Typed and untyped disquotational truth

Cezary Cieśliński
Institute of Philosophy, the University of Warsaw,
Poland

Abstract

We present an overview of typed and untyped disquotational truth theories with the emphasis on their (non)conservativity over the base theory of syntax. Two types of conservativity are discussed: syntactic and semantic. We observe in particular that TB - one of the most basic disquotational theories - is not semantically conservative over its base; we show also that an untyped disquotational theory PTB is a syntactically conservative extension of Peano Arithmetic.

1 Disquotational truth theories

Disquotationalists believe that the whole content of the notion of truth is captured by the so called schema T:

(T) $\quad$ 'p' is true iff p.

The general intuition is that the result of adding "is true" to a name of a sentence (or an utterance, or a proposition) is equivalent – in some weaker or stronger sense – to the sentence (the utterance, the proposition) itself. Apart from that, no further explanation of the meaning of "is true" is needed. Thus e.g. Field in [5] states that for a given person and an utterance u, "the claim that u is true (true-as-he-understands-it) is cognitively equivalent (for the person) to u itself (as he understands it)" (p. 250); and he proceeds to assert that "the cognitive equivalence of the claim that u is disquotationally true to u itself provides a way to understand disquotational truth independent of any nondisquotational concept of truth or truth conditions" (p. 251). Another example of a philosophical position arising from disquotational intuitions is Horwich's minimalism (see his [10]). According to Horwich, all the facts about truth can be explained on the basis of the so called "minimal theory",

whose axioms have a disquotational form.[1] Horwich claims that the minimal theory fully characterizes the content of the notion of truth. In particular, to understand the truth predicate it is simply enough to be ready to accept the substitutions of the relevant T-schema. Accordingly, truth has no hidden nature which could and should be revealed by scientific enquiry ([10], p. 2). In effect all the traditional, substantive conceptions of truth (like correspondence, coherence and warranted assertability theory) turn out to be useless at best and probably misleading.

In this context disquotationalists quite often cite Alfred Tarski, recalling his famous Convention T. Why are we ready to call Tarski's definition a definition of truth, and not of some other property? The reason is that it permits us to derive all the instances of the T-schema for sentences of the object language. Tarski gives us a clear hint: his predicates (for various languages) are *truth* predicates, because they satisfy the same condition of material adequacy, i.e. they conform to Convention T. After we recognize this, there is only one last step to be taken: one can declare that all the apparatus of classical semantics (compositional approach, involving an inductive characterization of reference and satisfaction) is really unnecessary in the context of our general project of explaining the intuitive notion of truth. "Truth has a certain purity" (see [10], p. 12) – we can explain it directly in terms of T-schemata without any appeal to other semantic notions.

What's the form of the theory of truth most adequate to disquotational intuitions? The most direct approach consists in stipulating that the set of axioms of our theory of truth will take the form of a collection of chosen instantiations of a T-schema.[2] Of course due to the contradictions of the liar type, we can't have all substitutions on our list – some restrictions are necessary. Nevertheless, the disquotationalist will claim that no other sort of axioms is really needed.

There are two basic variants of disquotational theories of truth. One possibility consists in adopting as axioms the substitutions of the local T-schema:

(L) $\quad T(\ulcorner\varphi\urcorner) \equiv \varphi$

The second option is to adopt a schema of uniform disquotation:

[1] It's worth mentioning that Horwich attributes truth to propositions, not sentences or utterances.

[2] Admittedly, it is not the only possible option. Cf. Beall [1], where disquotationalism is understood as a view that truth is a "fully transparent device" (p.3); more exactly: it's "a device introduced via rules of intersubstitution: that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable in all (nonopaque) contexts" (p. 1). On this approach, adopting T-biconditionals as axioms might be just one of the possible ways to give justice to the disquotationalist's intuitions.

(U) $\forall a_1...a_n[T(\ulcorner\varphi(a_1...a_n)\urcorner) \equiv \varphi(a_1...a_n)]$

In the second case our axioms are formulas obtained from (U) by substituting concrete variables for $a_1...a_n$ and concrete formulas for a schematic letter φ. In fact (L) can be viewed as a special case of (U), with φ being a sentence and the sequence of quantified variables being empty.

