Journal of Philosophical Logic (2020) 49:439–474
https://doi.org/10.1007/s10992-019-09525-9

Reference and Truth

Lavinia Picollo[1]

Received: 19 July 2019 / Accepted: 30 July 2019 / Published online: 29 August 2019

Abstract
I apply the notions of alethic reference introduced in previous work in the construction of several classical semantic truth theories. Furthermore, I provide proof-theoretic versions of those notions and use them to formulate axiomatic disquotational truth systems over classical logic. Some of these systems are shown to be sound, proof-theoretically strong, and compare well to the most renowned systems in the literature.

Keywords Semantic paradoxes · Disquotation · Reference · Self-reference · Well-foundedness · Formal truth theories

1 Introduction

Tarski [32, 33] has bequeathed to us a pessimistic but valuable lesson about truth. Given a language L with a monadic predicate T and a name $\ulcorner\varphi\urcorner$ for each sentence φ of the language, for T to adequately express the notion of truth of sentences of L, each sentence φ should be materially equivalent to $T\ulcorner\varphi\urcorner$. However, if the underlying logic is classical and the language is 'expressive enough', the equivalence must fail in some cases, on pain of triviality. For instance, if L allows for self-referential expressions and, in particular, contains a liar sentence λ, that says of itself that it is not true, λ and $\neg T\ulcorner\lambda\urcorner$ will turn out to be equivalent. Thus, the equivalence between λ and $T\ulcorner\lambda\urcorner$ is untenable. As a consequence, an adequate truth predicate for the whole language is not possible in L, if working within classical logic.

Tarski proposed to restrict the equivalence between each φ and $T\ulcorner\varphi\urcorner$ (expressed as a biconditional) – the so-called T-schema – to the T-free sentences of the language. In this way, λ is easily excluded. Tarskian *typed* truth theories have been pervasive for many years and play a prominent role in diverse areas of logic and philosophy.

✉ Lavinia Picollo
l.picollo@ucl.ac.uk

[1] Department of Philosophy, University College London, 19 Gordon Square, London WC1H 0AW, UK

Nonetheless, the shared view nowadays is that these theories are far too restrictive. After all, many sentences containing the truth predicate – e.g. expressions of the form $T\ulcorner T\ulcorner\varphi\urcorner\urcorner$ where φ is T-free – seem to be entirely unproblematic.

Less drastic solutions have been subsequently explored. Theories which weaken the logic but keep some form of equivalence between each sentence and its truth ascription – originating in the work of Kripke [20] and Martin & Woodruff [23] – are very popular. Others have investigated the possibility of remaining classical but imposing less strict restrictions on the admissible instances of the T-schema. The goal is to find a unified criterion that allows us to keep as many unproblematic instances as possible, leaving the paradoxical ones out.

A promising strategy in this direction is to restrict the T-schema to sentences that exhibit 'safe' reference patterns. The orthodox view, championed by Russell, Poincaré, and Tarski, has it that the cause of semantic paradox is self-reference. Despite the popularity of this view, the restriction of the T-schema to non-self-referential sentences has not yet been explored. The main reason is that the notions of self-reference and reference have proved to be quite elusive in the past, and are now surrounded by an aura of scepticism.[1] Additionally, the self-reference diagnosis has been (relatively) recently challenged by a new paradox that is *prima facie* free of self-reference: the Visser-Yablo paradox.[2] However, the bearing of this new antinomy is not yet entirely clear: the lack of a precise notion of self-reference has hindered the evaluation of the Visser-Yablo paradox and thus the self-reference diagnosis of paradoxicality.

In [27] I provide a systematic and rigorous account of reference *in the context of truth* – or "alethic reference", as I call it – and define self-reference and other reference patterns in terms of this notion. I hope the intuitive appeal of the definitions I put forward there helps dissipate the scepticism the corresponding notions are immersed in, at least with regard to the semantic paradoxes. Furthermore, I show that the expressions involved in the Yablo-Visser paradox are not self-referential according to the new notions, refuting the self-reference diagnosis on the roots of paradox. Although this ruins the prospects of restricting the T-schema to non-self-referential expressions, the more general project of restricting it to sentences that exhibit safe reference patterns need not be abandoned, as all paradoxical expressions might share other reference patterns.

The purpose of this paper is to confirm this hypothesis. I provide both semantic and axiomatic theories of truth in which the T-schema is restricted to *well-founded* expressions, and prove they are encompassing and sound. Moreover, I show this condition can be relaxed even further, i.e. that the reference patterns that underlie paradoxical expressions can be given a more fine-grained characterization. In short, I deploy the reference notions introduced in [27] to formulate different restricting criteria for admissible instances of T-schema in terms of their underlying reference patterns and show that this strategy is successful, as it results in classically consistent, sound, and fairly attractive theories of truth.

[1] See Picollo [26] for an overview of the state of the matter.

[2] See Herzberger [15], Visser [34], and Yablo [35, 36].

Section 2 provides a technical introduction to formal theories of truth, followed by an overview of the state of the issue regarding different restrictive criteria for the truth predicate that have been proposed in the literature. I also give a compact presentation of the approach to reference in the context of truth introduced in [27]. Section 3 employs these notions to construct semantic truth theories. Section 4 provides simpler, proof-theoretic versions of the notions of reference (and related concepts) from Section 2 and formulates several axiomatic truth systems based on them, some of which are shown to be sound and proof-theoretically strong. Finally, in Section 5 I conclude by evaluating these systems in light of Leitgeb's [22] criteria for formal theories of truth.

2 Preliminaries

2.1 Formal Truth Theories

Let $\mathscr{L}$ be the language of first-order Peano arithmetic (PA) and let $\mathscr{L}_{\mathrm{T}}$ augment $\mathscr{L}$ with a monadic predicate T, to express truth. $\mathscr{L}$ contains $=$, $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\forall$, and $\exists$ as logical constants, an individual constant 0, a monadic function symbol S, two dyadic function symbols $+$ and $\times$, and a finite stock of extra function symbols for primitive recursive (p.r.) functions to be specified. All other logical and non-logical symbols occurring in formulae are to be understood as the usual abbreviations. Let $\mathbb{N}$ be the standard model for $\mathscr{L}$, with ω as its domain. Note that, for each $n \in \omega$, $\mathscr{L}$ has a term $\overline{n}$ denoting n (the numeral of n), that consists of n occurrences of S followed by 0.

We work with a fixed (effective and monotonic) coding of expressions of $\mathscr{L}_{\mathrm{T}}$ by numbers in ω.[3] If σ is a string of symbols of $\mathscr{L}_{\mathrm{T}}$, we write $\ulcorner\sigma\urcorner$ for the numeral of its code. We often identify expressions of $\mathscr{L}_{\mathrm{T}}$ with their codes if there's no room for confusion. Unless otherwise indicated, by "formula" and "sentence" we mean formula of $\mathscr{L}_{\mathrm{T}}$ and sentence of $\mathscr{L}_{\mathrm{T}}$, respectively.

Although $\mathscr{L}$ speaks primarily about natural numbers, our coding allows it to express many syntactic properties, relations, and functions about the expressions of $\mathscr{L}_{\mathrm{T}}$. Thus, our truth theories can be formulated in $\mathscr{L}_{\mathrm{T}}$, with background syntactic principles formulated in $\mathscr{L}$.

Formal truth theories can be either semantic or axiomatic. Semantic theories consists of a model or family of models $\langle \mathbb{N}, \Gamma \rangle$ expanding $\mathbb{N}$ to a model of $\mathscr{L}_{\mathrm{T}}$, where Γ is the extension of T in the model. In evaluating them, we look at the truth principles that hold in every model of the family. Epistemic considerations are also relevant, however: it is important that there is a way to know which truth principles belong to the theory, to the extent possible. If the theory is too complex, this would hardly be the case. This is not to say that complex semantic constructions are of no value. On the contrary, they can serve as witnesses of the consistency of collections of truth principles, or play a heuristic role in the formulation of more constructive theories, that is, axiomatic ones.[4]

[3] See my companion paper for more details about the coding.

[4] See Field [5] and Halbach & Horsten [11].

By contrast, axiomatic truth theories result from adding truth-theoretic axioms to a syntax theory, which we assume to be PA.[5] We assume PA contains the defining recursion equations for each extra function symbol in $\mathscr{L}$. As is well known, PA is strong enough to represent every recursive relation between numbers and, therefore, expressions of $\mathscr{L}_{\mathrm{T}}$, and to weakly represent every recursively enumerable relation. Let PAT consist of the axioms of PA formulated in $\mathscr{L}_{\mathrm{T}}$ with induction for the whole language. Call an axiomatic truth theory in $\mathscr{L}_{\mathrm{T}}$ any recursive extension of PAT. Of course, some theories will be highly incomplete and others simply unsound, but the terminology is convenient. As before, the merits of a theory lie in the truth principles the theory entails. To know what principles hold in an axiomatic theory it is enough to find a proof.[6]

2.2 Tarski's Theorem (and ways to circumvent it)

Ideally, any truth theory (whether semantic or axiomatic) would satisfy Tarski's condition of material adequacy, according to which all instances of the following schema hold in the theory:

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

These are known as "T-biconditionals". Alternatively, we could work with a more general variant of the T-biconditionals, the *uniform* T-biconditionals:

(Uniform T-schema) $\qquad \forall t_1 \ldots \forall t_n (\mathrm{T}\ulcorner\varphi(\dot{t}_1, \ldots, \dot{t}_n)\urcorner \leftrightarrow \varphi(t_1^\circ, \ldots, t_n^\circ))$

Both the T-schema and its uniform version are known as "disquotational" principles. Note that the set of closed terms of $\mathscr{L}_{\mathrm{T}}$ is p.r., as well as the value relation that holds between each term of the language and the number it denotes (as $\mathscr{L}$ contains only finitely many function symbols). Let $\mathrm{CTerm}(x)$ and $x^\circ = y$ represent in PA ("represent", from now on) this set and relation, respectively.[7] Moreover, the substitution function that takes a formula φ, a term t, and a variable v and returns the formula that results from replacing all free occurrences of v in φ with t, is also p.r. and, thus represented by a term $x(y/z) \in \mathscr{L}$. We write $\forall t\, \varphi$ for $\forall v\, (\mathrm{CTerm}(v) \rightarrow \varphi)$ and $\exists t\, \varphi$ for $\exists v\, (\mathrm{CTerm}(v) \wedge \varphi)$, for a suitable variable v. Finally, $\ulcorner\varphi(\dot{t})\urcorner$ abbreviates $\ulcorner\varphi\urcorner(t/\ulcorner v\urcorner)$, for some suitable term variable t, provided that v is the only free variable in φ. Thus, the instances of the Uniform T-schema quantify over closed terms, entailing all substitution instances of the standard T-schema *uniformly*.

