mathcal{L}_{\mathbb{N}}$, she will be in a position to believe all compositional axioms.

An evaluation of such proposals is outside the scope of this paper. There is, however, a natural way of extending (PT) that is in fact based on the intrinsically intensional reading of FIX proposed by deflationist theorists and that, as such,

[8] The quantified claim $\forall\varphi \, A(\varphi)$ is intended to abbreviate $\forall x(\mathrm{Sent}_{\mathcal{L}_{\mathbb{N}}}(x) \rightarrow A(x))$, where $\mathrm{Sent}_{\mathcal{L}_{\mathbb{N}}}(x)$ formalizes the predicate 'x is a sentence of $\mathcal{L}_{\mathbb{N}}$'. Existential quantification is treated similarly. For the sake of readability, I do not distinguish between the logical constants and their corresponding syntactic operations: that is, I also write $\varphi \rightarrow \psi$ for $\mathrm{imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)$, where imp represents in $\mathcal{L}_{\mathbb{N}}$ the syntactic operation on codes of formulas $\varphi, \psi \mapsto \varphi \rightarrow \psi$.

does not require any justification independent from the deflationary theory of truth itself. This is the aspect of FIX to which I now turn.

1.1 Modality

The equivalence between A and 'A is true' is not material. However, the traditional deflationist reading of FIX assigns to the Tr-biconditionals a certain intensional status. For my purposes it is sufficient to give a general structural account of the kind of intensionality involved in the disquotation schema, without committing myself to a specific notion such as conceptual necessity (Field, 1994, §6) or analyticity (Halbach, 2003).[9]

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 I would like the objects of truth and the objects of necessity to be the same. In addition, this choice enables one to state desirable laws, such as 'what is necessary is 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 $\mathcal{L}_{\mathrm{Tr}}^{\Box}$ the language $\mathcal{L}_{\mathrm{Tr}} \cup \{\Box\}$.

A structural account of the modal status of disquotation arises naturally from a generalization of Halbach's modalized disquotationalism (Halbach, 2003).[10] 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:[11]

$$\forall \varphi \in \mathcal{L}_{\mathrm{Tr}}^{\Box} (\mathrm{Prov}_{\mathrm{fol}}(\varphi) \rightarrow \Box \varphi). \tag{log1}$$

where $\mathrm{Prov}_{\mathrm{fol}}(\cdot)$ is a canonical provability predicate for first-order logic in $\mathcal{L}_{\mathrm{Tr}}^{\Box}$. On the other, one requires a closure condition that is reminiscent of the modal

[9] Of course, by endorsing the modal status of truth principles, deflationism clearly departs from theorists who regard the of meaning expressions of one's language involving truth (and reference) as a contingent matter (Lewy, 1947; Putnam, 1985). The disquotationalist's truth predicate is regarded as a primitive tool of quasi-logical character, mainly employed to disquote and generalize, and may be paraphrased 'true-as-I-understand-it', or 'true-from-my-present-perspective' (Field, 1994). The contingency of linguistic meaning does not immediately affect a truth predicate understood in this sense.

[10] The present approach is a generalization of modalized disquotationalism because Halbach mostly deals with typed principles for truth and necessity. The principles I discuss are type-free extensions of Halbach's principles.

[11] 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 $\mathcal{L}_{\mathrm{Tr}}$ are considered to be closed under classical logic.

axiom K:

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

The core principle of generalized modal disquotationalism states that all instances of the schema (PT), that is Tr-biconditionals for positive formulas, are modalized:

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

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.[12]

Finally, modalized disquotationalism requires $\Box$ to be factive.

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

As an additional axiom, one may add a necessitation principle:

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

Also principles (mpt), (fact) seem straightforwardly compatible with an understanding of $\Box$ as truth-analyticity (Halbach, 2003), or conceptual necessity (Field, 1994, §3). Both logical truths and disquotation principles are in fact necessary. If one adds (nec), this should be on the grounds that truths about the syntactic structure of language are also necessary. If one extended the theory with contingent vocabulary (and axioms), then (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 $\mathcal{L}_{\mathrm{Tr}}^{\Box}$-theory PA+(log1)-(fact), and $\mathrm{MPT}_{\mathrm{nec}}$ the $\mathcal{L}_{\mathrm{Tr}}^{\Box}$-theory PA+(log1)-(nec).

The restrictions to the schemata of MPT are in place to avoid paradox. Some of them can be lifted to obtain stronger principles. Since we aim to a negative result, we won't need them.