Comment. Using (U) instead of (L) retains a lot of the disquotationalist spirit, although one could complain that it is more a satisfaction than a truth schema. The intended meaning of "$T(\ulcorner\varphi(a_1...a_n)\urcorner)$" is after all that a formula $\varphi(x_1...x_n)$ is satisfied by objects $a_1...a_n$. Admittedly, in an arithmetical context, where every object has a standard numeral denoting it, we can express this thought employing just a one place truth predicate: we say in effect that the result of substituting in φ numerals for $a_1...a_n$ is true. However, in other contexts, where nameability assumption can't be employed, we would have to use a satisfaction predicate instead.

In what follows both types of disquotational axioms will be discussed, in two variants: typed and untyped one. I will concentrate on arithmetical context, taking PA as the base theory of syntax and stressing each time the arithmetical strength of the resulting theory.

2 Conservativeness

The main emphasis will be on (non)conservativity results. Conservativeness has been one of the major issues in recent debates about truth theories. Should we expect from a theory of truth that it be conservative over its base? The opinions have been divided. On the one hand, some philosophers of deflationary bent claimed that truth is an innocent and metaphysically thin notion. An explication of this claim has been proposed by Horsten in [9] and elaborated by Shapiro [16] and Ketland [13]. On this view, an adequate theory of truth for a given language should conservatively extend a base theory of syntax for this language. The motivation for accepting conservativeness demand is succinctly formulated in the following fragment of Shapiro's paper:

> How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers?

(see [16] p. 499.) As we see, the underlying intuition is that if by invoking the notion of truth we can learn more about natural numbers, then the notion of truth is not thin. In the next step the notion of conservativeness is used to analyze the situation in more detail. A representative fragment from Shapiro's paper runs as follows:

> I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence Φ in the original language (so that Φ does not contain the truth predicate) such that Φ is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet Φ false, but it is not logically possible for the axioms of B' to be true and Φ false. This undermines the central deflationist theme that truth is in-substantial. (Shapiro [16], p. 497)

Observe that although in the quoted passage the claim of insubstantiality of truth is explicated in terms of conservativeness, Shapiro remains noncommittal about a particular form of a conservativeness demand. (In fact various versions of the demand can be considered; see below, Definition 1.)

Others have argued that the deflationists have no reason to embrace conservativeness as a condition on truth theories; in addition, a theory of truth with this property would be too weak.[3] I am not going to engage into this debate here; I stress only that results about arithmetical strength of truth theories are philosophically important no matter what one's standpoint in the debate is.[4]

Let us introduce now the definition of two notions of conservativeness.

Definition 1 *Let T_1 and T_2 be theories in languages L_1 and L_2 (with $L_1 \subseteq L_2$). Then:*

(a) T_2 is syntactically conservative over T_1 iff $T_1 \subseteq T_2$ and $\forall \psi \in L_1[T_2 \vdash \psi \rightarrow T_1 \vdash \psi]$.

(a) T_2 is semantically conservative over T_1 iff every model M of T_1 can be expanded to a model of T_2 (interpretations for new expressions of L_2 can be provided in M in such a way as to make T_2 true).

The two notions of conservativeness do not coincide. Semantic conservativeness is a more general notion: it gives via completeness theorem the

[3]See e.g. Halbach [8], p. 188: "As far as I can see, neither have deflationists subscribed to conservativeness explicitly nor does it follow from one of their other doctrines. (...) But if the deflationist understands his claim that truth is not a substantial notion as implying that his truth theory has no substantial consequences, he commits a mistake."

[4]For a philosophical discussion of conservativeness as a demand for deflationary truth theories, see also [3], [12] and [17].

syntactic version, but the opposite implication does not hold. Examples of truth theories being syntactically, but not semantically conservative over their base theories will be given below. Both notions are invoked by Shapiro in [16]. Later however most of the authors writing on the subject concentrated almost exclusively on the syntactic notion, ascribing to the deflationist a commitment to syntactic conservativeness. One of the few pleas for semantic conservativeness can be found in McGee [14].

3 Typed disquotation

I will discuss typed disquotational theories in two variants: local and uniform one.

