Minimalism, Reference, and Paradoxes

LAVINIA PICOLLO[1]

Abstract: The aim of this paper is to provide a minimalist axiomatic theory of truth based on the notion of reference. To do this, we first give sound and arithmetically simple notions of reference, self-reference, and well-foundedness for the language of first-order arithmetic extended with a truth predicate; a task that has been so far elusive in the literature. Then, we use the new notions to restrict the T-schema to sentences that exhibit 'safe' reference patterns, confirming the widely accepted but never worked out idea that paradoxes can be characterised in terms of their underlying reference patterns. This results in a strong, ω-consistent, and well-motivated system of disquotational truth, as required by minimalism.

Keywords: minimalism, disquotation, reference, paradoxes, well-foundedness

1 Introduction

The core of minimalism, one of the most popular versions of deflationism about truth nowadays, consist of the following two theses: first, that the meaning of the truth predicate is exhausted by the T-schema, this is

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

where T stands for the truth predicate, φ is a sentence and $\ulcorner\varphi\urcorner$ a quotational name for it.[2] Second, that the truth predicate is just a logico-linguistic device that exists in the language solely to allow us to express certain things—mainly generalisations—we simply cannot express otherwise. The latter prompts the construction of 'logics' or axiomatic theories of truth. The former thesis

[1] I'm obliged to Eduardo Barrio, Volker Halbach, Hannes Leitgeb, Thomas Schindler, the Buenos Aires Logic Group, and the MCMP logic group for their extremely useful comments, suggestions, and corrections on previous stages of this work.

[2] Actually, Horwich (1998), the main exponent of minimalism, takes propositions to be truth bearers rather than sentences. In his account $\ulcorner\varphi\urcorner$ should be understood as a canonical name of the proposition expressed by φ.

suggests the instances of the T-schema—i.e. the T-biconditionals—as axioms.

Unfortunately, as is well-known, if the language is capable of self-reference and the underlying logic is classical, the full T-schema leads to paradox. For we can formulate a liar sentence λ, that "says of itself" that it's *untrue*. Thus, we have that

$$\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner, \tag{1}$$

which obviously contradicts the T-biconditional for λ. As a consequence, minimalists choose to let some T-biconditionals go, as follows:

> [...] the principles governing our selection of excluded instances are, in order of priority: (a) that the minimal theory not engender 'liar-type' contradictions; (b) that the set of excluded instances be as small as possible; and—perhaps just as important as (b)—(c) that there be a constructive specification of the excluded instances that is as simple as possible. (Horwich, 1998, p. 42)

Theories consisting exclusively of instances of the T-schema are called *disquotational*. The search for a constructive and encompassing policy for selecting jointly-consistent instances of this principle is what we call the *minimalist project*.

The task is not as easy as it may seem. The most natural option, namely letting the instances that lead to contradiction go, is not available, as McGee (1992) has shown. There is not one but many different maximal consistent sets of T-biconditionals, all of which are highly complex—not even arithmetically definable. A stricter criterion than mere consistency is needed.

Horwich himself puts forward a plausible restriction:

> The intuitive idea is that an instance of the equivalence [T-] schema will be acceptable, even if it governs a proposition concerning truth (e.g. "What John said is true"), as long as that proposition (or its negation) is grounded—i.e. is entailed either by the non-truth-theoretic facts, or by those facts together with whichever truth-theoretic facts are 'immediately' entailed by them (via the already legitimised instances of the equivalence schema), or ... and so on. (Horwich, 2005, p. 81)

However, he doesn't specify in which way we should understand 'grounded' or 'entailed'. Moreover, the notions of *grounding* (Kripke, 1975) and *dependence on non-truth-theoretic facts* (Leitgeb, 2005) that are available in the literature, even though they can lead to a unique set of acceptable instances of the T-schema, are far from supporting a constructive specification.

Perhaps the criterion that fares best so far is that of T-positiveness: only sentences in which the truth predicate occurs positively (i.e. under the scope of an even number of negation symbols) are allowed in the T-schema (Halbach, 2009). This is a recursive restriction that results in an ω-consistent powerful system when formulated over Peano arithmetic, called PUTB.[3] However, T-positiveness is a highly artificial restriction. It leaves out many intuitively harmless instances of the T-schema, and is inconsistent with appealing truth principles, like consistency and the fact that Modus Ponens and Conditional Proof preserve truth.

According to the orthodox view on paradoxes driven by Poincaré, Russell and Tarski, among others, semantic paradoxes and other pathological expressions are characterised by a common reference pattern, namely, *self-reference*. That certainly seems to be the case for liar sentences. This view has never been thoroughly investigated, mainly because of the elusiveness of a sound notion of reference for formal languages. If true, self-reference could be employed as a plausible restriction on the T-schema. Moreover, since reference has a syntactic vein, the resulting criterion could be in principle simple enough to give axiomatic disquotational theories.