Unfortunately, neither of these principles can be implemented unrestrictedly, as the language is 'expressive enough' to allow for paradoxical expressions such as liar sentences.

Let $\mathbf{v}$ abbreviate the string of variables $v_1, \ldots, v_n$ different from x and y.

[5] Robinson Arithmetic, a weaker subsystem of PA, would also do. I choose PA instead to facilitate the comparison between our truth theories and other systems that are found in the literature.

[6] However, to know whether a schematic principle holds, the schema of a proof is required.

[7] Since $\mathscr{L}$ contains enough function symbols for the proof of the Strong Diagonal Lemma (cf. Theorem 2) to go through, it cannot have a function symbol for the value relation (which is also a function), on pain of triviality. Nonetheless, I write $x^\circ = y$ instead of, e.g. $\mathrm{Val}(x, y)$, to preserve readability, as is customary.

Theorem 1 (Diagonalization) *For every formula $\varphi(x, \mathbf{v})$ there is a formula $\psi(\mathbf{v})$ s.t. the (universal closure of the) following equivalence is a theorem of* PAT*:*[8]

$$\psi(\mathbf{v}) \leftrightarrow \varphi(\ulcorner\psi\urcorner, \mathbf{v}) \tag{1}$$

In equivalences of the form (1), $\psi(\mathbf{v})$ is said to be a fixed point of $\varphi(x, \mathbf{v})$. Let φ in Theorem 1 be $\neg \mathrm{T}x$. Then there is a sentence λ such that the following is a theorem of PAT:

$$\lambda \leftrightarrow \neg \mathrm{T}\ulcorner\lambda\urcorner \tag{2}$$

λ is normally understood as saying of itself that it is untrue, as a liar sentence. Given (2), no consistent extension of PAT can contain an instance of the T-schema for λ and, *a fortiori*, full disquotation is untenable. This is what Tarski's undefinability result consists in. Likewise, if we opt for a semantic account instead, no model $\langle \mathbb{N}, \Gamma \rangle$ of $\mathscr{L}_{\mathrm{T}}$ can validate unrestricted versions of disquotation, since all theorems of PAT are true in $\langle \mathbb{N}, \Gamma \rangle$, including (2). Thus, we say λ is paradoxical.

To avoid paradox, Tarski opted to restrict disquotation to sentences without T, but more permissive restrictions are possible without stepping into triviality. Thus, we wonder with Leitgeb [21, p. 156], "What kinds of sentences with truth predicate may be inserted plausibly and consistently into the T-scheme?" An idea that suggests itself is to restrict our disquotational principles to non-paradoxical expressions, that is, to those that can be consistently inserted in the T-schema. Alas, McGee [24] has shown that maximal consistent sets of T-biconditionals decide each and every sentence of $\mathscr{L}_{\mathrm{T}}$, which means they are far too complex for an axiomatization. Moreover, McGee's result shows there are uncountably many of those sets, so picking one of them amounts to an arbitrary choice. Consider the following 2-liar cycle:

$$\begin{aligned} \lambda_1 &\leftrightarrow \neg \mathrm{T}\ulcorner\lambda_2\urcorner \\ \lambda_2 &\leftrightarrow \mathrm{T}\ulcorner\lambda_1\urcorner \end{aligned} \tag{3}$$

Theorem 1 guarantees that both biconditionals are provable in PAT if we diagonalize the predicate $\neg \mathrm{T}\ulcorner \mathrm{T}\dot{x}\urcorner$. Given a formula φ with exactly one free variable v, let $\ulcorner\varphi(\dot{v})\urcorner$ be short for $\ulcorner\varphi\urcorner(\dot{v}/\ulcorner v\urcorner)$, where $\dot{x}$ is a term of $\mathscr{L}$ for the p.r. function that maps each natural number to the code of its numeral. Since v is free in $\ulcorner\varphi(\dot{v})\urcorner$, we can quantify over it. Clearly, the biconditionals in (3) are inconsistent with the T-biconditionals for λ_1 and λ_2. However, if one of these T-biconditionals is dropped, consistency is restored. Every maximal consistent set will contain one of them but not the other.

A more promising suggestion has been made by Leitgeb [21]. Roughly, his idea is to restrict the T-schema to grounded sentences, that is, sentences whose truth value ultimately depends on expressions not containing the truth predicate. Since the truth value of the liar and of sentences in liar cycles seems to depend, directly or indirectly, on these very same sentences, the latter are considered ungrounded and

[8] See Carnap [2] and Montague [25]. See Picollo [27] for a proof.

excluded from our disquotational principles. The truth value of other expressions such as $0 = 0$, $\mathrm{T}\ulcorner 0 = 0\urcorner$, and $\mathrm{T}\ulcorner\lambda\urcorner \vee \neg\mathrm{T}\ulcorner\lambda\urcorner$, instead, is ultimately fixed by non-semantic facts, so their corresponding T-biconditionals hold in the theory. Leitgeb's theory consists of a model $\langle \mathbb{N}, \Phi_{\mathrm{lf}} \rangle$ of $\mathscr{L}_{\mathrm{T}}$ where only grounded sentences belong to Φ_{lf} and all instances of disquotation for grounded sentences are true. The theory is quite natural, elegant, and seemingly free from adhocness. However, it is also fairly complex.[9] This means that an $\mathbb{N}$-categorical axiomatization is not possible, i.e. the class of purely truth-theoretic principles the theory entails is not recursively axiomatizable.[10] Schindler [30] provides a nice axiomatic system, but it is naturally far from capturing the original semantic construction.

Others have put forward simpler, syntactic restrictions on disquotation. Halbach [9] explores the restriction of uniform disquotation to formulae in which T occurs only positively – i.e. in the scope of an even number of negations and conditional antecedents. The resulting system is known as PUTB, for "Positive Uniform Tarski Biconditionals". Schindler [31], in turn, considers restricting the Uniform T-schema to translations of formulae of the language of second-order arithmetic without second-order parameters. Both criteria seem somewhat unnatural and ad hoc.

I would like to propose an alternative path, namely, to restrict disquotation to expressions that exhibit 'safe' reference patterns. According to the self-reference diagnosis, all paradoxical expressions share a common reference pattern, i.e. self-reference. This seems to be certainly the case of λ and the liar cycle given by λ_1 and λ_2. If this hypothesis were correct, a sensible plan would be to restrict disquotation to non-self-referential expressions. The prior lack of adequate and precise notions of reference and self-reference has prevented us so far from exploring this route. Luckily, this situation has been remedied. In my companion paper [27] I give a systematic and formal account of reference in the context of truth, designed specifically for the study of the reference patterns underlying paradoxical sentences. Moreover, since according to my account reference has a syntactic aspect, restricting disquotation to non-self-referential expressions could turn out to be simple enough for the formulation of axiomatic truth systems.

Unfortunately, the notions I put forward in [27] reveal that the self-reference diagnosis is not correct. As it turns out, there are semantic paradoxes that are free of self-reference. Such is the case of the Visser-Yablo paradox, consisting of an infinite list of sentences, each of which says of the ones coming after that they are untrue. The existence of this list can be proved in PAT by Theorem 1. Diagonalizing the formula $\forall z\,(z > w \rightarrow \neg\mathrm{T}x(\dot{z}/\ulcorner w\urcorner))$, we obtain a predicate $\mathrm{Y}(w)$ such that

$$\forall w\,(\mathrm{Y}(w) \leftrightarrow \forall z\,(z > w \rightarrow \neg\mathrm{T}\ulcorner\mathrm{Y}(\dot{z})\urcorner)) \tag{4}$$

[9] Leitgeb shows that Φ_{lf} is Π^1_1-complete, i.e. the simplest formula expressing this set begins with a string of second-order universal quantifiers followed by a formula of $\mathscr{L}$.