3.1 Typed uniform disquotation

Adopting the typed approach, we obtain a Tarskian hierarchy of truth predicates and a family of theories characterizing the notion of truth for languages with truth predicates of all lower levels. Let L_0 be the language of Peano arithmetic; let L_{n+1} be the extension of L_n with a new one place predicate "T_n". Denote by $Ind(L_n)$ the set of all induction axioms for formulas of the language L_n. Then we define ("UTB" reads "uniform Tarski biconditionals"):

Definition 2

- $UTB_0 = PA$
- $UTB_{n+1} = UTB_n \cup \{\forall a_1 ... a_n [T_n(\ulcorner \varphi(a_1 ... a_n) \urcorner) \equiv \varphi(a_1 ... a_n)] : \varphi \in L_n\} \cup Ind(L_{n+1})$

(Observe that UTB_n is always in the language L_n.) Then the following result can be obtained:

Theorem 3 *For every n, UTB_{n+1} is syntactically conservative over UTB_n.*[5]

Proof. Assume that $UTB_{n+1} \vdash \varphi$, $\varphi \in L_n$. Consider all disquotational axioms employing T_n in a (fixed) proof of φ. Let $\psi_0 ... \psi_i$ be all formulas mentioned in these axioms in the scope of T_n (i.e. every such an axiom has a form "$\forall a_1 ... a_n [T_n(\ulcorner \psi_k(a_1 ... a_n) \urcorner) \equiv \psi_k(a_1 ... a_n)]$" for some $k \leq i$). Taking m as a maximal quantifier rank of $\psi_0 ... \psi_i$, we observe that that there is a predicate "$Tr_m(x)$" of the language L_n, which is a truth predicate for formulas of L_n with a quantifier rank smaller or equal m.[6] Since UTB_n proves

[5] Cf. Halbach [6], p. 55, where the proof is given that UTB_1 is conservative over PA.

[6] On partial truth predicates, see Kaye [11], p. 119ff.

all biconditionals of the form "$\forall a_1...a_n[Tr_m(\ulcorner\psi_k(a_1...a_n)\urcorner) \equiv \psi_k(a_1...a_n)]$" for $k \leq i$, the proof of φ can be reconstructed in UTB_n by substituting "Tr_m" for "T_n" and by supplying proofs for the resulting biconditionals when necessary. □

Before analyzing the semantic conservativeness property, I want to remind the reader an important notion of a recursively saturated model.

Definition 4

- *Let Z be a set of formulas with one free variable x and parameters $a_1...a_n$ from a model M. Z is realized in M iff there is an $a \in M$ such that every formula in Z is satisfied in M under a valuation assigning a to x.*
- *Z is a type of M iff every finite subset of Z is realized in M.*
- *M is recursively saturated iff every recursive type of M is realized in M.*

It is a well known fact that every infinite model is elementarily equivalent with a recursively saturated structure (see e.g. Kaye [11], Proposition 11.4, p. 14).

Theorem 5 *UTB_{n+1} is not semantically conservative over UTB_n.*[7]

Proof. The proof consists in observing that only recursively saturated models of UTB_n can be expanded to models of UTB_{n+1}. Given a model M_1 of UTB_n, assume that it's possible to expand it to a model $M_2 = (M, T_0...T_n)$ in such a way that $M_2 \models UTB_{n+1}$. Let $p(x, a_1...a_n)$ be a recursive type over M_1. Let "$s \in p$" be an arithmetical formula representing in PA the recursive set of formulas (without parameters) used in forming the type $p(x, a_1...a_n)$. Then we have:

$$\forall k \in NM_2 \models \exists z \forall \varphi(x, y_1...y_n) < k[\varphi(x, y_1...y_n) \in p \rightarrow T_n(\varphi(z, a_1...a_n))]$$

So by overspill, there is a nonstandard $b \in M_2$ such that:

$$M_2 \models \exists z \forall \varphi(x, y_1...y_n) < b[\varphi(x, y_1...y_n) \in p \rightarrow T_n(\varphi(z, a_1...a_n))]$$

Then such a z realizes our type $p(x, a_1...a_n)$ in M_1.[8]

□

[7] Cf. [11], p. 228, Proposition 15.4.

[8] Although we worked in M_2, the transition to M_1 is made possible by the fact that all formulas in our type belong to the language L_n, i.e. they do not contain "T_n", so if they are satisfied in M_2, they are also satisfied in M_1.

3.2 Typed local disquotation

In an analogous manner, we define now the hierarchy of typed theories based on the local disquotational schema.

Definition 6

- $TB_0 = PA$
- $TB_{n+1} = TB_n \cup \{T_n(\ulcorner\varphi\urcorner) \equiv \varphi : \varphi \in L_n\} \cup Ind(L_{n+1})$

Since local disquotation is a special variant of uniform disquotation, some results from the last subsection carry over to the present case. In particular, Theorem 3 applies without any changes – TB_n is syntactically conservative over TB_k for $k < n$. As for Theorem 5, although its proof doesn't carry over to our present case, the result still holds.