FACT 1 (Halbach, 2003, Thm. 4). MPT *and* $\mathrm{MPT}_{\mathrm{nec}}$ *are consistent.*

The intensional equivalence between A and ' "A" is true' articulated by

[12] To be precise (mpt) is a notationally simplified formulation of

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

where $\mathrm{eq}(x,y)$ sends the codes of two formulas to the code of their biconditional, sub is a substitution function and num is the numeral function just mentioned. Similar conventions apply to (fact) below.

MPT overcomes the weakness of the material reading of the Tr-biconditionals. 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

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

and one application of (fact).[13] A similar argument enables one to obtain in MPT the useful principles for truth ascriptions:[14]

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

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

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

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

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

Once again, by endorsing the principles of MPT, I do not want to claim that this is a *sufficient* formal analysis of FIX. In fact, in the following we will encounter good reasons for requiring an even stronger form of equivalence between the two sides of the Tr-biconditionals. For the sake of the argumentation it is sufficient to hold that the modal status of the Tr-biconditionals 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

The factivity principle (fact) can be straightforwardly paraphrased as 'if a sentence is necessary/truth-analytic, then it's true'. The schematic formulation of factivity principles is the standard choice in modal or epistemic logic when the

[13] In particular, (2) is obtained by the following theorems of MPT:

$$\forall \varphi, \psi \in \mathcal{L}^+ \, \Box\big(\mathrm{Tr}(\varphi \mathbin{\dot{\wedge}} \psi) \leftrightarrow \varphi \wedge \psi\big), \tag{3}$$

$$\forall \varphi \in \mathcal{L}^+ \, \Box(\mathrm{Tr}\dot{\varphi} \leftrightarrow \varphi), \tag{4}$$

$$\forall \psi \in \mathcal{L}^+ \, \Box(\mathrm{Tr}\dot{\psi} \leftrightarrow \psi). \tag{5}$$

[14] Given our conventions, (6) is an abbreviation of the longer:

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

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

modality in question is given in the form of a sentential operator. 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 means to endorse infinite sets 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 crucially translates in the possibility of analyzing in our language the rejection of schemata.

To accommodate this idea in MPT, one may attempt to replace (fact) with the axiom:

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

However, this would not be 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 (log1) introduces under the scope of $\Box$ sentences of a different language: for instance, since MPT is a classical theory in $\mathcal{L}_{\mathrm{Tr}}$, the sentence $\neg \mathrm{Tr} t \vee \mathrm{Tr} t$ will be an $\mathcal{L}_{\mathrm{Tr}}$-theorem of first-order logic. Therefore, so will be $\Box(\neg \mathrm{Tr} t \vee \mathrm{Tr} t)$. But $\neg \mathrm{Tr} t$ is a *negative* sentence, and the truth predicate of MPT has nothing to say about these sentences.[15] This is not ideal.

A slight modification of (tfact) will however deliver a more palatable principle. There is a natural mapping of the language $\mathcal{L}_{\mathrm{Tr}}$ into the language $\mathcal{L}^+$ that essentially replaces negative occurrences of truth predicates with the falsity predicate F of $\mathcal{L}^+$ – the details of such mapping are provided in Appendix A.[16] I denote with φ^* the $\mathcal{L}^+$-sentence resulting from the translation of the $\mathcal{L}_{\mathrm{Tr}}$-sentence φ. One can then turn (tfact) into the more plausible

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

It is clear that (tfact*) deals satisfactorily with negative theorems of MPT such as $\neg \mathrm{Tr} \ulcorner 0 \neq 0 \urcorner$.

However, there is little hope to give an adequate version of modal disquotationalism including (tfact*). The proof of the following result in given in Appendix A:

[15] A similar point holds for $\mathrm{MPT}_{\mathsf{nec}}$ and innocuous, negative sentences such as $\neg \mathrm{Tr} \ulcorner 0 \neq 0 \urcorner$.

[16] 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 $\mathcal{L}_{\mathrm{Tr}}$ 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.

PROPOSITION 1. *The truth predicate of any* $T \supseteq$ PA *in* $\mathcal{L}^{\Box}_{Tr}$ *containing* (PT), (6)-(9), (log1), (tfact*), *is inconsistent.*

Proposition 1 suggests that the truth predicate of any reasonable modal account of disquotation is bound to be inconsistent; there are sentences of $\mathcal{L}^+$ that the deflationist's truth predicate deems both true and false. I believe this 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 advanced here is that the theory of truth and the theory of logic have nothing to do with one another. (Horwich, 1998, p. 74)

In general, the classical deflationist approach 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 need to be dropped at the level of the internal logic of truth. It's then reasonable to infer from Horwich's passage – and analogous discussions – that, whereas it might be acceptable that certain general claims of logical character such as

'every sentence is either true or false'

do not follow from the theory of truth in their unrestricted form, one should not obtain truth-theoretic consequences of logical nature that contradict one's background logic. But this is exactly what happens in the case of modalized disquotationalism.