[10] The notion of $\mathbb{N}$-categoricity has been introduced and put forward as a reasonable criterion for the axiomatizability of semantic theories by Fischer et al. [6].

is provable in PAT. Instantiating w in each numeral results in the following biconditionals, i.e. the list:

$$\begin{aligned} \mathrm{Y}(0) &\leftrightarrow \forall z\,(z > 0 \rightarrow \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}) \\ \mathrm{Y}(\overline{1}) &\leftrightarrow \forall z\,(z > \overline{1} \rightarrow \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}) \\ &\cdots \\ \mathrm{Y}(\overline{n}) &\leftrightarrow \forall z\,(z > \overline{n} \rightarrow \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}) \\ &\cdots \end{aligned} \tag{5}$$

By reductio ad absurdum, the T-biconditionals for $\mathrm{Y}(\overline{n})$ entail $\neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\overline{n})^{\urcorner}$ for each $n \in \omega$, as well as $\neg\forall z\, \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}$. Nonetheless, $\forall z\, \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}$ does not follow in PAT plus the T-biconditionals. This means that the theory is consistent, albeit ω-inconsistent.[11] Note, however, that no model $\langle \mathbb{N}, \Gamma \rangle$ of $\mathscr{L}_{\mathrm{T}}$ can make all T-biconditionals for each $\mathrm{Y}(\overline{n})$ true at the same time: since each $\neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\overline{n})^{\urcorner}$ would have to be true in the model, so would $\forall z\, \neg \mathrm{T}^{\ulcorner}\mathrm{Y}(\dot{z})^{\urcorner}$. For these reasons, the Visser-Yablo paradox is not considered to be a a paradox in the strict sense, but an ω-paradox. Despite not directly trivializing our axiomatic theories, it is still problematic, as no semantic truth theory can validate all T-biconditionals for the sentences in the list. Furthermore, although (4) is consistent with the T-biconditionals, an outright inconsistency can be obtained by combining (4) with an instance of *uniform* disquotation for $\mathrm{Y}(w)$.

According to my account of alethic reference, no sentence in the Visser-Yablo list is self-referential but they are all *unfounded*, as will be seen in Section 2.3. This shows that restricting disquotation to non-self-referential expressions is not a viable project; banning self-reference is not enough. However, as will be shown in Section 3, there are other reference patterns shared by unparadoxical expressions that prove to be sensible restrictions on disquotation. Moreover, their relative simplicity will allow us in Section 4 to formulate proof-theoretic approximations that we then deploy in the formulation of sound and encompassing axiomatic theories of disquotational truth. But before we get to that, I will briefly introduce my general account of alethic reference and related notions.

2.3 Alethic reference

Before giving a formal definition of alethic reference, I will briefly discuss its motivation and some of its general features.

Alethic reference (just "reference", from now on) is introduced in order to study the reference patterns underlying the semantic paradoxes. Accordingly, it is a relation

[11] See Hardy [14] and Ketland [18, 19]. A theory formulated in an extension of $\mathscr{L}$ is said to be ω-inconsistent just in case there is a formula $\varphi(x)$ such that, for each $n \in \omega$, $\varphi(\overline{n})$ is a theorem and, at the same time, the theory entails $\neg\forall x\, \varphi(x)$. An ω-inconsistent theory may be nonetheless consistent, as inferring $\forall x\, \varphi(x)$ from the set of all its instances $\varphi(\overline{n})$ would require an infinitary rule, not admissible in finitary systems such as the ones we are working with.

between sentences of $\mathscr{L}_{\mathrm{T}}$. Sentences can refer to one another in two different ways: by mention and by quantification. For a sentence φ to refer by mention – "m-refer" – to a sentence ψ the former must contain a closed term denoting (the code of) ψ; and for it to refer to ψ by quantification – "q-refer" – (the code of) ψ must fall under the range of the quantifiers in φ, which might be restricted by predicates in a sense to be specified. Since we are only interested in reference in the context of truth, to determine the m- and q-referents of a sentence, we will focus only on terms occurring in the scope of T. As a consequence, only sentences containing T can (alethically) refer. Finally, it is worth pointing out that since reference is strongly tied to the presence of terms in sentences, the notion cannot be closed under logical equivalence, on pain of triviality (cf. footnote 13). In other words, reference is not extensional, but hyperintensional.

We turn now to the definitions.

Definition 1 (M-reference) Let φ and ψ be sentences. φ *m-refers* to ψ iff φ contains a subsentence of the form $\mathrm{T}t$ and $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.

A sentence m-refers only to those sentences denoted by *closed* terms that occur immediately after T. For instance, let $\dot{\neg}$ be a function symbol of $\mathscr{L}$ representing the p.r. function that maps each formula to its negation – and similarly for other logical connectives – and let $\mathrm{Bew}_{\mathsf{PA}}(x)$ weakly represent provability in PA in a natural way.[12] According to the definition of m-reference, while $\neg\mathrm{T}\ulcorner 0 = 0\urcorner$ m-refers to $0 = 0$, $\mathrm{T}\dot{\neg}\ulcorner 0 = 0\urcorner$ m-refers to *the negation* of $0 = 0$ – but not to $0 = 0$ itself – and $\mathrm{Bew}_{\mathsf{PA}}(\ulcorner 0 = 0\urcorner)$ doesn't m-refer to any expression. Thus, proper subterms in the scope of T and T-free sentences don't play a role in m-reference.

Note also that while $\mathrm{T}\ulcorner 0 = 0\urcorner \rightarrow \mathrm{T}\ulcorner 0 = 0\urcorner$ m-refers to $0 = 0$, $0 = 0 \rightarrow 0 = 0$ doesn't, so m-reference is not closed under logical – let alone material – equivalence.[13] For this reason, one should not expect the equivalences delivered by Theorem 1 to bring forth self-m-referential sentences. However, since $\mathscr{L}$ contains a function symbol for substitution, there is a stronger version of this result that does so, provided that $\mathrm{T}x$ is a subformula of φ.

Theorem 2 (Strong Diagonalization) *For every formula* $\varphi(x, \mathbf{v})$ *there is a term* t *s.t.* $t = \ulcorner\varphi(t, \mathbf{v})\urcorner$ *is a theorem of* PA.[14]

We say that $\varphi(t, \mathbf{v})$ is a strong fixed point of $\varphi(x, \mathbf{v})$. Strongly diagonalizing, for instance, the predicate $\neg\mathrm{T}x$, we obtain in PA the following identity:

$$\mathrm{l} = \ulcorner\neg\mathrm{T}\mathrm{l}\urcorner$$

that is, a sentence that m-refers to itself. $\neg\mathrm{Tl}$ is a 'strong' liar sentence.

[12] See Halbach & Visser [12, 13] for a discussion on natural representations.

[13] If it were, every sentence φ would refer to every other sentence ψ, as φ and, e.g. $\varphi \wedge (\mathrm{T}\ulcorner\psi\urcorner \rightarrow \mathrm{T}\ulcorner\psi\urcorner)$ are logically equivalent, and the latter obviously m-refers to ψ.

[14] See Jeroslow [17] for a proof.

Q-reference is considerably more complicated. The underlying idea is the following. While, e.g. $\forall x\, \mathrm{T}x$ q-refers unrestrictedly to every expression, sentences such as PA's global reflection principle,

$$(\mathrm{GRfn}_{\mathsf{PA}}) \qquad \forall x\, (\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$$

q-refer only to the theorems of PA. In general, universal quantifiers followed by conditionals are to be understood as *restricted*, i.e. ranging over just the sentences satisfying the antecedent, with one proviso: that the antecedents are T-free. In this way, we are able to determine what the sentence q-refers to.[15] Otherwise, the quantifiers are taken to be unrestricted.

Giving a formal definition of q-reference requires some additional setup. We first need to introduce the procedure of *normalizing* formulae of $\mathscr{L}_{\mathrm{T}}$; q-reference for sentences is then defined in terms of their normalizations. Since distinct sentences can have the same normalization, the procedure therefore induces an equivalence relation – *having the same normalization* – under which q-reference is closed. Triviality is avoided, however, since this equivalence relation is more fine grained than logical equivalence.

The normalization of an expression is the result of a series of logically valid transformations that deliver a formula in *alethic disjunctive normal form* (ADNF). Call "prime" any atomic or universal formulae or their negation. A formula is in ADNF just in case it contains no conditionals and no existential or dummy quantifiers, and every subformula of the form $\forall v\, \varphi$ is s.t. (i) φ is a disjunction (of length ≥ 0) of conjunctions (of length ≥ 0) of primes and, (ii) if φ is of the form $\psi \vee \chi$, ψ contains all and only the T-free disjuncts of φ (if any) and χ all of the T-containing disjuncts (if any). The point of normalization is that one can easily see whether sentences in ADNF take the form of a restricted quantified claim. Consider a sentence $\forall v\, \varphi$ in ADNF containing T. If φ is a disjunction $\psi \vee \chi$ with some T-free disjuncts, $\forall v\, \varphi$ can be seen to be equivalent to the restricted quantificational claim $\forall v\, (\neg\psi \rightarrow \chi)$, for the disjuncts of ψ must be T-free whereas the disjuncts of χ must not. In this way, the reference-restricting conditional $\neg\psi \rightarrow \chi$ is made explicit, and $\neg\psi$ is guaranteed to encapsulate all truth-free restrictions imposed on $\forall v$. If, instead, φ is not a disjunction or contains no T-free disjuncts, the quantifiers in $\forall v\, \varphi$ are unrestricted.