However, Yablo (1985, 1993) challenged the orthodox view with a *prima facie* non-self-referential semantic paradox. This antinomy gave rise to a lively debate on its referential status that put in evidence the lack of sound and precise notions of reference and self-reference in the literature to assess paradoxes in formal languages (cf. Cook, 2006; Leitgeb, 2002). Until we come up with such notions, neither the orthodox view nor the referential status of Yablo's paradox can be evaluated properly.

The first goal of this paper is to remedy this situation. After some technical preliminaries in section 2, section 3 provides precise and intuitively appealing definitions of reference, and thus self-reference and well-foundedness, for formal languages of truth. As it turns out, according to

[3] PUTB can relatively interpret the Ramified Theory of Truth up to the ordinal ϵ_0, $\mathrm{RT}_{<\epsilon_0}$, an axiomatic version of Tarski's hierarchy of semantic theories, and the Kripke-Fererman theory KF, an axiomatisation of Kripke's fixed-point semantic theory with the strong Kleene valuation scheme. In fact, it can be show that all three systems have the same proof-theoretic power. For an introduction to the systems and proofs of the quoted results see (Halbach, 2011), instead.

our definitions, the orthodox view is wrong, for Yablo's paradox isn't self-referential. Nonetheless, we show it is still possible to characterise the semantic paradoxes in terms of their referential patterns: they are all non-well-founded, as Horwich notices. This will become evident in section 4. Since the new notions are of a proof-theoretic nature, we employ them in the construction of an axiomatic theory given by well-founded T-bicondicionals. We show that this system is sound and at least as strong as the best regarded axiomatic theories in the literature. Thus, in section 5 we conclude it's a good candidate for minimalism, the second and main aim of this note.

2 Technical preliminaries

Let $\mathcal{L}$ be the language of first-order Peano arithmetic (PA), with $\neg$, $\rightarrow, \forall$ and $=$ as primitive logical symbols. Formulae containing $\wedge$, $\vee$, $\leftrightarrow$ and $\exists$ are understood as abbreviations. $\mathcal{L}$ contains one individual constant 0, the successor function symbol S, and finitely many other function symbols for primitive recursive (p.r.) functions, to be specified. $\mathcal{L}$ has no predicate symbols besides identity. Other relation symbols such as $<$ are mere abbreviations. For each $n \in \omega$, the complex term given by n occurrences of S followed by 0 is the numeral of n, which we note $\bar{n}$. $\mathbb{N}$ is the standard model of $\mathcal{L}$, with ω as its domain.

$\mathcal{L}_T$, our language of truth, expands $\mathcal{L}$ with a new predicate symbol T for truth. PAT is the result of formulating PA in $\mathcal{L}_T$, taking all the instances of induction given by formulae of this language as axioms. If $\Gamma \subseteq \omega$, let $\langle \mathbb{N}, \Gamma \rangle$ be the expansion of $\mathbb{N}$ to $\mathcal{L}_T$, assigning Γ to T as its extension.

The expressions of $\mathcal{L}_T$ can be codified with natural numbers *à la* Gödel, so that $\mathcal{L}$ and its extensions can be understood as talking about these expressions and sequences (instead of numbers). Given a particular coding and an expression σ of $\mathcal{L}_T$, $\#(\sigma)$ is the code of σ and $\ulcorner \sigma \urcorner$ is the numeral of this code. We assume a standard coding, this is effective and monotonic.[4] Usually, we identify expressions with their codes, for perspicuity.

As is well known, for any $n \in \omega$ the (semi-)recursive subsets of ω^n can be defined in $\mathcal{L}$ and (weakly) represented in PA.[5] Let $ClTerm(v)$ represent the recursive set of closed terms of $\mathcal{L}_T$. If TH $\subseteq \mathcal{L}_T$ is a recursively axiomatisable system, $Bew_{\text{TH}}(v)$ *weakly* represents the set of its theorems. If TH is

[4] I.e. if a string of symbols σ occurs in another string σ', then $\#(\sigma) < \#(\sigma')$.

[5] Actually, this is possible already in Robinson arithmetic, a subsystem of PA. We use the latter for uniformity.

PA, we omit the subscript. We assume that all predicates $Bew_{\text{TH}}(v)$ satisfy Löb's derivability conditions (cf. Löb, 1955).

For any expression σ, let $\vec{\sigma}$ abbreviate $\sigma_1, \ldots, \sigma_n$. The diagonalisation function, that takes a formula $\varphi(v, \vec{v})$ and returns $\forall v(v = \ulcorner\varphi\urcorner \rightarrow \varphi)$, is represented in PA by $Diag(u, v)$. The evaluation function, that takes a term t of $\mathcal{L}_T$ and returns the numeral of the number it denotes, is also recursive and representable in PA by $val(u, v)$.