It may be objected that argument just given simply amounts to additional evidence for the inconsistency of truth. Inconsistency in this context may be understood in at least three ways: as an inconsistency view of the ordinary concept of truth, as a fully fledged dialetheist account of truth, or as semi-classical approach in which one accepts an internally inconstent truth predicate in a classical environment. The remarks just made are intended to rule out to the latter option. The first option is simply not compatible with the framework considered: I explicitly started with a (minimal) consistent set of biconditionals *because of paradox*, and showed that even such a restricted set leads to inconsistency. So this is radically different from accepting all instances of the schema (T) and learning to live up with their inconsistency (Azzouni, 2006, ch. 5). We are

left with the full dialetheist approach, as defended for instance in Beall (2009). Although my argument does not affect such a view, there are good reasons not to be persuaded by it. They are essentially related to the impact that such frameworks have on applied mathematics: contrary to what the advocates of nonclassical solution to paradox claim – usually, by appeal to so-called recapture strategies (Field, 2008) –, the non-classicality of truth 'spreads' to mathematical principles that we regard as uncontroversial and compromises their universal applicability (Halbach and Nicolai, 2018).

A final word about the scope of the argument given. Modal disquotationalism amounts to the deflationist's best attempt so far for articulating the modal status of FIX. What I said, of course, does not rule out the possibility of finding a better modal rendering of FIX that does not lead to inconsistency. 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 and its interplay with EXPRESS and QUANTIFY. It will be the focus of the following sections.

2 Express, quantify

The links between EXPRESS and QUANTIFY are clear if one looks at examples in the deflationist canon. Following Quine (Quine, 1990, p. 80), the infinite conjunction

(10) $$\begin{aligned}&(\text{snow is white} \rightarrow \text{snow is white}) \wedge \\ &(\text{grass is green} \rightarrow \text{grass is green}) \wedge \ \ldots\end{aligned}$$

is taken to be equivalent, via FIX, to

(11) $$\begin{aligned}&(\mathrm{Tr}\ulcorner\text{snow is white} \rightarrow \text{snow is white}\urcorner) \wedge \\ &(\mathrm{Tr}\ulcorner\text{grass is green} \rightarrow \text{grass is green}\urcorner) \ \wedge \ldots\end{aligned}$$

The conjunction of EXPRESS and QUANTIFY then should enable one to conclude:

(12) $$\text{for all sentences } \sigma\text{: } \sigma \rightarrow \sigma \text{ is true.}$$

There are different options to understand the relationship between (10) and (12) or, more generally, between an infinite conjunction or disjunction of infinite sets of sentences of the deflationist's language and the truth theoretic general claim that is intended to 'express' it.

An approach that appears to be particularly 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' $\{A(\ulcorner B \urcorner) \rightarrow B \mid \ B \text{ a sentence of } \mathcal{L}_{\mathbb{N}}\}$ proves the same theorems without the truth predicate as T plus (i) the set of biconditionals $\mathrm{Tr}\ulcorner B \urcorner \leftrightarrow B$, for B truth-free, and (ii) the single sentence $\forall x(A(x) \rightarrow \mathrm{Tr}(x))$. A similar result holds for disjunctions. The claim is then that this result captures faithfully the interplay of EXPRESS and QUANTIFY.

This proposal articulates a clear and precise rendering of EXPRESS. However, it focuses on the mere material equivalence of A and $\mathrm{Tr}\ulcorner A \urcorner$, which we have already seen to be insufficient in the previous section. Moreover, 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. However, even if one only takes into account conjunctions or disjunctions separately, Halbach's result can already be obtained from the principle $\mathrm{Tr}\ulcorner A \urcorner \rightarrow A$, for A an *arbitrary* sentence of $\mathcal{L}_{\mathrm{Tr}}$, a principle that is often called Tr-out.[17] Now any T augmented with Tr-out only is consistent: one can for instance read Tr as 'is provable in T'. This indicates 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. Anil Gupta argued that 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). This is because there are certain tasks that the disquotationalist's truth predicate is bound to perform, or features that it should possess, which cannot obtain if EXPRESS is read in a weaker way, such as material or necessary equivalence.

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

(13) true beliefs about how to achieve goals tend to facilitate success in achieving them.

To affirm their explanatory role without violating EXPLAIN, deflationists have argued that general claims such as (13) *only* express an infinite conjunction of simple facts such as:

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

[17] That Halbach's result only needs Tr-out is observed also in print, by Picollo and Schindler (2017).