Since formulae in ADNF cannot contain conditionals or existential quantifiers, the first step in normalizing an expression is to replace these connectives with negations, conjunctions, disjunctions, and universal quantifiers, making use of the standard definitions. Let $\tau : \mathscr{L}_{\mathrm{T}} \rightarrow \mathscr{L}_{\mathrm{T}}$ carry out these replacements, that is, $\tau(\varphi \rightarrow \psi) := \neg\tau(\varphi) \vee \tau(\psi)$ and $\tau(\exists v\, \varphi) := \neg\forall v\, \neg\tau(\varphi)$. Then, the normalization process is done in stages. It consists of successive transformations of each subfor-

[15] Recall there is no standard interpretation of T yet. Our goal is, instead, to define reference to formulate restrictive criteria for disquotation in terms of this notion and *then* learn about truth and its extension.

mula of the form $\forall v\, \varphi$ into ADNF, starting from those of lesser *depth*. Let φ contain no conditionals, existential, or dummy quantifiers.

$$dep(\forall v\, \varphi) = \begin{cases} 1 & \text{if } \varphi \text{ is atomic} \\ dep(\forall v\, \psi) & \text{if } \varphi := \neg\psi \\ max\{dep(\forall v\, \psi), dep(\forall v\, \chi)\} & \text{if } \varphi := (\psi \wedge \chi) \\ max\{dep(\forall v\, \psi), dep(\forall v\, \chi)\} & \text{if } \varphi := (\psi \vee \chi) \\ dep(\forall u\, \psi) + 1 & \text{if } \varphi := \forall u\, \psi \end{cases}$$

For each $i \in \omega$, the i-normalization of φ is the result of successively applying the following transformations to each subformula $\forall v\, \psi$ of depth i:

1. Replace every subformula of the form $\neg(\psi_1 \vee \psi_2)$ and $\neg(\psi_1 \wedge \psi_2)$ with $(\neg\psi_1 \wedge \neg\psi_2)$ and $(\neg\psi_1 \vee \neg\psi_2)$ resp. until they don't occur any longer, starting with the innermost.
2. Erase all double negations.
3. Replace every subformula of the form $\psi_1 \wedge (\psi_2 \vee \psi_3)$ and $(\psi_2 \vee \psi_3) \wedge \psi_1$ with $(\psi_1 \wedge \psi_2) \vee (\psi_1 \wedge \psi_3)$ and $(\psi_2 \wedge \psi_1) \vee (\psi_3 \wedge \psi_1)$ resp. until they don't occur any longer, starting with the innermost.
4. In every subformula of the form $\forall v(\psi_1 \vee \cdots \vee \psi_m)$ (where each ψ_i, $1 \leq i \leq m$, is not itself a disjunction), rearrange the disjuncts into $\chi_1 \vee \chi_2$ such that the ones not containing T (if any) occur in χ_1, whilst the others (if any) occur in χ_2.

Definition 2 (Normalization) The normalization φ^* of a formula φ is the result of erasing all dummy quantifiers in $\tau(\varphi)$ and then, if there are any quantifiers left, performing successive i-normalizations starting with $i = 1$ and stopping after $i = max\{dep(\forall v\, \psi) : \forall v\, \psi \text{ is a subformula of } \varphi\}$.

Since every step in the i-normalization of a formula involves only finitely many transformations and formulae contain finitely many quantifiers, the normalization process always terminates. Moreover, it delivers a logically equivalent expression.

We are finally in a position to provide an adequate and precise definition of q-reference. Let $\mathbf{n}$ abbreviate $n_1, \ldots, n_m \in \omega$, and $\overline{\mathbf{n}}$ abbreviate $\overline{n_1}, \ldots, \overline{n_m}$.

Definition 3 (Q-reference) Let φ, ψ be sentences. φ q-refers to ψ iff φ^* has a subsentence of the form $\forall \mathbf{v}\, \chi$ s.t. T occurs in χ, and one of the following holds:

1. $\chi := (\chi_1 \vee \chi_2)$, T doesn't occur in χ_1, and $\chi_2[\overline{\mathbf{n}}/\mathbf{v}]$ q-refers or newly m-refers to ψ,[16] for some $\mathbf{n} \in \omega$ s.t. $\mathbb{N} \vDash \neg\chi_1[\overline{\mathbf{n}}/\mathbf{v}]$.
2. Either $\chi := (\chi_1 \vee \chi_2)$ and T occurs in χ_1 or χ is not a disjunction or a universal statement, and $\chi[\overline{\mathbf{n}}/\mathbf{v}]$ q-refers or newly m-refers to ψ, for some $\mathbf{n} \in \omega$.

In brief, the q-referents of universally quantified claims are the sentences their (possibly restricted) instances m- or q-refer to. More precisely, a sentence whose

[16] To keep m- and q-reference apart, we require here that m-reference is achieved through 'new' closed terms, that is, that the latter are not present in the sentence itself but are a product of the instantiation of the quantifiers at issue.

normalization contains a subsentence of the form $\forall \mathbf{v}\, \varphi$ q-refers to what each instance $\varphi[\overline{\mathbf{n}}/\mathbf{v}]$ newly m-refers or q-refers to, unless it φ a reference-restricting conditional, in which case only the instances with true antecedents are to be considered.[17] For instance,

$$\forall x \forall y\, (\mathrm{T}\ulcorner 0 = 0 \urcorner \wedge \mathrm{T}x \dot{\rightarrow} y) \tag{6}$$

which is already in ADNF, q-refers to the m-referents of each instance $\mathrm{T}\ulcorner 0 = 0 \urcorner \wedge \mathrm{T}\overline{m} \dot{\rightarrow} \overline{n}$, provided that the terms delivering m-reference are a product of the instantiation of the quantifiers $\forall x \forall y$, which must be eliminated together 'at once'. Thus, (6) q-refers just to every conditional sentence. On the other hand,

$$\forall x\, (x = \ulcorner 0 = 0 \urcorner \rightarrow \forall y\, (y = \dot{\neg} x \rightarrow \mathrm{T}y)) \tag{7}$$

which normalizes into $\forall x\, (x \neq \ulcorner 0 = 0 \urcorner \vee \forall y\, (y \neq \dot{\neg} x \vee \mathrm{T}y))$, q-refers to the *q*-referents of each instance $\forall y\, (y \neq \dot{\neg} \overline{n} \vee \mathrm{T}y))$ provided that $\mathbb{N} \vDash \overline{n} = \ulcorner 0 = 0 \urcorner$: i.e. (7) q-refers to the negation of $0 = 0$.

Just as the definition of m-reference accounts for the self-referentiality of certain sentences delivered by the Strong Diagonal Lemma, Definition 3 implies that 'weakly' diagonalizing certain predicates – e.g. in which $\mathrm{T}x$ occurs as a subformula – also delivers self-referential sentences.[18] This implies that λ is self-q-referential. The Visser-Yablo sentences, instead, only q-refer to sentences coming later in the sequence; they are not self-q-referential.

Mention and quantification exhaust the ways in which a sentence can directly refer to other expressions. Thus, we have the following definition of direct reference – "d-reference".

Definition 4 (D-reference) Let φ, ψ be sentences. φ directly refers to ψ iff it m- or q-refers to ψ.

Observation 3 *For all $\varphi, \psi, \chi \in \mathscr{L}_{\mathrm{T}}$:*

1. *If $\varphi \in \mathscr{L}$, φ doesn't d-refer to ψ.*
2. *If $\mathbb{N} \vDash s = t$, φ and $\varphi[s/t]$ d-refer to the same sentences.*
3. *φ and $\neg\varphi$ d-refer to the same sentences.*
4. *$\varphi \vee \chi$, $\varphi \wedge \chi$, and $\varphi \rightarrow \chi$ d-refer to ψ iff either φ or χ do.*
5. *If v is not free in φ, φ, $\forall v\, \varphi$, and $\exists v\, \varphi$ d-refer to the same sentences.*
6. *$\forall v\, \varphi$ and $\forall u\, \varphi[u/v]$ d-refer to the same sentences, if u is free for v in φ,*
7. *The following pairs of logical equivalents d-refer to the same sentences:*
   - *φ and $\neg\neg\varphi$,*
   - *$\varphi \vee \psi$ and $\psi \vee \varphi$,*

[17] Several worked examples, together with additional motivation and discussion, can be found in Picollo [27].

[18] See Picollo [27, Section 3.3].

- $\varphi \wedge \psi$ *and* $\psi \wedge \varphi$,
- $(\varphi \vee \psi) \vee \chi$ *and* $\varphi \vee (\psi \vee \chi)$,
- $(\varphi \wedge \psi) \wedge \chi$ *and* $\varphi \wedge (\psi \wedge \chi)$,
- $\varphi \rightarrow \psi$ *and* $\neg\psi \rightarrow \neg\varphi$,
- $\varphi \vee (\psi \wedge \chi)$ *and* $(\varphi \vee \psi) \wedge (\varphi \vee \chi)$,
- $\varphi \wedge (\psi \vee \chi)$ *and* $(\varphi \wedge \psi) \vee (\varphi \wedge \chi)$,
- $\neg(\varphi \vee \psi)$ *and* $\neg\varphi \wedge \neg\psi$,
- $\neg(\varphi \wedge \psi)$ *and* $\neg\varphi \vee \neg\psi$,
- $\varphi \rightarrow \psi$ *and* $\neg\varphi \vee \psi$,
- $\exists \mathbf{v}\, \varphi$ *and* $\neg\forall \mathbf{v}\, \neg\varphi$.

Definition 4 allows us to characterize many 'mixed' reference patterns. Consider, for instance, sentences λ_1 and λ_2 in the 2-liar cycle in (3). While λ_1 m-refers to λ_2, the latter only q-refers to the former. Thus, we can say they directly refer to each other. However, in a very clear sense they also refer to themselves, albeit *indirectly*. Otherwise, we would get a semantic paradox without self-reference on the cheap. The following notions are intended to deal with this and other similar cases.

Definition 5 (Chain of reference) A (possibly infinite) sequence of sentences s.t. each sentence in the sequence d-refers to the one coming after, if any.

Definition 6 (Reference) Let φ, ψ be sentences. φ refers to ψ iff there's a chain of reference starting with φ and ending with ψ.

We can employ the notions just introduced to define salient reference patterns, such as the following.