We assume $\mathcal{L}$ contains the following function symbols for p.r. functions, and PA their corresponding definitions: $\underset{\cdot}{\neg} v$ for the function that maps φ into $\neg\varphi$, $u(v/w)$ for the substitution function, that takes a formula φ and two terms t and s and replaces s in φ with t, and $\dot{v}$ for the numeral function that assigns to each number n its numeral $\bar{n}$. $\mathcal{L}$ cannot contain a function symbol for the evaluation function for its own terms, on pain of triviality. However, we write $u^\circ = v$ for the evaluation function as short for $val(u, v)$.

Let $\forall v(\psi(\ulcorner\varphi(\dot{v})\urcorner))$ abbreviate $\forall v(\psi(\ulcorner\varphi\urcorner(\dot{v}/\ulcorner u\urcorner)))$, which allows us to quantify over the free occurrences of v in $\varphi[v/u]$ when φ is between corner quotes. Also, let $\forall t\varphi$ abbreviate $\forall v(ClTerm(v) \rightarrow \varphi)$. As before, instead of $\forall t(\psi(\ulcorner\varphi\urcorner(t/\ulcorner v\urcorner)))$ we write $\forall t(\psi(\ulcorner\varphi(\underset{\cdot}{t})\urcorner))$ to quantify over terms within Gödel quotes.

Later it will become useful to have in mind the proof of the following well-known result.

Theorem 1 (Weak diagonal lemma) *For any formula $\varphi(v, \vec{v}) \in \mathcal{L}_T$ there is a formula $\psi(\vec{v}) \in \mathcal{L}$ s.t.*

$$\text{PAT} \vdash \psi(\vec{v}) \leftrightarrow \varphi(\ulcorner\psi(\vec{v})\urcorner, \vec{v})$$

Proof. The result of applying the diagonalisation function to

$$\forall u(Diag(v, u) \rightarrow \varphi(u, \vec{v}))$$

is the formula

$$\forall v(v = \ulcorner\forall u(Diag(v, u) \rightarrow \varphi(u, \vec{v}))\urcorner \rightarrow \forall u(Diag(v, u) \rightarrow \varphi(u, \vec{v}))) \quad (2)$$

Let a be the numeral of the Gödel code of (2). (2) is equivalent in PAT to

$$\forall u(Diag(\ulcorner\forall u(Diag(v, u) \rightarrow \varphi(u, \vec{v}))\urcorner, u) \rightarrow \varphi(u, \vec{v}))$$

which is equivalent to $\varphi(a, \vec{v})$. $\square$

It's possible to strengthen this result using function symbols as follows:

Theorem 2 (Strong diagonal lemma) *For any formula $\varphi(v, \vec{v})$ of $\mathcal{L}_T$ there is a term t s.t.*

$$\text{PA} \vdash t = \ulcorner \varphi(t, \vec{v}) \urcorner$$

It is commonly thought that both diagonal lemmata deliver self-referential expressions. For instance, applying strong diagonalisation to the predicate $\neg Bew(v)$ we obtain a term g s.t.

$$\text{PA} \vdash g = \ulcorner \neg Bew(g) \urcorner \tag{3}$$

$\neg Bew(g)$ is a Gödel sentence of PA and it is usually understood as "saying of itself" that it isn't provable in PA. As is well known, this sentence is true and therefore unprovable in PA.

Finally, recall that formulae in $\mathcal{L}$ can be classified according to their quantificational—also called *arithmetical*—complexity into sets Σ_n, Π_n and $\Delta_n \subseteq \mathcal{L}$, with $n \in \omega$. These sets constitute the *arithmetical hierarchy*. If φ is logically equivalent to a formula where all quantifiers are bound, φ is both Σ_0 and Π_0. If φ is logically equivalent to a formula of the form $\forall \vec{v} \psi$, where $\psi \in \Sigma_n$, then $\varphi \in \Pi_{n+1}$. If φ is logically equivalent to a formula of the form $\neg \forall \vec{v} \psi$ where $\psi \in \Pi_n$, then $\varphi \in \Sigma_{n+1}$. Finally, if φ is both Π_n and Σ_n, we say that $\varphi \in \Delta_n$. Note that the sets in the hierarchy are cumulative, for it's always possible to add superfluous quantifiers at the beginning of a formula.

Recursive sets can be defined in $\mathcal{L}$ by Δ_0-formulae, and semi-recursive sets by Σ_1-formulae. Non-semi-recursive sets can only be defined by more complex formulae, if at all. Every Δ_0-formula is decidable in PA. If $\varphi \in \Sigma_1$ is true in the standard model, then $\text{PA} \vdash \varphi$, this is, PA is Σ_1-complete. For other, more complex expressions, we have no guarantees.

3 Alethic reference

In this section we focus on the reference of sentences of $\mathcal{L}_T$ to sentences of the same language. This isn't just any kind of reference but reference *through the truth predicate* or, as we call it, *alethic reference*. Intuitively, an expression alethically refers to all sentences that syntactically fall, as it were, under the scope of the truth predicate. This will become clear soon. The notion we provide, is, as we show, of a low arithmetical complexity, though this doesn't come without costs.