Therefore, if by doing Y S will achieve X, S 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, they would also explain this infinite conjunction.[18] 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 necessarily equivalent claims.

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).

I see no immediate way of formulating a logico-mathematical analysis of the kind of cognitive/conceptual equivalence discussed by deflationists. And one can safely predict that this is not an easy task. However, Gupta's analysis suggests a way out in the form of a necessary condition for such an equivalence. The infinite conjuction/disjunction and the corresponding quantified claims should at least be *equivalent for all theoretical/explanatory purposes.* It is then natural to understand the purported equivalence of (10) and (12) – or of (15) and (17) – in terms of formal criteria of *equivalence* for theoretical concepts. In the next section I consider notions of theoretical equivalence widely employed to compare scientific concepts. I take this analysis to be a necessary component of the deflationist's claim that an infinite conjunction and the corresponding truth-theoretic generalization should be faithful to the congnitive or conceptual equivalence involved in the Tr-biconditionals. If it turned out that, for instance, (10) and (12) are indeed equivalent in this sense, one would immediately ob-

[18] See (Horwich, 1998, pp. 22-23), (Williams, 1986, p. 232), and (Gupta, 1993, p. 65) for a discussion. In what follows, I will only require a *somewhat weaker* form of equivalence between 13 and 14 than sameness of meaning.

tain 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. This may not yet be a full vindication of their conceptual equivalence, but at least it would put the deflationist in the comfortable position of being able to consider (10) and (12) as equivalent for all relevant theoretical and explanatory purposes. By contrast, if they turned out to be inequivalent in this sense, then we would have evidence to doubt their conceptual/cognitive equivalence.

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 *bi-interpretability (or weak intertranslatability) and definitional equivalence (or synonymy/strong intertranslatability)*, and on some variants of them (Visser, 2006; Halvorson, 2019). Bi-interpretability is a slightly weaker notion than definitional equivalence, and will be mainly employed to strengthen some negative results. Crucially, the notion of bi-interpretability is essentially equivalent to the – also well-known – notion of Morita Equivalence (Halvorson, 2019, §7).

Both notions can be defined in terms of relative interpretations.[19] 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.

Bi-interpretability 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* (cf. Appendix B for the full definition). Informally, U and V are bi-interpretable if the interpretations K and L are such that, for any basic concept C of U, their interaction yields a new concept $C^{L\circ K}$ which is equivalent, up to U-provable isomorphism, to C, and similarly for primitives of V.

Even more succinctly, in bi-interpretations each theory can see that the interpretations involved can be combined to yield an *isomorphism* between the theory's primitive and interpreted concepts. Similarly, definitionally equivalent theories are such that the interpretations involved can be combined to yield an *identity* between the theory's primitive and interpreted concepts.

[19] I employ the definition of definitional equivalence given by Visser (2006). This definition may not coincide with the definition of Halvorson (2019) in full generality, but it does in the specific cases I will consider.

One can then readily see how definitional equivalence entails the explanatory equivalence of the two theories. Working in the theory U, suppose one's explanation E is supported by the $\mathcal{L}_U$-argument $\langle \varphi_0, \ldots, \varphi_n, \varphi \rangle$. Letting K, L be as above, to the question whether E is equally supported by the $\mathcal{L}_V$-argument $\langle \varphi_0^K, \ldots, \varphi_n^K, \varphi^K \rangle$, the U-theorist can answer by analyzing the status of the argument via her understanding of the V-concepts, that is by analyzing the status of the argument $\langle \varphi_0^{L \circ K}, \ldots, \varphi_n^{L \circ K}, \varphi^{L \circ K} \rangle$. Now, by definitional equivalence, this latter argument is simply $\langle \varphi_0, \ldots, \varphi_n, \varphi \rangle$, which indeed supports E by assumption. A parallel argument works for the V-theorist. Also, the example can be easily extended to bi-interpretations by taking into account isomorphisms.

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-definitional equivalence and Θ-bi-interpretability are then given in the obvious way by replacing interpretations with Θ-interpretations.