Definition 7 (Self-reference) A sentence is self-referential iff it refers to itself.

Definition 8 (Well-foundedness) A sentence φ is well-founded iff all chains of reference starting with φ are finite. Otherwise, we say that φ is unfounded.

Thus, sentences that don't d-refer to any expressions – e.g. the theorems of PA – are well-founded. And sentences that only refer to well-founded expressions are well-founded too. On the other hand, every self-referential expression is obviously unfounded. But there are also unfounded sentences that don't refer to themselves, such as the Visser-Yablo sentences in (5). Thus, not all paradoxical expressions are self-referential. However, the notions introduced in this section can still be deployed in the formulation of restrictions to our disquotational truth principles. As will be seen in the next section, all paradoxical expressions are unfounded, for theories in which disquotation is restricted to well-founded sentences will be shown to be (ω-) consistent. What is more, we will show that this characterization can be refined even further, prompting more encompassing truth systems.

3 Semantic Theories of Truth

In this section we deploy the notions of alethic reference introduced in Section 2.3 to prove the existence of models of $\mathscr{L}_{\mathrm{T}}$ expanding $\mathbb{N}$ which verify large and well-motivated sets of instances of disquotation. In other words, we put forward semantic truth theories. In turn, these will serve as witnesses to the consistency and arithmetical soundness of the axiomatic systems we introduce in Section 4.

3.1 Well-Founded Truth

Our first theory consists of a single model $\langle \mathbb{N}, \Phi_{\mathrm{wf}} \rangle$ of $\mathscr{L}_{\mathrm{T}}$ in which the extension assigned to T, Φ_{wf}, contains all and only true well-founded sentences. Thus, all instances of the T-schema for well-founded sentences hold in the model. Moreover, since $\langle \mathbb{N}, \Phi_{\mathrm{wf}} \rangle$ expands the standard model of arithmetic, the Uniform T-schema restricted to formulae with only well-founded numerical instances also holds in $\langle \mathbb{N}, \Phi_{\mathrm{wf}} \rangle$.

The set Φ_{wf} is obtained via a usual Kripke-style construction, by considering a transfinite sequence of sets $\Phi_\alpha \subseteq \omega$, with $\alpha \in \mathrm{On}$ (the class of all ordinals), of which Φ_{wf} is shown to be a fixed point. In order to construct this sequence, we first consider another sequence of sets Ψ_α whose fixed point, Ψ_{wf}, is the set of well-founded sentences. Thus, this sequence 'stratifies' this set. If φ is a sentence, let Θ_φ be the set of sentences φ d-refers to. For each $\alpha \in \mathrm{On}$, Ψ_α is defined as follows:

$$\Psi_\alpha := \begin{cases} \varnothing & \text{if } \alpha = 0 \\ \{\varphi : \Theta_\varphi \subseteq \Psi_\beta\} & \text{if } \alpha = \beta + 1 \\ \bigcup_{\beta<\alpha} \Psi_\beta & \text{if } \alpha \text{ is a limit} \end{cases}$$

Lemma 1 *For all $\alpha, \beta \in \mathrm{On}$, if $\alpha < \beta$, then $\Psi_\alpha \subseteq \Psi_\beta$.*

Proof By transfinite induction on β. Let $\alpha < \beta$ and $\varphi \in \Psi_\alpha$. If β is 0 or a limit ordinal, the result follows trivially. Let $\beta = \xi + 1$. By inductive hypothesis (i.h.), $\varphi \in \Psi_\xi$. Thus, $\xi \neq 0$. If $\xi = \zeta + 1$, $\Theta_\varphi \subseteq \Psi_\zeta$. Again, by i.h., $\Psi_\zeta \subseteq \Psi_\xi$, so $\Theta_\varphi \subseteq \Psi_\xi$. Therefore, $\varphi \in \Psi_\beta$. The case in which ξ is a limit can be proved in a similar way. □

Proposition 1 *There is an $\alpha \in \mathrm{On}$ s.t., for every $\beta > \alpha$, $\Psi_\alpha = \Psi_\beta$ and Ψ_α is the set of well-founded sentences.*

Proof The first conjunct follows immediately from Lemma 1 and cardinality considerations. Therefore, the sequence reaches a fixed point, Ψ_{wf}. We show that Ψ_{wf} is the set of well-founded sentences.

Let $\varphi \in \Psi_{\mathrm{wf}}$. Thus, $\varphi \in \Psi_\alpha$, for some $\alpha \in \mathrm{On}$. We show by transfinite induction on α that all sentences in Ψ_α are well-founded. Assume that, for every $\beta < \alpha$, Ψ_β contains only well-founded sentences. If $\alpha = 0$, the result follows trivially. If $\alpha = \xi + 1$, then ξ is a set of well-founded sentences, by i.h.. Then, all members of Ψ_α d-refer just to well-founded sentences and are also well-founded. If α is a limit ordinal, the result follows trivially from the i.h. as well.

Now let φ be a well-founded sentence. By Definition 8, either φ doesn't refer or every chain of reference starting with φ is finite. Thus, the following function from the set of well-founded sentences to On is well defined:

$$f(\varphi) := \begin{cases} 0 & \text{if } \Theta_\varphi = \varnothing \\ sup\{f(\psi) + 1 : \varphi \text{ d-refers to } \psi\} & \text{otherwise} \end{cases}$$

I show that $\varphi \in \Psi_{f(\varphi)+1}$ by transfinite induction on $f(\varphi)$. If $f(\varphi) = 0$, $\Theta_\varphi = \varnothing$, so $\Theta_\varphi \subseteq \Psi_0$, so $\varphi \in \Psi_1$. Assume the result holds for every $\alpha < f(\varphi)$ and let $f(\varphi) = \beta + 1$. The case in which $f(\varphi)$ is a limit is similar. Then, β is the supremum of all $f(\psi)$ s.t. φ d-refers to ψ. By i.h. and Lemma 1, $\Theta_\varphi \subseteq \Psi_{\beta+1}$, which means $\varphi \in \Psi_{f(\varphi)+1}$. By Lemma 1 and the first conjunct of this proposition, we can conclude that $\varphi \in \Psi_{\text{wf}}$. □

We have now paved the way for the construction of the sequence of Φ_α, with $\alpha \in$ On, that will give us the desired extension of the truth predicate. At each ordinal α, Φ_α contains only sentences that are well-founded at this stage, i.e. that belong to Ψ_α.

$$\Phi_\alpha := \begin{cases} \varnothing & \text{if} \alpha = 0 \\ \{\varphi : \langle \mathbb{N}, \Phi_\beta \rangle \vDash \varphi\} \cap \Psi_\alpha & \text{if} \alpha = \beta + 1 \\ \bigcup_{\beta<\alpha} \Phi_\beta & \text{if } \alpha \text{ is a limit ordinal} \end{cases}$$

To show that the sequence reaches a fixed point we first need to prove two lemmata. According to the first, the truth value of a statement in a model is not affected by the mere addition or removal of sentences it doesn't d-refer to from the extension of the truth predicate. This establishes a link between direct reference and Leitgeb's [21] notion of dependence.[19] More specifically, it follows that φ depends on Θ_φ. The converse doesn't hold, that is, Θ_φ is not a subset of every set φ depends on as, e.g. $\text{T}\ulcorner\lambda\urcorner \rightarrow \text{T}\ulcorner\lambda\urcorner$ d-refers to λ but depends on $\varnothing$. Just like aboutness, dependence is not as tied to the syntactic structure of sentences as reference is.

Lemma 2 *Let Γ be a set of sentences. If $\Theta_\varphi \subseteq \Gamma$, then for every set of sentences Δ,*

$$\langle \mathbb{N}, \Delta \rangle \vDash \varphi \text{ iff } \langle \mathbb{N}, \Gamma \cap \Delta \rangle \vDash \varphi.$$

Proof Since every sentence is logically equivalent to its normalization (cf. Definition 2), we can prove the result for the normalization of φ, φ^*, by induction on its complexity (number of logical operators). If $\varphi^* \in \mathscr{L}$, φ^* is true in every expansion of $\mathbb{N}$ or in none, so the result follows trivially. Thus, we assume φ^* contains T.

If φ^* is an atomic sentence, then it's of the form Tt, where either t denotes a sentence in $\mathbb{N}$ or it doesn't. If it doesn't, then φ^* is false both in $\langle \mathbb{N}, \Delta \rangle$ and in $\langle \mathbb{N}, \Gamma \cap \Delta \rangle$. If t denotes a sentence ψ, then, by Definition 1, $\psi \in \Theta_\varphi \subseteq \Gamma$. Thus, $\psi \in \Delta$ iff $\psi \in \Gamma \cap \Delta$, so we have the desired result.

[19] A sentence φ is said to depend on a set of sentences Γ iff, for every set of sentences Δ, $\langle \mathbb{N}, \Delta \rangle \vDash \varphi$ iff $\langle \mathbb{N}, \Gamma \cap \Delta \rangle \vDash \varphi$.

Assume the claim holds for every sentence of lower complexity than φ^*. Since d-reference is closed under logical connectives (cf. Observation 3), if φ is a negation, conjunction, or disjunction, the result follows trivially from the i.h.