A sentence in a first-order language can refer to an object either by mentioning it or by quantifying over it. In the first case, the expression must contain a term t that denotes the object. Since we're only interested in alethic reference, we have the following definition.

Definition 1 *Let φ and ψ be sentences of $\mathcal{L}_T$. φ refers by mention to ψ, or* m-refers*, for short, iff φ contains a subsentence Tt and* $\textsc{pa} \vdash t = \ulcorner\psi\urcorner$.

Note that if t actually denotes the code of ψ then PA will be able to prove it, for identity statements don't contain quantifiers. Definition 1 covers many cases, like the liar sentence that obtains applying the strong diagonal lemma to $\neg Tv$, that is

$$\textsc{pa} \vdash l = \ulcorner\neg Tl\urcorner, \tag{4}$$

that intuitively m-refers to itself. In general, any sentence that result from strongly diagonalising formulae that contain Tv as a subformula will m-refer to themselves. On the other hand, if we strongly diagonalise formulae that don't satisfy this condition, we might not get self-referential expressions. For instance, diagonalising $T\underset{\cdot}{\neg}v$ we get

$$\textsc{pa} \vdash l' = \ulcorner T\underset{\cdot}{\neg}l'\urcorner. \tag{5}$$

$T\underset{\cdot}{\neg}l'$ is an alternative liar sentence that doesn't refer to itself according to definition 1 but only to its negation. The latter is actually the self-m-referential one. This follows from (5) and the fact that $\neg T\underset{\cdot}{\neg}l'$ contains $T\underset{\cdot}{\neg}l'$ as a subsentence.

Sentences of $\mathcal{L}_T$ can also refer to other sentences by quantifying over them. For instance,

$$\forall x(Bew(x) \rightarrow Tx) \tag{6}$$

intuitively refers to all theorems of arithmetic, while

$$\forall x Tx \tag{7}$$

seems to refer to everything. Conditionals allow us to restrict reference by quantification. Thus, if a universal quantifier or a string of universal quantifiers is followed by a conditional expression, we would like to say that it refers to whatever satisfies the antecedent, and otherwise it refers to everything.

However, things are not so simple. In the first place, talking about satisfaction introduces too much complexity into our notion, for to know whether an arbitrary code satisfies a certain formula we would have to look into the set of arithmetically true statements, which is not arithmetically definable. Thus, we turn to the notion of *provability* instead. After all, what matters to avoid paradoxes is that we cannot *derive* a contradiction or an unsound claim. Consequently, the resulting notion of reference via quantification—or *q-reference*, for short—will be tied to a particular system, the system whose provability predicate we employ in the definition. We work in PA, but any extension of Robinson arithmetic works as well.

Secondly, recall we're only interested in alethic reference here, so what matters is what actually falls under the scope of T. While in (6) all theorems of arithmetic fall under the scope of T, in $\forall x(Bew(x) \rightarrow T\dot{\neg}x)$ only their negations do. Analogously, in (7) all sentences fall under T but in $\forall xT\dot{\neg}x$ only negations do. And the same can be said of more complex expressions. For instance, in $\forall x(Bew(x) \rightarrow \forall y(y = \dot{\neg}x \rightarrow \neg Ty))$, again, only negations of PA's theorems fall under the scope of the truth predicate. Thus, we define q-reference recursively. Roughly, a universal expression q-refers to whatever its instances m- or q-refer to, unless the universal quantifier is followed by a conditional, in which case we consider only the instances given by numerals that provably satisfy the antecedent.

Finally, note that if quantification is restricted by a conditional expression in which the truth predicate occurs both in the antecedent and the consequent—e.g. $\forall x(Tx \rightarrow Tx)$, our theory has no means to know which sentences fall in the scope of T; since the idea is to axiomatise truth in terms of reference, not vice versa. Sentences of this kind could exhibit dangerous reference patterns without us knowing. Therefore, we just treat them as non-conditional expressions.

Now we turn to the formal definition of alethic q-reference.

Definition 2 *Let φ, ψ be sentences of $\mathcal{L}_T$. φ q-refers to ψ in* PA *iff T occurs in φ and one of the conditions 1-3 holds:*

1. *$\varphi := \forall\vec{v}\chi$ and*
   - (a) *$\chi := Tt$ or $\chi := \neg\delta$ and, for some $\vec{k} \in \omega$, $\chi[\vec{\overline{k}}/\vec{v}]$ q-refers to ψ or has a new occurrence of Ts as a subsentence s.t.* PA $\vdash s = \ulcorner\psi\urcorner$*; or*
   - (b) *$\chi := \delta \rightarrow \gamma$ and*
     - i. *both δ and γ contain T and for some $\vec{k} \in \omega$, $\chi[\vec{\overline{k}}/\vec{v}]$ q-refers to ψ or contains a new occurrence of Tt as a subsentence s.t.* PA $\vdash t = \ulcorner\psi\urcorner$*, or*

ii. only γ (δ) contains T and there exist $\vec{k} \in \omega$ and $1 \leq i \leq n$ s.t. PA $\vdash \delta[\vec{\vec{k}}/\vec{v}]$ *($\neg\gamma[\vec{\vec{k}}/\vec{v}]$) and ($\delta \to \gamma)[\vec{\vec{k}}/\vec{v}]$ q-refers to ψ or contains a new occurrence of Tt as a subsentence s.t.* PA $\vdash t = \ulcorner\psi\urcorner$.