Although definitional equivalence is a standard notion of theoretical equivalence, one may wonder whether there are alternative notions of theoretical equivalence that are relevant in this context. For instance, whether there are any among the notions of theoretical reductions normally employed to compare theories of truth. The answer is negative: mutual interpretability, proof-theoretic equivalence, mutual faithful interpretability, mutual truth-definability (Halbach, 2014, I.6), are all inadequate to capture a strict form of equivalence between theories.[20]

2.2 *Expressing* general claims

I now turn to the main claims of this section. Let us work with a sufficiently expressive (consistent) base theory B containing the usual machinery for formal

[20] In particular, a theory S is always mutually interpretable with $S+$'S is inconsistent'. Mutual faithful interpretability suffers a similar fate: for many natural theories, such as finitely axiomatized sequential consistent theories, interpretability collapses into faithful interpretability (Visser, 2005, Cor. 5.6). Proof-theoretic reducibility may obliterate the distinction between base syntax theory and additional concepts: for instance, the typed Tarski-biconditionals over PA are proof-theoretically equivalent with PA, and similarly for many choices of truth axioms. Mutual truth-definability can identify substantially different truth predicates, for instance, the truth predicates of the disquotational theory PUTB, also discussed above, and the compositional theory KF (see also Nicolai (2017) for more details on this point). Such drawbacks are overcome by definitional equivalence and bi-interpretability. An interesting question which is left open by the present study is whether the notion of categorical equivalence – see again Halvorson (2019) – yields different verdicts.

syntax – Kalmar's Elementary Arithmetic is a safe lower bound (Nicolai, 2017). Given an infinite set of sentences S of $\mathcal{L}_B$ that is *definable* in B[21] – 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 $\mathcal{L}_{\mathrm{Tr}}$ with principles entailing suitable Tr-biconditionals *and* the single sentence $\forall x(\varphi_S(x) \rightarrow \mathrm{Tr}\, x)$.

In the following, given a base theory B, I take B^{Tr} to be a consistent, finite extension of B with truth theoretic principles that entails at least the Tr-sentences for $\mathcal{L}_B$. The first result is as follows:

Proposition 2. *Let B be finitely axiomatizable, and let S be an infinite set of sentences of $\mathcal{L}_B$ such that $\{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in \mathcal{L}_B\}$ cannot be finitely axiomatized over B. Then, $B + \{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in \mathcal{L}_B\}$ cannot be bi-interpretable with $B^{Tr} + \forall x(\varphi_S(x) \rightarrow Tr\, x)$.*

The proof is a straightforward application of Lemma 1 of Appendix B: if the two theories were bi-interpretable, then $\{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in \mathcal{L}_B\}$ 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 $\forall x(\varphi_S(x) \rightarrow \mathrm{Tr}\, x)$ cannot be definitionally equivalent.

Proposition 2 relies on the finiteness of B^{Tr}, both in the support theory B and in the truth-theoretic component of B^{Tr}. The first of these assumptions can be relaxed, as I will now show. The second will be considered right after.

The notions of $\mathcal{L}_B$-definitional equivalence and $\mathcal{L}_B$-bi-interpretability are defined in the same way as definitional equivalence and bi-interpretability, except that one replaces the interpretations involved with $\mathcal{L}_B$-interpretations. Such notions fit our discussion quite naturally: the infinite conjunction of the set of sentences S and its truth theoretic version are added to a common background theory providing uncontroversial syntactic tools. It then 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. By employing $\mathcal{L}_B$-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 $\mathcal{L}_{Tr}$-sentences that entails at least all the Tr-sentences for $\mathcal{L}_B$ over B. Furthermore, let S be an infinite set of sentences of $\mathcal{L}_B$ such that $\{\varphi_S(\ulcorner\psi\urcorner) \rightarrow \psi \mid \psi \in \mathcal{L}_B\}$ cannot be finitely axiomatized*

[21] By definability in B is intended here essentially the notion of weak representation in B of the set of sentences in S. Following a standard approach in the literature (Halbach, 1999), the restriction to *definable* sets of sentences formally captures the range of sets of sentences that are available in the deflationist's language, which are clearly less than all sets of sentences there are. I thank an anonymous referee for demanding clarifications on this point.

over B. Then, $B + \{\varphi_S(\ulcorner\psi\urcorner) \to \psi \mid \psi \in \mathcal{L}_B\}$ cannot be $\mathcal{L}_B$-bi-interpretable with $B + X + \forall x(\varphi_S(x) \to Tr\, x)$.

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

I now turn to the requirement – common to Propositions 2 and 3 – for the cluster of truth principles to be finitely presented. It may be objected that deflationists typically resort to an infinite formulation of their axioms for truth.[22] On closer inspection, there are two main issues connected with this objection. The first concerns the very existence of deflationary theories that are directly affected by the results. The second concerns the scope of the results: even if they indeed impacted on *some* deflationist theories of truth, they may still leave untouched the best, or most promising such theories.