Let $\varphi^* := \forall \mathbf{v}\, \psi$, where ψ is not a universal statement. φ^* contains no dummy quantifiers. Let $\psi := (\psi_1 \vee \psi_2)$ s.t. T occurs in every disjunct of ψ_2 but in none of ψ_1. Otherwise, the proof is similar but with less complications. Since $\psi_1 \in \mathscr{L}$, for each $\mathbf{n} \in \omega$ either $\mathbb{N} \vDash \psi_1[\overline{\mathbf{n}}/\mathbf{v}]$ or $\mathbb{N} \vDash \neg\psi_1[\overline{\mathbf{n}}/\mathbf{v}]$. Note that if $\mathbb{N} \vDash \psi_1[\overline{\mathbf{n}}/\mathbf{v}]$, then $\langle \mathbb{N}, \Delta\rangle \vDash \psi[\overline{\mathbf{n}}/\mathbf{v}]$ for every set of sentences Δ.

Assume then that $\mathbb{N} \vDash \neg\psi_1[\overline{\mathbf{n}}/\mathbf{v}]$. Thus, for every Δ, $\langle \mathbb{N}, \Delta\rangle \vDash \psi[\overline{\mathbf{n}}/\mathbf{v}]$ iff $\langle \mathbb{N}, \Delta\rangle \vDash \psi_2[\overline{\mathbf{n}}/\mathbf{v}]$. Since $\Theta_{\forall \mathbf{v}\, \psi} \subseteq \Gamma$ and, by Definition 3, $\Theta_{\forall \mathbf{v}\, \psi}$ is the union of all $\Theta_{\psi_2[\overline{\mathbf{n}}/\mathbf{v}]}$ s.t. $\mathbb{N} \vDash \neg\psi_1[\overline{\mathbf{n}}/\mathbf{v}]$, we have that $\Theta_{\psi_2[\overline{\mathbf{n}}/\mathbf{v}]} \subseteq \Gamma$. By i.h., $\langle \mathbb{N}, \Delta\rangle \vDash \psi_2[\overline{\mathbf{n}}/\mathbf{v}]$ iff $\langle \mathbb{N}, \Gamma \cap \Delta\rangle \vDash \psi_2[\overline{\mathbf{n}}/\mathbf{v}]$. By our last assumption, this entails that $\langle \mathbb{N}, \Delta\rangle \vDash (\psi_1 \vee \psi_2)[\overline{\mathbf{n}}/\mathbf{v}]$ iff $\langle \mathbb{N}, \Gamma \cap \Delta\rangle \vDash (\psi_1 \vee \psi_2)[\overline{\mathbf{n}}/\mathbf{v}]$ or, what is the same, $\langle \mathbb{N}, \Delta\rangle \vDash \psi[\overline{\mathbf{n}}/\mathbf{v}]$ iff $\langle \mathbb{N}, \Gamma \cap \Delta\rangle \vDash \psi[\overline{\mathbf{n}}/\mathbf{v}]$.

Thus, in any case we also have that $\langle \mathbb{N}, \Delta\rangle \vDash \forall \mathbf{v}\, \psi$ iff $\langle \mathbb{N}, \Gamma \cap \Delta\rangle \vDash \forall \mathbf{v}\, \psi$. □

Lemma 3 *For all $\alpha, \beta \in$ On, if $\alpha < \beta$, then $\Phi_\alpha \subseteq \Phi_\beta$.*

Proof By transfinite induction on β. If β is not a successor ordinal, the result follows trivially. Let $\beta = \xi + 1$, and assume that, for all $\zeta < \xi$, $\Phi_\zeta \subseteq \Phi_\xi$ and $\alpha < \beta$. Thus, $\alpha \leq \xi$. If $\alpha = \xi$, trivially $\Phi_\alpha \subseteq \Phi_\xi$, and if $\alpha < \xi$, $\Phi_\alpha \subseteq \Phi_\xi$ by i.h.. Thus, it remains to be shown that $\Phi_\xi \subseteq \Phi_\beta$.

Let $\varphi \in \Phi_\xi$. Therefore, $\Phi_\xi \neq \varnothing$, so ξ is either a successor or a limit ordinal. The proof for both cases is similar. I only show it for $\xi = \zeta + 1$. Since $\varphi \in \Psi_\xi$, φ d-refers only to sentences in Ψ_ζ and $\langle \mathbb{N}, \Phi_\zeta\rangle \vDash \varphi$. Moreover, note that $\Phi_\zeta = \Psi_\zeta \cap \Phi_\xi$: by construction of Φ_ζ, $\Phi_\zeta \subseteq \Psi_\zeta$ and, by i.h., $\Phi_\zeta \subseteq \Phi_\xi$. For the other direction, assume for contradiction that $\psi \in \Psi_\zeta \cap \Phi_\xi$ and $\psi \notin \Phi_\zeta$. Thus, $\zeta \neq \varnothing$. Let $\zeta = \theta + 1$; the case in which ζ is a limit ordinal is analogous. Since $\psi \notin \Phi_\zeta$ but $\psi \in \Psi_\zeta$, we have that $\langle \mathbb{N}, \Phi_\theta\rangle \nvDash \psi$, that is, $\langle \mathbb{N}, \Phi_\theta\rangle \vDash \neg\psi$. Since d-reference is closed under negation (cf. Observation 3), we also have that $\neg\psi \in \Psi_\zeta$. Thus, $\neg\psi \in \Phi_\zeta$. By i.h., $\neg\psi \in \Phi_\xi$. As $\psi \in \Phi_\xi$ too, both $\langle \mathbb{N}, \Phi_\zeta\rangle \vDash \psi$ and $\langle \mathbb{N}, \Phi_\zeta\rangle \vDash \neg\psi$, which is impossible.

Therefore, $\langle \mathbb{N}, \Psi_\zeta \cap \Phi_\xi\rangle \vDash \varphi$. By Lemma 2, $\langle \mathbb{N}, \Phi_\xi\rangle \vDash \varphi$. Also, since φ d-refers only to sentences in Ψ_ζ, $\varphi \in \Psi_\xi$, and by Lemma 1, $\varphi \in \Psi_\beta$. Thus, $\varphi \in \Phi_\beta$. □

Proposition 2 *There is an ordinal $\alpha \in$ On s.t. $\Phi_\alpha = \Phi_{\alpha+1}$.*

Proof By Lemma 3 and cardinality considerations. □

The sequence of sets Φ_α stabilizes at some ordinal, it has a fixed point. Let Φ_{wf} be this fixed point. The following result establishes that all instances of the T-schema for well-founded sentences hold in $\langle \mathbb{N}, \Phi_{\mathrm{wf}}\rangle$.

Proposition 3 *For every well-founded sentence φ, $\langle \mathbb{N}, \Phi_{\mathrm{wf}}\rangle \vDash \mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$.*

Proof Let φ be a well-founded expression. By Proposition 1, there is an ordinal α s.t. $\varphi \in \Psi_\alpha$. Thus, $\alpha \neq \varnothing$, so α is either a successor or a limit ordinal. I only prove the result for the latter case, as the proof for the former is analogous. If α is a limit, by Lemma 3 there is a successor ordinal $\beta < \alpha$ s.t. $\varphi \in \Psi_\beta$. As d-reference is closed under negation, we also have that $\neg\varphi \in \Psi_\beta$. Let $\beta = \xi + 1$. Either $\langle \mathbb{N}, \Phi_\xi \rangle \vDash \varphi$ or $\langle \mathbb{N}, \Phi_\xi \rangle \vDash \neg\varphi$, so either $\varphi \in \Phi_\beta$ or $\neg\varphi \in \Phi_\beta$. Therefore, by Lemma 3, $\varphi \in \Phi_{\mathrm{wf}}$ or $\neg\varphi \in \Phi_{\mathrm{wf}}$. Let ζ be the smallest ordinal s.t. $\Phi_\zeta = \Phi_{\mathrm{wf}}$, i.e. the fixed-point ordinal.

If $\varphi \in \Phi_\zeta$, then $\langle \mathbb{N}, \Phi_\zeta \rangle \vDash \mathrm{T}\ulcorner\varphi\urcorner$, but also $\varphi \in \Phi_{\zeta+1}$, since $\Phi_\zeta = \Phi_{\zeta+1}$. Thus, $\varphi \in \{\varphi : \langle \mathbb{N}, \Phi_\zeta \rangle \vDash \varphi\} \cap \Psi_\zeta$, which means that $\langle \mathbb{N}, \Phi_\zeta \rangle \vDash \varphi$. So, we get $\langle \mathbb{N}, \Phi_\zeta \rangle \vDash \mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$.

If, instead, $\neg\varphi \in \Phi_\zeta$, then $\neg\varphi \in \Phi_{\zeta+1}$, so $\neg\varphi \in \{\varphi : \langle \mathbb{N}, \Phi_\zeta \rangle \vDash \varphi\} \cap \Psi_\zeta$. This implies that $\langle \mathbb{N}, \Phi_\zeta \rangle \vDash \neg\varphi$ and, thus, $\langle \mathbb{N}, \Phi_\zeta \rangle \nvDash \varphi$. Therefore, $\varphi \notin \Phi_{\zeta+1}$, which means that $\varphi \notin \Phi_\zeta$, i.e. $\langle \mathbb{N}, \Phi_\zeta \rangle \nvDash \mathrm{T}\ulcorner\varphi\urcorner$. Again, we have that $\langle \mathbb{N}, \Phi_\zeta \rangle \vDash \mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$. □

As a consequence, for every well-founded sentence, either it belongs to Φ_{wf} or its negation does. Note also that, by Lemma 1 and Proposition 1, Φ_{wf} contains only well-founded sentences.