2. $\varphi := \neg\chi$ and χ q-refers to ψ.

3. $\varphi := \chi \to \delta$ and either χ or δ q-refer to ψ.

By *a new occurrence of Tt in $\chi[\vec{\vec{k}}/\vec{v}]$* in the above definition we mean that Tt occurs in the result of replacing all occurrences of Tt in χ with $0 = 0$ (or any sentence not containing T) and then instantiating the variables $\vec{v}$ with $\vec{\vec{k}}$. This is needed to avoid cases of m-reference passing as cases of q-reference—e.g. in $\forall x T\ulcorner\lambda\urcorner$.

According to definition 2, the liar sentence λ introduced in (1) q-refers to itself, as well as all sentences that are obtained by weakly diagonalising a predicate $\varphi(v)$ containing Tv as a subformula. Looking at the proof of theorem 1, we see that the real form of these sentences is

$$\forall u(u = \ulcorner\forall v(Diag(u, v) \to \varphi(v))\urcorner \to \forall v(Diag(u, v) \to \varphi(v))) \qquad (8)$$

Applying the clause (b)ii. of definition 2 twice, we get that (8) is q-self-referential. But just like in the case of m-reference, if Tv isn't a subformula of $\varphi(v)$, our definition cannot guarantee that the weak diagonalisation of this predicate will be a self-referential expression.

Note that the notion of q-reference could clash with some of our intuitions. If $g = \ulcorner\neg Bew(g)\urcorner$ as in (3), strongly diagonalising the predicate $\forall x(x = y \wedge \neg Bew(g) \to \neg Tx)$ delivers a term l^* s.t.

$$\text{PA} \vdash l^* = \ulcorner\forall x(x = l^* \wedge \neg Bew(g) \to \neg Tx)\urcorner \qquad (9)$$

Since $\neg Bew(g)$ is true in the standard model, intuitively we would say $\forall x(x = l^* \wedge \neg Bew(g) \to \neg Tx)$ q-refers to itself. However, we're thinking about reference *in* PA, so this won't be the case. For PA cannot prove its own Gödel sentence, on pain of triviality. This is a direct consequence of adopting provability instead of satisfaction for defining reference. As we will see later, this issue can be circumvented to some extent.

Putting the notions of m- and q-reference together isn't enough to define reference *simpliciter*. Consider the following identities:

$$\begin{aligned} l_1 &= \ulcorner Tl_2\urcorner \\ l_2 &= \ulcorner\neg Tl_1\urcorner. \end{aligned} \qquad (10)$$

This statements can be proved in PA by slightly tweaking theorem 2. Together, they give rise to a paradox akin to the liar. Sentences Tl_2 and $\neg Tl_1$ m-refer only to each other but, intuitively, also refer to themselves, though *indirectly*. Alethic reference is a transitive relation.

Definition 3 *Let φ, ψ be sentences of $\mathcal{L}_T$. φ directly refers to ψ in* PA *iff it m- or q-refers to ψ in* PA.

Definition 4 *A sequence of sentences $\chi_0, \dots, \chi_n \in \mathcal{L}_T$, $n \in \omega$, is a chain of reference in* PA *iff, for each $i < n$, χ_i directly refers to χ_{i+1} in* PA.

Definition 5 *Let φ, ψ be sentences of $\mathcal{L}_T$. φ refers to ψ in* PA *iff there's a chain of reference in* PA *starting with φ and ending with ψ.*

According to this definition, both Tl_2 and $\neg Tl_1$ refer to themselves, as we wanted.

It's worth noticing that the notion of reference we present is not extensional but *hyperintensional*: there are logically equivalent sentences that don't refer to the same things. For instance, $0 = 0$ and $T\ulcorner\lambda\urcorner \vee \neg T\ulcorner\lambda\urcorner$ are logically equivalent but, while the former doesn't refer to anything, the latter refers to λ. Unlike grounding or dependence, reference is based at least partly on syntactic features of sentences and, therefore, extensionality fails.

The notion of reference we introduced can be used to define relevant reference patterns, such as the following two.

Definition 6 *A sentence $\varphi \in \mathcal{L}_T$ is self-referential in* PA *iff it refers to itself in* PA.