The first issue is readily addressed. The literature abounds with examples of finite axiomatizations of the truth predicate that are directly motivated or advocated from a deflationary point of view.[23] Propositions 2 and 3 directly affect all those proposals. The second issue is more challenging: are there any adequate deflationist theories that aren't affected by these results?

A schematic, essentially infinite version of disquotation can be formulated in two main ways. As a set of biconditionals, or as inference rules of introduction and elimination. Depending on which conditional one employs in the formulation of the former and which logic one assumes in the background, the two formulations may come apart.[24] As explained above, I am here mainly concerned with *classical* disquotationalism, so I will not distinguish between the two formulations. Now, all of the available consistent, recursively given, essentially infinite set of Tr-biconditionals fall prey of the deductive weakness of material disquotation, which is debated at length in §1.[25] As we have seen, much work has been devoted to the recovery of some basic general claims that are out of reach for material disquotationalism, *in primis* the recovery of compositional

[22] See for instance (Horwich, 1998, p. 30).

[23] Just to mention a few: the compositional, typed theory of truth CT^- (aka $\mathrm{CT}\uparrow$) played a substantial role in the debate about the conservativeness argument and was explicitly endorsed by prominent deflationists (Field, 1999). A finite set of *positive, typed* axioms of truth has been advocated, from a deflationary standpoint, in Fischer and Horsten (2015). Their motivation can be extended to the type-free systems $\mathrm{KF}_{\mathrm{int}}$ and $\mathrm{KF}_{\mathrm{tot}}$ from Cantini (1989). The ω-consistent theories FS_n advocated by Halbach and Horsten (2005), are formulated with a finite set of compositional axioms and a finite set of Global Reflection Principles. Inferential deflationism, as articulated in Horsten (2012), clearly resorts to a finite presentation of the truth principles. Even modalized disquotationalism MPT, once properly regimented with a factivity *axiom* – as suggested in §1 – amounts to a finite set of sentences.

[24] The inferential formulation is common to many disquotational theories in many-valued logic. In fact, an anonymous referee has emphasized how the inferential formulation highlights the *logical* character assigned to truth by deflationism. I believe there is much to say also about this deflationist slogan; in forthcoming work with Johannes Stern this issue is studied with care.

[25] This is what Cieśliński calls *the generalization problem* (Cieśliński, 2017, §5).

principles. This is because they guarantee, either in typed or type-free form, the deduction of some desirable general claims of semantic or logical nature.

We might call *generalization core* this finite set of general claims. The generalization core should entail at least typed Tr-biconditionals, and include compositional principles and perhaps other truth theoretic general claims, such as the truth of all instances of some logical or mathematical axiom schemata.[26] Of course, this generalization core does not exhaust the kind of general claims the disquotationalist would like to express via truth, but contains only some desirable general claims of logico-mathematical character that she would like to establish *directly*. In fact, according to our analysis of EXPRESS, infinite lot of sentences that are available in the disquotationalist's language and that she would like to affirm should be equivalent, in a strong sense, to the claim that all such sentences are true. In the absence of a better formal rendering of this strong equivalence, I proposed – inspired by Gupta (1993) – to understand this relationship in terms of their theoretical/explanatory equivalence. Propositions 2 and 3 now tell us that, on the background of this minimal (and desired) generalization core, the infinite conjunction of an infinite lot of sentences and its truth-theoretic counterpart *cannot be theoretically/explanatorily equivalent*, let alone conceptually equivalent.

In other words, if deflationists already embrace a finite formulation of their truth principles, they are directly affected by the results above. If they instead resort to a schematic formulation of their Tr-biconditionals, *and* – as they do – endorse strategies to recover compositional principles and reach a minimal core of general claims, they should allow for the possibility of evaluating the theoretical/explanatory equivalence of an infinite set of sentences and its corresponding truth-theoretic generalization on the background of this finite generalization core. And once this is done, the verdict is fairly uncontroversial, the two are not theoretically/explanatorily equivalent.

Crucially, in evaluating Propostions 2 and 3, one should not forget that the focus is on the relationships between the infinite set of sentences one wishes to endorse and the corresponding truth-theoretic generalizations. It is not known whether the results can be lifted to an infinite formulation of the truth theory,

[26] As an anonymous referee points out, if one conceives of instances of logical axiom schemata in the language with the truth predicate, or even nonlogical axiom schemata in the language with the truth predicate, as *truth-theoretic* axioms, one would need to include them among the axioms of any adequate theory of truth, and no generalizing core would be finite. However, it is not uncontroversial that one should think of those schemata in this way. Logical principles do not acquire a different status depending on the language in which they are instantiated. For non-logical schemata, the issue is certainly more complex, as witnessed by the literature on the conservativess argument (Shapiro, 1998; Field, 1999; Ketland, 1999; Heck, 2018; Nicolai, 2015), and theorists disagree precisely on the nature of such nonological schemata. However, it is reasonable to say that they are at best of a mixed nature, mathematical/syntactc and truth-theoretic, and therefore there is room for evaluating the logical properties of a finite formulation of purely truth theoretic axioms only.

but this fact should not distract us. There is a finite set of truth principles that the disquotationalist strives to obtain to achieve some minimal generalizing power. Given those principles, our analysis gives a precise sense in which the strong sense of equivalence required by EXPRESS cannot obtain.