Theorem 7 *TB_{n+1} is not semantically conservative over TB_n.*[9]

Before giving the proof, I would like to remind the reader some basic concepts, which will be used later on.

Explanation 1 (the notion of coding). A set Z of natural numbers is coded in a model M by an element a of this model iff $Z = \{n : M \models n \in a\}$. Expression of the form "$x \in y$" is taken to be an arithmetical formula used for the purposes of coding; it can be e.g. "$p_x \mid y$" ("the x^{th} prime divides y"). In the standard model it is exactly finite sets which are coded. The situation is different in nonstandard models, where some infinite sets will be coded as well.[10]

Explanation 2 (the notion of a prime model). Let S be a consistent extension of PA in the language L with new predicates $\tilde{A}_1 ... \tilde{A}_n$, with full induction for L. Let $M^* = (M, A_1 ... A_n)$ be a model for S. A *prime model* K of S is obtained from M^* in the following manner:

- The universe of K is defined as $\{a \in M$: a is definable in M^* by some formula of $L\}$
- The operations of K are the operations of M^* restricted to the universe of K

[9] After obtaining Theorem 7, I found out that the result was proved earlier by Fredrik Engström. Engström's work is unpublished.

[10] For more about coded sets, see e.g. Kaye [11] p. 141ff.

- for $i \leq n, A_i^K = A_i \cap K$.

It is possible to show that K is an elementary submodel of M^*, with all elements of K being definable in K.[11]

We now start with the following lemma:

Lemma 8 *For every $n \in N$, the following conditions are equivalent for an arbitrary nonstandard model $M^* = (M, T_0...T_{n-1})$ of TB_n:*

(a) M^* can be expanded to a model of TB_{n+1}

(b) M codes $Th(M^*)$.

Proof. For the direction from (b) to (a), assume that a is a code of $Th(M^*)$ in M. Define: $T_n = \{x \in M : M \models \text{“}x \in a\text{”}\}$. Then $(M, T_0...T_n) \models TB_{n+1}$ as required (observe in particular, that T_n is inductive, since it's definable with parameters in M). For the opposite direction, assume that M^{**} is an expansion of M^* satisfying TB_{n+1}. Then we have:

$$\forall k \in N M^{**} \models \exists z \forall s[s \in z \equiv (s < k \wedge T_n(s))]$$

Therefore by overspill there is a nonstandard $a \in M$ such that:

$$M^{**} \models \exists z \forall s[s \in z \equiv s < a \wedge T_n(s)]$$

(Observe that overspill can be used, because we assumed that T_n is inductive.) Picking such a z, we obtain a code for $Th(M^*)$ in M, as required. □

With Lemma 8 at hand, the proof of Theorem 7 is immediate.

Proof of theorem 7. Let $M^* = (M, T_0...T_{n-1})$ be a prime nonstandard model of TB_n. We show that it can't be expanded to a model of TB_{n+1}. For an indirect proof, assume that it can. Then by Lemma 8, M codes $Th(M^*)$, and since it's prime, a code c of $Th(M^*)$ is definable in M^*. Take a formula $\alpha(x)$ defined as:

$$\alpha(x) := \exists z[\psi(z) \wedge x \in z]$$

with $\psi(x)$ being a formula of L_n which defines c in M^*. It's easy to observe that $\alpha(x)$ is a truth predicate of the language L_n for L_n sentences in M^*, which contradicts Tarski's indefinability theorem. □

[11] More information about prime models can be found in Kaye [11], p. 91ff.

4 Untyped disquotation

If we decide to drop the typing restrictions, the situation may change drastically, depending on our choice of the substitution class for the T-schemata. Even a seemingly weaker schema (L) can generate quite powerful theories once a suitable set of instances is selected. The key observation was made by Vann McGee in [15]. Consider an arbitrary sentence φ of the arithmetical language extended with the truth predicate (it will be denoted as L_T). Let PAT be a theory in the language L_T whose all extralogical axioms are just those of PA. Then there is a substitution of (L) which is provably (in PAT) equivalent with φ. The method of finding an appropriate substitution of (L) is effective; it is also possible to "decode" effectively the sentence φ given a corresponding substitution of (L). In effect we obtain the following:

Theorem 9 *Let H be an arbitrary recursive set of sentences of the language L_T. Then there is a recursive set G of substitutions of (L) such that H and G are (over PAT) recursive axiomatizations of the same theory.*

Superficially, Theorem 9 might look like a great news for the disquotationalist. Whatever your favourite theory of truth is, you can always axiomatize it by substitutions of (L). Nothing else is needed! However, in fact McGee's result leads the disquotationalist nowhere. The main problem is that the disquotationalist wants to treat the substitutions of the T-schemata as epistemologically basic. Whatever more substantial principles of truth we accept, he plans to justify them by recourse to the T-schemata, and not the other way round. (In particular, it won't do to justify the acceptance of a given set of substitutions by saying that they are equivalent to the axioms of our favourite (substantial) theory of truth.) Unfortunately, McGee's result shows, that in general there is nothing basic about the schema (L). *False* sentences are provably equivalent to substitutions of (L); arithmetical truths unknown to us are also provably equivalent to such substitutions. It seems that (in many cases) accepting a given substitution of (L) requires a special argument, which goes beyond a mere saying that it is a substitution of a disquotational schema.

In short: the disquotationalist needs to characterize a set S satisfying the following conditions: (1) S is a recursive set of substitutions of a T-schema (the local or the uniform one) (2) we have good reasons to treat elements of S as epistemologically basic. In particular, we do not accept S because of its equivalence (over PAT) with some substantial truth theory of our choice.

The disquotationalist's predicament is that it seems quite difficult to find a comprehensive set S satisfying (1) and (2).[12] The difficulty will be illustrated below, by considering a concrete candidate for the role of such an S: a set of *positive* substitutions of a T-schema.

4.1 Untyped uniform disquotation

The proposal is described by Volker Halbach in [7]. It arises from an analysis of the way paradoxes are produced. The initial insight is that in paradoxical reasonings we apply the truth predicate to sentences containing a negated occurrence of this predicate (see [7], p. 788). This is plainly the case with the liar sentence: the standard, diagonal construction of the liar produces a sentence with "T" within a scope of one negation. In effect one could try to avoid the paradoxes by restricting the set of substitutions of (U): from now on we admit *positive* substitutions only, i.e. our axiom is whatever can be obtained from (U) by substituting a positive formula for a schematic letter φ.

The notion of a positive formula is defined for a language containing $\neg, \wedge$ and $\vee$ as the only connectives. (Implication is not a primitive symbol. A reflection on Curry's paradox forces us to treat apparently positive occurrences of "T" in an antecedent of an implication as negative.[13]) From now on we stipulate that L_T (the extension of the language of PA with the truth predicate) contains just those connectives. Then we define:

Definition 10

(a) A formula φ of L_T is positive iff every occurrence of "T" in φ appears in the scope of an even number of negations.

(b) PUTB ("positive uniform Tarski biconditionals") is a theory axiomatized by all axioms of PAT with extended induction and all substitutions of the uniform truth schema (U) by positive formulas.

Halbach's main theorem characterizes the arithmetical strength of PUTB. Far from being conservative over PA, PUTB is arithmetically very strong –

[12] One path could consist in considering maximal conservative sets of substitutions of a T-schema. It has been shown however, that there are uncountably many such sets and none of them is axiomatizable. See Cieśliński [4].

[13] In Curry's paradox we consider a sentence ψ satisfying the condition: $\psi \equiv [T(\ulcorner \psi \urcorner) \rightarrow 0 = 1]$. It turns out then that adopting a T-biconditional for ψ results in a contradiction. However, a Curry sentence ψ constructed by diagonalization contains an occurrence of the truth predicate which is not negated.

it is in fact arithmetically equivalent to the Kripke-Feferman theory KF, one of the strongest truth theories discussed in contemporary literature.[14]

Theorem 11 $\forall \psi \in L_{PA}[PUTB \vdash \psi \equiv KF \vdash \psi]$

The proof consists in showing that PUTB defines the truth predicate of KF, i.e. there is a formula $\alpha(x)$ such that PUTB proves all sentences obtained from axioms of KF by replacing the truth predicate $T(x)$ with $\alpha(x)$. Together with the information that $PUTB \subseteq KF$, this implies Theorem 11. For details, see [7] (lemma 4.3 and theorem 5.1). It's also worth mentioning, that nevertheless PUTB is truth-theoretically weaker than KF – it doesn't prove compositional truth axioms (Halbach [7], lemma 6.1 and below).