According to this definition, sentences such as λ in (1), $\neg Tl$ in (4) and Tl_2 and $\neg Tl_1$ in (10) turn out to be self-referential.

Definition 7 *A sentence $\varphi \in \mathcal{L}_T$ is well-founded in* PA *iff there is no indefinitely extensible chain of reference in* PA *starting with φ.*

Every self-referential expression is obviously non-well-founded. But there are also non-well-founded sentences that don't refer to themselves. Yablo's paradox (Yablo, 1985, 1993) consist of an infinite sequence of sentences, each of which says of the ones coming after that they are untrue. In $\mathcal{L}_T$, Yablo's sentences can be formalised as $\forall x > \bar{n} \neg T\upsilon(x)$, where $\upsilon(v) = \ulcorner \forall x > \dot{v} \neg T\upsilon(x) \urcorner$. This identity statement is provable in PA by strong diagonalisation, guaranteeing the existence of the list in our formal setting.

According to definitions 6 and 7, no sentence in the sequence is self-referential, though they are all non-well-founded. It can be shown that an ω-inconsistency follows from the set of T-biconditionals for sentence in Yablo's list, so the paradox is actually an ω-paradox (cf. Ketland, 2005). If our definitions are correct, this shows that the orthodox view on semantic paradoxes is mistaken: there are non-self-referential (ω-)paradoxes. But this doesn't spell doom to our approach, for semantic paradoxes could share a reference pattern other than self-reference; for instance, non-well-foundedness. Later we will see this is actually the case.

It's easily seen that m-reference is recursive. Since the only proper non-recursive notion involved in the definition of q-reference is the semi-recursive notion of provability, and it occurs only positively, q-reference is also semi-recursive. By a similar reasoning, direct reference, reference and self-reference are semi-recursive as well. Well-foundedness, on the other hand, is more complex. Nonetheless, all of these notions can be defined in $\mathcal{L}$ and most of them at least weakly represented in PA. This sets reference further apart from the usual notions of grounding and dependence, and is enough to allow our notion to play a role in a disquotational axiomatisation of truth.

Being q-reference strictly semi-recursive, PA can prove all positive cases, but some negative ones won't be provable. For instance, PA has no means to know that

$$\forall x(x = \ulcorner 0 = 0 \urcorner \to Tx) \tag{11}$$

does not q-refer to itself. That would mean PA knows that $\neg Bew(\ulcorner \forall x(x = \ulcorner 0 = 0 \urcorner \to Tx) \urcorner = \ulcorner 0 = 0 \urcorner)$, this is, its own consistency. Since we want to be able to determine which sentences exhibit safe referential patterns to take them as instances of the T-schema, and (11) clearly does, we must add axioms to inform our theory of *some* negative cases of q-reference—by Gödel's theorem, it's impossible to have them all. The simplest principle we can add is

$$\forall x(Bew(\underset{\cdot}{\neg} x) \to \neg Bew(x)) \tag{QR}$$

Since QR is true-in-$\mathbb{N}$, PA $+$ QR, or QR(PA) for short, is ω-consistent. Given that PA knows that $\ulcorner \forall x(x = \ulcorner 0 = 0 \urcorner \to Tx) \urcorner \neq \ulcorner 0 = 0 \urcorner$ and, therefore, that $Bew(\ulcorner \forall x(x = \ulcorner 0 = 0 \urcorner \to Tx) \urcorner \neq \ulcorner 0 = 0 \urcorner)$, we can conclude in QR(PA) that $\neg Bew(\ulcorner \forall x(x = \ulcorner 0 = 0 \urcorner \to Tx) \urcorner = \ulcorner 0 = 0 \urcorner)$, which means that (11) doesn't q-refer to itself.

4 Well-founded truth

In the previous section we provided formal proof-theoretic notions of alethic reference, self-reference, and well-foundedness for sentences of $\mathcal{L}_T$ in PA. The next step is to use them in the formulation of axiomatic disquotational theories of truth.

In the spirit of Horwich's (2005, p. 81) idea cited in the introduction, the most natural choice is to relativise the T-schema to the predicate $Wf(v) \in \mathcal{L}$ that defines well-foundedness in PA according to definition 7. However, this wouldn't result in a consistent system. Coming back to our example in (9), recall that $\forall x(x = l^* \wedge \neg Bew(g) \rightarrow \neg Tx)$ $(= l^*)$ doesn't refer to anything in PA, for PA $\nvdash Bew(\ulcorner \neg Bew(g) \urcorner)$. Moreover, QR$(PA)$ can prove this, by internalising a proof of Gödel's theorem. Thus, QR$($PA$) \vdash Wf(l^*)$. But, as it turns out, the T-biconditional for $\forall x(x = l^* \wedge \neg Bew(g) \rightarrow \neg Tx)$ leads directly to paradox. The reason is that this sentence is well-founded in PA but *not in* QR$($PA$)$, where it's actually self-referential.