3.2 More Permissive Criteria

Well-foundedness is not the only restriction we can impose on instances of the T-schema in an expansion of $\mathbb{N}$ to $\mathscr{L}_{\mathrm{T}}$ using the notions of alethic reference: more permissive criteria may also be adopted. For instance, note that in the case of the Visser-Yablo paradox, every expression on the list is unfounded, but also there are infinitely many of them and, additionally, each of them d-refers to infinitely many others. It can be shown that, if only finitely many sentences on the list are considered, no ω-inconsistency arises. More generally, let Δ by a *finite* set of non-self-referential but unfounded sentences. One can easily find a model $\langle \mathbb{N}, \Gamma \rangle$ in which all instances of the T-schema for sentences in $\Psi_{\mathrm{wf}} \cup \Delta$ hold.

Proposition 4 *If Δ is a finite set of non-self-referential unfounded sentences, there's a $\Gamma \subseteq \Psi_{\mathrm{wf}} \cup \Delta$ s.t. all instances of the* T-schema *for sentences in $\Psi_{\mathrm{wf}} \cup \Delta$ are true in $\langle \mathbb{N}, \Gamma \rangle$.*

Proof Let $f : \Psi_{\mathrm{wf}} \cup \Delta \to \omega$ be s.t.

$$f(\varphi) := \begin{cases} 0 & \text{if } \Delta \cap \Theta_\varphi = \varnothing \\ max\{f(\psi) : \psi \in \Delta \cap \Theta_\varphi\} + 1 & \text{otherwise} \end{cases}$$

f is well defined, as Δ is finite and contains no self-referential sentences. Thus, every complete chain of reference restricted to members of Δ ends in a sentence that is not in Δ. For each $n \in \omega$, let $\Delta_n := \{\varphi \in \Delta : f(\varphi) = n\}$. Since Δ is finite, only finitely many of these sets are non-empty. Let $k \in \omega$ be the greatest number s.t. $\Delta_k \neq \varnothing$. Thus, $\Delta = \bigcup_{i \leq k} \Delta_i$.

For each $n \in \omega$, let

$$\Xi_n := \begin{cases} \Phi_{\mathrm{wf}} \cup \{\varphi \in \Delta_0 : \langle \mathbb{N}, \Phi_{\mathrm{wf}} \rangle \vDash \varphi\} & \text{if } n = 0 \\ \Xi_m \cup \{\varphi \in \Delta_m : \langle \mathbb{N}, \Xi_m \rangle \vDash \varphi\} & \text{if } n = m + 1 \end{cases}$$

I show that all instances of the T-schema for sentences in $\Psi_{\mathrm{wf}} \cup \Delta$ are true in $\langle \mathbb{N}, \Xi_k \rangle$. By Lemma 2, all instances of the T-schema for well-founded sentences are still true in $\langle \mathbb{N}, \Xi_k \rangle$, for $\Phi_{\mathrm{wf}} \subseteq \Xi_k$ and no well-founded sentence may refer to members of Δ, as they are unfounded. Now let $\varphi \in \Delta$, which means there's an $n \leq k$ s.t. $\varphi \in \Delta_n$. Let $\Xi_{-1} := \Phi_{\mathrm{wf}}$. Therefore, we have that $\langle \mathbb{N}, \Xi_k \rangle \vDash \varphi$ iff $\langle \mathbb{N}, \Xi_{n-1} \rangle \vDash \varphi$ iff $\langle \mathbb{N}, \Xi_n \rangle \vDash \mathrm{T}^{\ulcorner}\varphi^{\urcorner}$ iff, by Lemma 2, $\langle \mathbb{N}, \Xi_k \rangle \vDash \mathrm{T}^{\ulcorner}\varphi^{\urcorner}$, as φ cannot d-refer to any member of Ξ_k that isn't in Ξ_n. Consequently, the T-schema also holds for all sentences in Δ. □

More interestingly, it can be shown that for any *infinite* set Δ of non-self-referential unfounded sentences, each of which d-refers only to a *finite* number of expressions, there is a model $\langle \mathbb{N}, \Gamma \rangle$ in which all instances of the T-schema for sentences in $\Psi_{\mathrm{wf}} \cup \Delta$ are true.

Proposition 5 *If Δ is an infinite set of non-self-referential unfounded sentences, each of which d-refers only to a finite number of expressions, there is a $\Gamma \subseteq \Psi_{\mathrm{wf}} \cup \Delta$ s.t., for every $\varphi \in \Psi_{\mathrm{wf}} \cup \Delta$, $\langle \mathbb{N}, \Gamma \rangle \vDash \mathrm{T}^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi$.*[20]

Proof Let $\Delta := \{\varphi_0, \varphi_1, \dots\}$, $\Delta_n := \{\varphi_0, \dots, \varphi_n\}$, and

$$\mathscr{G}_n := \{\Xi : \Phi_{\mathrm{wf}} \subseteq \Xi \subseteq \Phi_{\mathrm{wf}} \cup \bigcup_{i \leq n} (\Theta_{\varphi_i} \cup \{\varphi_i\}) \;\&\; \langle \mathbb{N}, \Xi \rangle \vDash \mathrm{T}^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi, \text{for} \varphi \in \Delta_n\}$$

Each $\mathscr{G}_n$ is finite, for each $\bigcup_{i \leq n} (\Theta_{\varphi_i} \cup \{\varphi_i\})$ is. I show that, for every $n \in \omega$, we can find sets $\Xi_1 \subseteq \cdots \subseteq \Xi_n$ s.t. $\Xi_i \in \mathscr{G}_i$ for each $i \leq n$.

By Proposition 4, for each $n \in \omega$ there's a model $\langle \mathbb{N}, \Gamma_n \rangle$ with $\Phi_{\mathrm{wf}} \subseteq \Gamma_n \subseteq \Phi_{\mathrm{wf}} \cup \bigcup_{i \leq n} (\Theta_{\varphi_i} \cup \{\varphi_i\})$ in which all instances of the T-schema for sentences in $\Psi_{\mathrm{wf}} \cup \Delta_n$ are true. For $i \leq n$ let

$$\Gamma_n^i := \Gamma_n \cap (\Phi_{\mathrm{wf}} \cup \bigcup_{j \leq i} (\Theta_{\varphi_j} \cup \{\varphi_j\}))$$

Thus, for each $i \leq n$, $\Gamma_n^i \subseteq \Gamma_n^{i+1}$. By Lemma 2, we have that all instances of the T-schema for sentences in Δ_i hold in $\langle \mathbb{N}, \Gamma_n^i \rangle$, so $\Gamma_n^i \in \mathscr{G}_i$.

Let $\mathscr{G}$ be the smallest tree consisting of sequences of members of $\{\varnothing\} \cup \bigcup_{n \in \omega} \mathscr{G}_n$ s.t.:

- $\langle \varnothing \rangle \in \mathscr{G}$,
- if $\langle \Xi_0, \dots, \Xi_n \rangle \in \mathscr{G}$ and $\Xi_n \subseteq \Xi_{n+1} \in \mathscr{G}_{n+1}$, then $\langle \Xi_0, \dots, \Xi_n, \Xi_{n+1} \rangle \in \mathscr{G}$.

[20] This result has been proved in collaboration with Thomas Schindler.

Notice that $\mathscr{G}$ is locally finite, that is, each node has finitely many children, for each $\mathscr{G}_n$ is finite. Moreover, for every $n \in \omega$, $\mathscr{G}$ contains a sequence of length n, e.g. $\langle \Gamma_n^0, \ldots, \Gamma_n^n \rangle$, so $\mathscr{G}$ is an infinite graph. Finally, all sequences start with $\varnothing$, so all nodes are connected by $\varnothing$. By König's Lemma,[21] $\mathscr{G}$ contains an infinite sequence $\langle \Xi_0, \ldots, \Xi_n, \ldots \rangle$ s.t., for each $i \in \omega$, $\Xi_i \in \mathscr{G}_i$.

Let $\Gamma := \bigcup_{n \in \omega} \Xi_n$. By Lemma 2, all instances of the T-schema for well-founded sentences are still true in $\langle \mathbb{N}, \Gamma \rangle$ as, for each $n \in \omega$, $\Xi_n \cap \Psi_{\mathrm{wf}} = \Phi_{\mathrm{wf}}$. Let $\varphi_i \in \Delta$ and let $m > i$ be the least natural number s.t., for every $k \geq m$, Ξ_k doesn't contain elements of Θ_{φ_i} that are not already in Ξ_m – i.e. $\Xi_k \cap \Theta_{\varphi_i} = \Xi_m$. Such an m must exist, for Θ_{φ_i} is finite and sequences in $\mathscr{G}$ are monotonic. Since $i < m$ and $\Xi_m \in \mathscr{G}_m$, $\langle \mathbb{N}, \Xi_m \rangle \vDash \mathrm{T}^{\ulcorner}\varphi_i^{\urcorner} \leftrightarrow \varphi_i$. Given the choice of m, Lemma 2 entails that $\langle \mathbb{N}, \Gamma \rangle \vDash \mathrm{T}^{\ulcorner}\varphi_i^{\urcorner} \leftrightarrow \varphi_i$. □

Propositions 4 and 5 equip us with a more precise criterion for paradox. Whilst the former shows that infinitely many sentences are required for non-self-referential (ω-)paradox, by Proposition 5 we also know that at least some of the infinitely many paradoxical sentences must directly refer to infinitely many expressions.

Based on the notions of alethic reference and the results provided in this section, different reference patters may be employed to obtain a plethora of semantic truth theories, some more permissive, philosophically motivated, or elegant than others. Our restriction to well-founded sentences seems to fare quite well in these respects, although we have seen more permissive ones may also be adopted.

Unfortunately, though, reference and, *a fortiori*, well-foundedness and other reference patterns are far too complex to guide the development of axiomatic truth theories in a straightforward manner. For they are not definable in $\mathscr{L}$, as they make essential use of the notion of satisfaction in the standard model (cf. Definition 3). In the next section I provide mirroring notions of reference tailored specifically for that purpose; these are simplified counterparts of those introduced in the previous section.