Halbach ended his paper with an open question about the arithmetical strength of the theory taking as axioms all positive substitutions of the *local* truth schema (L). I will sketch the answer in the next subsection.

4.2 Untyped local disquotation

Let's consider now a case of a positive local disquotation. The basic definition is as follows.

Definition 12 *PTB ("positive Tarski biconditionals") is a theory axiomatized by all axioms of PAT with extended induction and all substitutions of the local truth schema (L) by positive sentences.*

We formulate now the main theorem about PTB.

Theorem 13 *PTB is syntactically conservative over Peano Arithmetic.*[15]

The proof consists in showing that:

(*) For every finite set S of axioms of PTB, for every recursively saturated model M of Peano arithmetic, M can be expanded to a model of S (i.e. an interpretation of the truth predicate can be found in M in such a way as to make all sentences in S true).

[14] A presentation of KF can be found in Halbach [6], starting on p. 195.

[15] The proof of Theorem 13 was presented on the "Truth be told" conference in Amsterdam (2011), and also in [2].

After (*) is obtained, Theorem 13 follows immediately.

Proof of Theorem 13 from (*). Assume that $PTB \vdash \varphi$ for an arithmetical sentence φ. Then there is a finite set S of axioms of PTB such that $S \vdash \varphi$. By (*), every recursively saturated model of PA can be expanded to a model of S; therefore φ is true in every recursively saturated model of PA. But every model of PA is elementarily equivalent with a recursively saturated model, therefore φ is true in every model of PA, which by completeness implies that $PA \vdash \varphi$. $\square$

Sketch of the proof of (*). A handy tool in this proof is a notion of a translation function $t(a, \psi)$, which takes as arguments a number a (possibly nonstandard) from a given model and a formula ψ belonging to the language with the truth predicate. The value of this function is an arithmetical formula (no "T" inside) with a parameter a – a *translation* of ψ. The translation is obtained by substituting all occurrences of "$T(t)$" in ψ by "$t \in a$" – an arithmetical formula used for the purposes of coding sets (see Explanation 1). In effect the translation interprets the truth predicate in ψ as referring to the set coded by a.

With the notion of a translation at hand, we can define, for a recursively saturated model M, a family of recursive types over M, a family of elements realizing these types and of models expanding M with an interpretation of the truth predicate. In what follows the predicates $Sent_{PA}$ and $Sent_T^+$ denote (respectively) the set of all sentences of the language of PA and the set of all positive sentences of the language L_T.

Definition 14

1.
- $p_0(x) = \{\varphi \in x \equiv \varphi : \varphi \in Sent_{PA}\} \cup \{\forall w(w \in x \Rightarrow w \in Sent_{PA}\}$
- d_0 *realizes* $p_0(x)$
- $T_0 = \{a : M \models a \in d_0\}$
- $M_0 = (M, T_0)$

2.
- $p_{n+1}(x, d_n) = \{\varphi \in x \equiv t(d_n, \varphi) : \varphi \in Sent_T^+\} \cup \{\forall z(z \in d_n \Rightarrow z \in x)\} \cup \{\forall z(z \in x \Rightarrow z \in Sent_T^+)\}$
- d_{n+1} *realizes* $p_{n+1}(x, d_n)$
- $T_{n+1} = \{a : M \models a \in d_{n+1}\}$
- $M_{n+1} = (M, T_{n+1})$

The idea behind Definition 14 is as follows. A number d_0 obtained at the start codes the set of all arithmetical sentences true in M – we denote it as T_0. Building a model M_0, we interpret the truth predicate with just this set.[16] In the next stage we obtain a number d_1 coding the set of all positive sentences of the language L_T, which become true in M once "T" is interpreted by a set T_0 (i.e. once "$T(t)$" is replaced by "$t \in d_0$"). And then we iterate the construction for all natural numbers.