To avoid this problem we restrict our attention to those sentences whose referenced expressions do not increase when we adopt more powerful systems. We call them *r-stable*. To formally characterise them, we need the following auxiliary notion:

Definition 8 *A sentence $\varphi \in \mathcal{L}_T$ is dr-stable iff all its subformulae of the form $\psi \rightarrow \chi$ where a free variable occurs in the scope of T and exactly one of ψ, χ contains T are s.t. the one not containing T is Δ_0.*[6]

For instance, $T\ulcorner \forall x(Bew(x) \rightarrow Tx) \urcorner$ and (11) are dr-stable, while

$$\forall x(Bew(x) \rightarrow Tx)$$

isn't, for $Bew(v) \notin \Delta_0$. If a dr-stable sentence φ doesn't directly refer to another sentence ψ in PA, φ cannot directly refer to ψ in a stronger theory either, since PA already decides all instances of Δ_0-formulae.

Definition 9 *A sentence $\varphi \in \mathcal{L}_T$ is r-stable iff it is dr-stable and refers only to dr-stable sentences.*

Thus, $T\ulcorner \forall x(Bew(x) \rightarrow Tx) \urcorner$ isn't r-stable, but (11) is, because it only refers to $0 = 0$. R-*un*stable expressions bear a certain analogy with blind

[6] By just considering Δ_0-expressions and not also their PA-equivalents we're leaving behind many sentences which have a stable direct reference. However, this doesn't matter for our purposes, since in the axioms of our truth system the restriction on the T-schema will be closed under PAT-equivalence.

truth ascriptions: in both cases we don't know what we are asserting and, *a fortiori*, if it's a paradox or not. Only for r-stable sentences we can be sure that their reference patterns are safe.

Since the set of Δ_0-expressions is obviously semi-recursive, so is the set of dr-stable sentences. Given that reference is also semi-recursive, r-stability has Π_2-complexity. Let $RSt(v) \in \Pi_2$ define this set. The theory we introduce next restricts the T-schema to r-stable and well-founded sentences and their equivalents *in a uniform way*.

Definition 10 WFUTB $\subseteq \mathcal{L}_T$ *extends* QR(PA) *with the new instances of induction for* $\mathcal{L}_T$*-formulae and the following schema, where* $\varphi \in \mathcal{L}_T$ *contains exactly* n *free variables:*

$$\begin{aligned}&\forall \vec{t} \forall x (RSt(x(\vec{t})) \wedge Wf(x(\vec{t})) \wedge \\ &\qquad \wedge Bew_{\text{PAT}}(\ulcorner \varphi(\dot{\vec{t}}) \urcorner \dot{\leftrightarrow} x(\vec{t})) \rightarrow (T \ulcorner \varphi(\dot{\vec{t}}) \urcorner \leftrightarrow \varphi(\vec{t}^{\circ})))\end{aligned}$$

WFUTB—for *Well-founded Uniform Tarski Biconditionals*—allows instances of the T-schema given, uniformly, by all sentences that are equivalent in PAT to an r-stable well-founded sentence. This includes of course, all r-stable well-founded expressions, but also, for example, $\forall x((Tl \rightarrow Tl) \wedge x = \ulcorner 0 = 0 \urcorner \rightarrow Tx)$ and $\neg \forall x(Tx \rightarrow Tx)$, which are not well-founded in PA. On the other hand, it excludes many intuitively safe instances, such as the one given by $\forall x(Bew(x) \rightarrow Tx)$. We get the following results:

Proposition 1 WFUTB *is* ω*-consistent.*

Proof. We just give a sketch. It can be shown that if a dr-stable sentence $\varphi \in \mathcal{L}_T$ doesn't refer directly to another sentence ψ, then there's a set $\Gamma \subseteq \mathcal{L}_T$ on which φ depends s.t. $\psi \notin \Gamma$, by induction on the logical complexity of φ.[7] It follows as a corollary that all r-stable well-founded sentences belong to Leitgeb's set Φ_{lf} of expressions that depend on non-semantic states of affairs (cf. Leitgeb, 2005, § 3), by transfinite induction on the ordinal level of the fixed-point construction that leads to Φ_{lf}. Since there's a model $\langle \mathbb{N}, \Gamma \rangle$ of $\mathcal{L}_T$ that verifies all instances of the T-schema given by sentences in Φ_{lf} (Leitgeb, 2005, theorem 17), $\langle \mathbb{N}, \Gamma \rangle \vDash$ WFUTB as well. □

Proposition 2 *The theory of Ramified Truth up to* ϵ_0 $\text{RT}_{<\epsilon_0}$ *is relatively interpretable in* WFUTB.

[7] For a definition of *dependence* and its basic properties, see (Leitgeb, 2005).