It is worth concluding this section by 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 Tr-biconditionals because of their simple and uncontroversial nature. As mentioned in the previous section, the choice of a consistent, type-free set of Tr-biconditionals involves a certain degree of arbitrariness. However, the result would still hold if for instance we replaced the Tarskian, typed biconditionals with (a finite theory entailing) (PT). We could also require in Proposition 2 the support theory to be any finite T such that $B \subseteq T \subseteq B^{\mathrm{Tr}}$, or consider consistent sets of sentences S of the full language $\mathcal{L}_{\mathrm{Tr}}$. Finally, both propositions still hold if we require only that the theory of truth only entails one direction of the disquotation schema.

2.3 The quantificational role of truth

Even if the deflationist's truth predicate cannot serve the expressive purposes one had hoped, there may still be room for the view that the truth predicate 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 arguments in the previous section do not settle the question whether a deflationist 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 the truth predicate 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 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 appeal to another negative result: the theoretical equivalence between disquotational truth and second-order quantification breaks 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 mutual 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 Tr-sentences that only involve sentences not containing Tr, on the one hand, and a form of *predicative*, second-order comprehension, on the other.[27] For simplicity, I here take B to be Peano arithmetic (PA), and add to its axioms the uniform Tr-schema

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

The resulting theory is known as 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^{-}.[28] 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

[27] 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.

[28] 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.

only the truth predicate via predication and vice versa. By employing the terminology introduced above, the two theories are mutually $\mathcal{L}_{\mathbb{N}}$-interpretable.

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 *bi-interpretable, 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.[29]

Back to QUANTIFY, one might hope to re-calibrate the role and purpose of the truth predicate by toning 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 the truth predicate as primitive, 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. Our best theory of modal disquotation, once formulated in accordance with the fundamental deflationary tenets, 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 that one would like to affirm and their corresponding truth theoretic generalizations stand in a close relationship that preserves their theoretical/explanatory status. In the most natural way of understanding this

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

relationship, that is via formal notions of theoretical equivalence, such a relationship 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 in a first-order setting, one requires the theoretical equivalence of truth and quantification. In the natural sense of definitional equivalence (and inter-translatability) or bi-interpretability (and Morita Equivalence), this equivalence cannot hold.

What has been said, of course, leaves open the possibility of regarding truth as a broadly logical *sui-generis* notion, and conceiving of deflationism as the formal and philosophical study of its principles. If, however, the deflationary approach to truth is charaterized by FIX, EXPRESS, QUANTIFY, and EXPLAIN, its chances of success are slim.

References

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

Beall, J. C. (2009). *Spandrels of Truth*. Oxford University Press.

Burge, Tyler (2011). Self and self-understanding. *The Journal of Philosophy*, 108:287–383.

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

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

Feferman, Solomon (1962). Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic*, 27(3):259–316.

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

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

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

Field, Hartry (2006). Compositional principles vs. schematic reasoning. *The Monist*, 89(1):9–27.

Field, Hartry (2008). *Saving truth from paradox*. Oxford University Press, Oxford.

Fischer, Martin and Horsten, Leon (2015). The expressive power of truth. *Review of Symbolic Logic*, 8(2):345–369.

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

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

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

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

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

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

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

Halbach, Volker and Horsten, Leon (2005). The deflationists' axioms for truth. In Beall, J. C. and Armour-Garb, Bradley, editors, *Deflationism and Paradox*. Oxford University Press.

Halbach, Volker and Nicolai, Carlo (2018). On the costs of nonclassical logic. *Journal of Philosophical Logic*, 47(2):227–257.

Halvorson, Hans (2019). *The Logic in Philosophy of Science*. Cambridge University Press.

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

Heck, Richard Kimberly (2018). The logical strength of compositional principles. *Notre Dame Journal of Formal Logic*, 59(1):1–33.

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

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

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

Horwich, Paul (2010). *Truth – Meaning – Reality*. Oxford University Press.

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

Lewy, Casimir (1947). Truth and significance. *Analysis*, 8(2):24–27.