At this moment two observations become useful. The first is that $T_0 \subseteq T_1 \subseteq T_2$.... The second is a general fact about positive formulas: if A and B are subsets of the universe of a model M with $A \subseteq B$, then every positive formula satisfied under some valuation in (M, A) will be also satisfied in (M, B). From these two observations it follows that given a finite set $S = \{T(\ulcorner\varphi_0\urcorner) \equiv \varphi_0 \ ... \ T(\ulcorner\varphi_k\urcorner) \equiv \varphi_k\}$ of axioms of PTB, there will be a natural number n such that $M_{n+1} \models S$. We simply find an n such that:

$$\forall i \leq k[M_n \models \varphi_i \vee \neg\exists l \in N M_l \models \varphi_i]$$

and then observe that all the equivalences from S are true in M_{n+1}. Since T_{n+1} is definable with parameters in M (by a formula "$x \in d_{n+1}$), M_{n+1} satisfies also all the induction axioms for the extended language. □

Although PTB is syntactically conservative over PA, it doesn't have the semantic conservativeness property. This follows easily from Theorem 7.

Corollary 15 *PTB is not semantically conservative over PA.*

Proof. Otherwise, since TB_1 can be treated as a subtheory of PTB, every model of PA could be expanded to a model of TB_1, contrary to Theorem 7. □

5 Justification of disquotational axioms

Is disquotationalism a philosophically attractive position? An answer to this question depends on the one hand, on the assessment of the strength of disquotational theories, and on the second, on the justification of disquotational axioms. In these final comments I want to concentrate on the second issue. As we saw, the key move consists in choosing a substitution class S for a

[16] Strictly speaking, T_0 will contain not only (codes of) arithmetical sentences true in M, but also these nonstandard numbers a, for which the formula "$a \in d_0$" is satisfied in M.

T-schema (local or uniform one). In view of Theorem 9, the choice of S must be well motivated – it won't do in general to accept S as a set of "mere substitutions of a T-schema" (see remarks below Theorem 9). What motivations can be offered?

Restoration of the consistency of disquotational theory is a natural aim. Naive, unrestricted T-schema generates a contradiction – that's a fact to which all truth theorists must react and the disquotationalist is no exception. Restoring the consistency of a theory of truth should be treated as a permissible motivation for the disquotationalist to proceed. The question is only how far it can take us.

I will discuss two worries concerning this type of justification. Eventually I will dismiss the first one; in contrast, the second seems to me a real issue.

Objection 1 (cf. Halbach [6], p. 311). The disquotationalist should not only guarantee that his theory is safe from a contradiction, but he must do it by using safe proof methods. In particular, since his aim is to characterize a satisfactory notion of truth, he shouldn't be allowed to use model theoretic arguments. A model theoretic notion of truth goes beyond the disquotational notion, therefore he can't employ it. Perhaps he will achieve his aim for typed theories: using the means available in PA, he can prove e.g. relative consistency of UTB_1 (i.e. he can prove in PA that if PA is consistent, then UTB_1 is consistent), so he is entitled to claim that his trust in the consistency of UTB_1 is no less warranted than his trust in PA itself. But the problem is that this approach fails as a general strategy. We can't do the same for strong disquotational theories, whose relative consistency is not provable in PA.

Answer. I can see no reason why model theoretic means shouldn't be available to the disquotationalist, discussing the notion of arithmetical truth. Observe that the notion of truth in a model can be expressed by set theoretical means; the completeness theorem can also be viewed as a set theoretical result. Questioning set theory is not a part of the disquotationalist's baggage; he questions rather a substantial notion of truth. As I take it, he is free to use "truth in a model" as a technical notion, useful for obtaining consistency results. He can just add: "this is not the same as the concept of (unrelativized) truth-as-we-understand-it, which I try to characterize by means of a T-schema. These are just two different concepts".

Although I take the above answer as basically correct, one qualification is needed. The real issue is not whether the disquotationalist can use set theory with its notion of model theoretic truth (he can!); the issue is rather

how he uses it. In particular, I would find his employment of the notion of truth in the standard (or the intended) model quite problematic. If a given philosophical argument hinged on identifying truth-as-we-understand-it with truth in the intended model, it would seem indeed that a stronger (non-disquotational) notion of truth is needed to make the argument work. As we will see, this provides a basis for the second and more serious objection.

Objection 2. It is not enough that a disquotational theory be consistent. If a T-schema is to be treated as epistemologically basic, some argument is needed to show that the obtained theory is arithmetically sound. By Theorem 9, false arithmetical sentences are also provably equivalent to substitutions of (L). Why should we trust that the disquotational theory of our choice doesn't produce false results?