Proof. We just give an idea of the proof.[8] We show that for each $\alpha < \epsilon_0$ there's a predicate $\theta_{\bar{\alpha}}(v) \in \mathcal{L}_T$ that satisfies in WFUTB the axioms that hold for $T_\alpha(v)$ in $\mathrm{RT}_{<\epsilon_0}$.[9] First, we obtain a binary predicate $\theta_y(x) \in \mathcal{L}_T$ by strongly diagonalising over the variable w a complex predicate that is basically the disjunction of the axioms of $\mathrm{RT}_{<\epsilon_0}$, where the predicates $T_\alpha(v)$ have been replaced by $Tw(\dot{y}/\ulcorner y \urcorner)(\dot{u}/\ulcorner x \urcorner)$ (and, correspondingly, α with y and v with u). Then we show by internal transfinite induction on α that the uniform T-schema holds in WFUTB for all predicates $\theta_{\bar{\alpha}}(v)$, where $\alpha < \epsilon_0$, which gives us the axioms of $\mathrm{RT}_{<\epsilon_0}$. This is done by uniformly showing in WFUTB that all instances of the predicates $\theta_{\bar{\alpha}}(v)$ given by sentences in which only predicates $\theta_{\bar{\beta}}(v)$ with $\beta < \alpha$ occur are r-stable and well-founded. □

As a corollary of propositions 1 and 2, WFUTB is a sound and powerful system. Since the Kripke-Feferman theory KF and PUTB have the same proof-theoretic strength as $\mathrm{RT}_{<\epsilon_0}$, WFUTB is at least as strong as these three well-regarded systems.

5 Conclusions

In this paper we have provided sound, precise, and arithmetically simple notions of reference, self-reference, and well-foundedness. Moreover, these concepts have been proved useful in the assessment of semantic paradoxes and in the formulation of axiomatic theories of truth.

We have also shown that a natural theory of disquotational truth that is ω-consistent, as powerful as KF and PUTB, and imposes only arithmetical restrictions on the T-schema is possible. Our system WFUTB is therefore (a) sound, (b) encompassing, and (c) employs a simple selective criterion of T-biconditionals. As a consequence, it's a perfect candidate for the minimalist search.

Perhaps other—more powerful—systems can be devised using the notions we introduced in section 3. It could well be that paradoxes shared more specific reference patterns than non-well-foundedness, which could be turned into broader selective criteria for instances of disquotation. We

[8]The proof is similar to the demonstration of Halbach's (2011, theorem 15.25).

[9]As is well known, natural numbers can codify ordinals up to ϵ_0 (and beyond). If $\alpha < \epsilon_0$, $\bar{\alpha}$ is the numeral of its code. PA is able to prove all instances of transfinite induction up to ϵ_0. For the details see (Pohlers, 2009, chapter 3).

believe this note not only provides answers to several issues such as finding a natural minimalist theory or assessing the orthodox view on semantic paradoxes, but also opens a new line of research on these topics.

References

Beall, J. C. (2005). Transparent Disquotationalism. In B. Armour-Garb & J. C. Beall (Eds.), *Deflationism and Paradox* (pp. 7–22). Oxford University Press.

Cook, R. T. (2006). There Are Non-circular Paradoxes (but Yablo's Isn't One of Them!). *The Monist*, *89*, 118–149.

Halbach, V. (2009). Reducing Compositional to Disquotational Truth. *Review of Symbolic Logic*, *2*, 786–798.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

Horwich, P. (1998). *Truth* (second ed.). Oxford: Blackwell.

Horwich, P. (2005). A Minimalist Critique of Tarski on Truth. In B. Armour-Garb & J. C. Beall (Eds.), *Deflationism and Paradox* (pp. 75–84). Oxford University Press.

Ketland, J. (2005). Yablo's Paradox and ω-inconsistency. *Synthese*, *145*, 295–307.

Kripke, S. (1975). Outline of a Theory of Truth. *Journal of Philosphy*, *72*, 690–716.

Leitgeb, H. (2002). What Is a Self-referential Sentence? Critical Remarks on the Alleged (Non)-circularity of Yablo's Paradox. *Logique et Analyse*, *177-178*, 3–14.

Leitgeb, H. (2005). What Truth Depends On. *Journal of Philosphical Logic*, *34*, 155–192.

Löb, M. H. (1955). Solution of a Problem of Leon Henkin. *Journal of Symbolic Logic*, *20*, 115–118.

McGee, V. (1992). Maximal Consistent Sets of Instances of Tarski's Schema. *Journal of Philosphical Logic*, *21*, 235–241.

Pohlers, W. (2009). *Proof Theory: the First Step into Impredicativity*. Berlin-Heidelberg: Springer.

Yablo, S. (1985). Truth and Reflexion. *Journal of Philosphical Logic*, *14*, 297–349.

Yablo, S. (1993). Paradox without Self-reference. *Analysis*, *53*, 251–252.

Lavinia Picollo

Lavinia Picollo

Ludwig-Maximilians University Munich
Germany
E-mail: Lavinia.Picollo@lrz.uni-muenchen.de