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

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

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

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

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

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

Putnam, Hilary (1985). A comparison of something with something else. *New Literary History*, 17(1):61–79.

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

Quine, Willard 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, Stewart (1998). Proof and truth. *Journal of Philosophy*, 95(10):493–521.

Visser, Albert (2005). Faith & falsity. *Annals of Pure and Applied Logic*, 131(1):103–131.

Visser, Albert (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, Michael (1986). Do we (epistemologists) need a theory of truth? *Philosophical Topics*, 14(1):223–242.

Wright, Crispin (2004). Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume*, 78(1):167–212.

Appendix A: Modality

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

$$(R(t_1, \ldots, t_n))^* := R(t_1, \ldots, t_n) \qquad \text{with } R \text{ a relation of } \mathcal{L}_{\mathbb{N}}$$

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

Proof of Proposition 1. T is a classical theory in $\mathcal{L}_{\mathrm{Tr}}$. Logic only (in $\mathcal{L}_{\mathrm{Tr}}$) entails $\mathrm{Tr}l \vee \neg\mathrm{Tr}l$, where $\neg\mathrm{Tr}l$ is a liar sentence with l a closed term such that $l = \ulcorner\neg\mathrm{Tr}l\urcorner$ is provable in T. By (log1), T will also prove $\Box(\mathrm{Tr}l \vee \neg\mathrm{Tr}l)$; therefore $\mathrm{Tr}(\mathrm{Tr}l \vee \neg\mathrm{Tr}l)^*$ by (tfact*). By the nature of the mapping $(\cdot)^*$, therefore,

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

Since both $\mathrm{Tr}l^*$ and $\mathrm{F}l^*$ are $\mathcal{L}^+$-sentences, and T contains (PT), we can distribute the truth predicate over the disjunction to obtain

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

Now by the nature of l, (6), (7), (9), each disjunct entails $\mathrm{Tr}l^* \wedge \mathrm{F}l^*$.

□

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 returns 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'. Model-theoretically, a $K\colon T \to W$ provides a method for constructing, in any model $\mathcal{M} \vDash W$, an internal model $\mathcal{M}^K \vDash T$.

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

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

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

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

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

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:

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

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

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

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 – the identity interpretation on 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.*

Given the model-theoretic interpretation of relative interpretability mentioned after Definition 1, bi-interpretability also be characterized model-theoretically. With reference to Definition 3, U and V are bi-interpretable if and only if there are $K\colon U \to V$ and $L\colon V \to U$, a $\mathcal{L}_V$-formula F_0, and a $\mathcal{L}_U$-formula such that: for any model $\mathcal{M} \vDash V$, F_0 defines an isomorphism between $\mathcal{M}$ and $\mathcal{M}^{K\,L}$, and for any model $\mathcal{N} \vDash U$ F_1 defines an isomorphism between $\mathcal{M}$ and $\mathcal{N}^{L\,K}$.

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 bi-interpretability. Thus $U_0 \vdash A$ by (i). □

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 $\mathcal{L}_B^+$ that finitely expands $\mathcal{L}_B$. Moreover, let T_1 be $B + Y$, where Y is finitely axiomatizable over B in $\mathcal{L}_B^+$. If there is a T_0-isomorphism I 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 (in $\mathcal{L}_B^+$).*

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 $\mathcal{L}_{T_0}$-sentence C, if $T_0 \vdash C$, then $B + A \vdash C^K$. But then also $T_0^* \vdash C^{K\,L}$, 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 is folklore (Nicolai, 2017).

Proof of Proposition 4 . It is useful for to move to a relational formulation of $\mathcal{L}_{\mathbb{N}}$. This is a generalization of an unpublished argument by Albert Visser and Ali Enayat. It is originally contained in Nicolai (2017).

Seeking a contradiction, let's assume that ACA^- and UTB are bi-interpretable. Now let us consider the two-sorted structure $(\mathbb{N}, \mathrm{P}(\omega)) \vDash \mathrm{ACA}^-$. By assumption, given the model-theoretic characterisation of bi-interpretability, in $(\mathbb{N}, \mathrm{P}(\omega))$ we can find an internal model $(\mathcal{M}, S) \vDash \mathrm{UTB}$ – $S \subset M$ being the extension of Tr – that, in turn, contains a model $(\mathcal{N}, \mathcal{R}) \vDash \mathrm{ACA}^-$ 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 $\mathcal{L}_{\mathrm{Tr}}$ 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 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 fact that $(\mathcal{M}, S)$ interprets the uncountable model $(\mathbb{N}, \mathrm{P}(\omega))$. □