Comment. An advocate of typed disquotation – say, of UTB_1 – could retort that PA proves not only relative consistency, but also conservativeness of UTB_1 over PA. In effect we (as users of PA) are entitled to trust UTB_1 just as we are entitled to trust PA. No strong notion of truth is needed to establish that.[17]

However, the situation of an advocate of an untyped theory like PUTB is more problematic. A natural move could consist in arguing that PUTB admits a standard interpretation – that it's possible to interpret the truth predicate of PUTB in the intended model of arithmetic. But the problem with this rejoinder is that in its employment of the notion of truth in the standard model, it goes beyond the legitimate disquotational means. The disquotationalist can't argue "my axioms are trustworthy, since they produce true arithmetical results, which I know because they are true under the intended interpretation". In saying this he commits himself to a stronger notion of arithmetical truth than the disquotational one. And that's his predicament.

6 Problems

I end with listing what I take to be the main problems in this area of research. The problems are:

[17] In this respect the situation of an adherent of a typed disquotational theory is quite comfortable; his problems lie elsewhere: in the deductive weakness of his theory.

(1) Is there a natural substitution class for (U), which could be used to obtain not only the arithmetical content of KF, but also its compositional principles?

(2) Are there any plausible candidates for the role of a natural substitution class for (L), producing an arithmetically strong theory?

(3) Is there a disquotationally acceptable answer to the question "why should we trust positive disquotational axioms"?

(4) Is there an argument for conservativity of PTB over PA, which can be formalized in PA?

Questions (1)-(3) are philosophical; question (4) is formal. (1) relates to the fact that PUTB, although arithmetically strong, is quite weak in proving compositional principles (see [7], pp. 793ff). Admittedly, it is not very clear what classes should count as "natural". The intuition is that principled, non ad-hoc reasons should stand behind selecting such a class. Question (2) is motivated by our observation, that PTB is conservative over PA, so positive substitutions of (L) do not take us very far (are there other good candidates worth considering?) Question (3) is in effect whether a good answer to Objection 2 can be given. For question (4), observe that the proof of conservativity of PTB over PA given here is semantic and doesn't translate easily to a syntactic argument.

References

[1] Beall, Jc (2009) *Spandrels of truth*, Oxford University Press, Oxford, New York.

[2] Cieśliński, Cezary (2011) "T-equivalences for positive sentences", *The Review of Symbolic Logic* 4, 319-325.

[3] Cieśliński, Cezary (2010) "Truth, Conservativeness, and Provability", *Mind* 119, 409-422.

[4] Cieśliński, Cezary (2007) "Deflationism, conservativeness and maximality", *Journal of Philosophical Logic* 36, 695-705.

[5] Field, Hartry (1994) "Deflationist Views of Meaning and Content", *Mind* 103, 24984.

[6] Halbach, Volker (2011) *Axiomatic theories of truth*, Cambridge University Press, Cambridge.

[7] Halbach, Volker (2009) "Reducing compositional to disquotational truth", *The Review of Symbolic Logic* 2, 786-798.

[8] Halbach, Volker 2001 "How innocent is deflationism?", *Synthese* 126, 167-194.

[9] Horsten, Leon (1995) "The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist Theory of Truth", in Cortois, P. (ed.) *The Many Problems of Realism*, vol. 3 of *Studies in the General Philosophy of Science*, Tilburg University Press, Tilburg, 173-87.

[10] Horwich, Paul (1990) *Truth*, Basil Blackwell, Oxford.

[11] Kaye, Richard (1991) *Models of Peano arithmetic*, Clarendon Press, Oxford.

[12] Ketland, Jeffrey (2010) "Truth, Conservativeness, and Provability: Reply to Cieslinski", *Mind* 119, 423-436.

[13] Ketland, Jeffrey (1999) "Deflationism and Tarski's Paradise", *Mind* 108, 69-94.

[14] McGee, Vann (2006) "In praise of the free lunch: why disquotationalists should embrace compositional semantics", in Vincent F. Hendricks, Stig Andur Pedersen, and Thomas Bollander (eds.) *Self Reference*, Stanford, CSLI, 95-120.

[15] McGee, Vann (1992) "Maximal consistent sets of instances of Tarski's schema (T)", *Journal of Philosophical Logic* 21, 235-41.

[16] Shapiro, Stewart (1998) "Proof and Truth: Through Thick and Thin", *Journal of Philosophy* 95, 493-521.

[17] Tennant, Neil (2010) "Deflationism and the Gödel-Phenomena: Reply to Cieslinski", *Mind* 119, 437-450.