4 Axiomatic Truth

The first part of this section is concerned with proof-theoretic versions of the notions presented in Section 2.3, that is, reference concepts that are relative to a particular proof-system. Section 4.2 explores a naïve truth theory formulated purely in terms of these proof-theoretic notions, shows it is unsound, and thus motivates the restriction of disquotation to reference stable expressions to avoid paradox. Finally, in Section 4.3 I formulate several axiomatic theories of disquotational truth deploying the concepts introduced in Sections 4.1 and 4.2. The resulting systems are shown to be sound and proof-theoretically strong in comparison with other systems that exists in the literature.

[21] König's Lemma establishes that every connected, locally finite, infinite graph contains an infinite path.

4.1 Proof-Theoretic Reference

Note that the set of sentences of $\mathscr{L}_{\mathrm{T}}$ is p.r., as is as the occurrence relation that holds between a string of symbols of the language and each of its substrings. Let $\mathrm{Sent}(x)$ and $\mathrm{Occ}(x, y) \in \mathscr{L}$ express this set and relation, respectively, in a natural way.[22] Thus, m-reference is also a p.r. relation (cf. Definition 1), as it can be expressed in $\mathscr{L}$ by the Δ_0^0 formula

$$\mathrm{MRef}(x, y) := \mathrm{Sent}(x \dot{\wedge} y) \wedge \exists t\, (\mathrm{Occ}(\dot{\mathrm{T}}t, x) \wedge t^{\circ} = y)$$

where $\dot{\mathrm{T}}$ is a function symbol of $\mathscr{L}$ for the p.r. function that maps each term t to the formula $\mathrm{T}t$ – and likewise for other predicates that will come up later.

Q-reference, instead, is much more complex, as it's not even arithmetical (cf. Definition 3), i.e. expressible in $\mathscr{L}$. As is well known, truth-in-$\mathbb{N}$ can be expressed by a Δ_1^1 formula, and this is best possible. Note that q-reference and truth-in-$\mathbb{N}$ are interdefinable over a small class of arithmetical notions. On the one hand, truth-in-$\mathbb{N}$ is the only notion occurring in the recursive definition of q-reference that involves second-order quantification, as the value and the occurrence relations, the normalization function, and syntactic properties of expressions such as being a disjunction or a universal statement are all p.r. and, thus, arithmetical. Therefore, q-reference is expressible by a Δ_1^1 formula. On the other hand, truth-in-$\mathbb{N}$ can be defined for each sentence $\varphi \in \mathscr{L}$ in terms of q-reference as follows: $\mathbb{N} \vDash \varphi$ iff $\forall x\, (x = \ulcorner \varphi \urcorner \wedge \varphi \rightarrow \mathrm{T}x)$ q-refers to φ. Thus, q-reference is Δ_1^1, namely, *hyperarithmetical*.

By the usual complexity considerations, this result extends to d-reference, reference, chains of reference, self-reference, and also well-foundedness (cf. Definitions 4 and 5–8), which means there is no formula in $\mathscr{L}$ that could be employed to restrict the instances of the T-schema or the Uniform T-schema in a first-order axiomatic system of disquotational truth. In other words, there is no straightforward way of axiomatizing the semantic theories introduced in Section 3.

Although there might be interesting, rather indirect, ways of axiomatizing this model, I opt instead to consider simpler, proof-theoretic versions of the concepts introduced in Section 2.3; in particular of well-foundedness. The strategy consists in replacing in Definition 3 the notion of truth-in-$\mathbb{N}$ with that of provability-in-PA, our base theory. Thus, all Δ_1^1 notions occurring in Definition 3 are replaced with Σ_1^0 ones. All further reference notions defined in terms of q-reference are modified accordingly. The new concepts can be seen as approximations to the original ones. They turn out to be expressible in $\mathscr{L}$ by formulae of fairly low quantificational complexity. This allows for the formulation of generous and well-motivated restrictions on disquotation, which result in powerful axiomatic truth theories, as will be seen soon.

Note that replacing in the definition of m-reference truth-in-$\mathbb{N}$ with provability-in-PA would not alter the extension of the relation, for all identity statements that

[22] That is, by means of Δ_0^0 formulae. Recall that every p.r. property (or relation) can be expressed in $\mathscr{L}$ by a Δ_0^0 expression and every recursively enumerable property by a Σ_1^0 expression. As PA decides all Δ_0^0 and proves all true Σ_1^0 sentences – i.e. PA is Σ_1^0-complete – those expressions represent and weakly represent, respectively, the properties they express.

are true in $\mathbb{N}$ are provable in PA and vice versa. As mentioned before, m-reference is already a p.r. notion. On the other hand, substituting provability-in-PA for truth-in-$\mathbb{N}$ in the definition of q-reference makes a significant difference: it results in a much simpler, recursively enumerable notion. As is well known, provability-in-PA is itself recursively enumerable. Thus, if we replace truth-in-$\mathbb{N}$ by provability-in-PA in the recursive definition of q-reference, it can be shown by induction on the depth of χ (cf. Section 2.3) that the altered notion is recursively enumerable as well, since both the definiendum and provability-in-PA occur only positively in the definiens, provability-in-PA is only preceded by a string of existential quantifiers, and all other notions occurring in the definition are p.r., as it has already been argued. Therefore, the resulting proof-theoretic notion of q-reference *relative to* PA can be expressed in $\mathscr{L}$ by a Σ_1^0 formula, $\mathrm{QRef}_{\mathsf{PA}}(x, y)$.

We can find arithmetical formulae expressing proof-theoretic versions of direct reference, reference *simpliciter*, self-reference, and well-foundedness relative to PA in terms of $\mathrm{MRef}(x, y)$ and $\mathrm{QRef}_{\mathsf{PA}}(x, y)$. For d-reference (cf. Definition 4), let $\mathrm{DRef}_{\mathsf{PA}}(x, y) := \mathrm{MRef}(x, y) \vee \mathrm{QRef}_{\mathsf{PA}}(x, y)$, which is trivially Σ_1^0.

For reference *simpliciter*, we first need to take a detour to define chain of reference. This requires some subtlety, for infinite sequences cannot be coded by natural numbers. However, finite sequences can. Moreover, note that, according to Definition 6, for φ to refer to ψ there must be a chain of reference starting with the former and ending with the latter, i.e. a *finite* chain. Thus, we can define reference in $\mathscr{L}$ exclusively in terms of finite chains. Let $\mathrm{Seq}(x) \in \Delta_0^0$ express in $\mathscr{L}$ the p.r. property of being a (finite) sequence, and let $\mathrm{len}(x)$ and $(x)_i \in \Delta_0^0$ express the p.r. function that maps each sequence to its length and the one that maps each sequence x and number i to the i-th entry of x, if i is smaller than the length of x, respectively. The Σ_1^0 formula

$$\mathrm{CRef}_{\mathsf{PA}}(x) := \mathrm{Seq}(x) \wedge \forall y < \mathrm{len}(x)\, \forall z < y\, \mathrm{DRef}_{\mathsf{PA}}((x)_z, (x)_{\mathrm{S}z})$$

expresses in $\mathscr{L}$ the property of being a finite chain of reference relative to PA (cf. Definition 5). Thus, reference relative to PA can be expressed in $\mathscr{L}$ by the Σ_1^0 formula

$$\mathrm{Ref}_{\mathsf{PA}}(x, y) := \exists z\, (\mathrm{CRef}_{\mathsf{PA}}(z) \wedge (z)_0 = x \wedge (z)_{\mathrm{len}(z)-1} = y)$$

Accordingly, self-reference relative to PA (cf. Definition 7) is expressed in $\mathscr{L}$ by $\mathrm{SRef}_{\mathsf{PA}}(x) := \mathrm{Ref}_{\mathsf{PA}}(x, x) \in \Sigma_1^0$. Well-foundedness, on the other hand, requires a bit more work, as it's defined in terms of possibly infinite chains of reference (cf. Definition 8). However, this is not a hindrance, for there is an equivalent formulation of the definition in terms of finite chains only. We say a finite chain of reference $\varphi_1, \ldots, \varphi_n$ can be *extended* just in case there is a sentence φ_{n+1} such that $\varphi_1, \ldots, \varphi_n, \varphi_{n+1}$ is also a chain of reference. Clearly, a sentence φ is well-founded just in case every finite chain of reference starting with φ can be extended only a finite number of times. Thus, well-foundedness relative to PA is expressed by the Π_3^0 formula

$$\begin{aligned}\mathrm{Wf}_{\mathsf{PA}}(x) := \forall y\, (\mathrm{CRef}_{\mathsf{PA}}(y) \wedge (y)_0 = x \rightarrow \\ \exists z \forall w\, (\mathrm{CRef}_{\mathsf{PA}}(w) \wedge \mathrm{len}(y) < \mathrm{len}(w) \wedge \forall k < \mathrm{len}(y)\, (y)_k = (w)_k \rightarrow \mathrm{len}(w) < z))\end{aligned}$$

This cumbersome formula states that for each finite chain of reference y starting with x there is a limit z to the length of every chain that extends y.

4.2 Unstable Reference

One would think our job is almost done now. It just remains to relativize disquotation to $\mathrm{Wf}_{\mathsf{PA}}(x)$ and then extend PAT with the resulting instances (see Section 2.1), that is,

$$\mathrm{Wf}_{\mathsf{PA}}(\ulcorner\varphi\urcorner) \rightarrow (\mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi) \tag{8}$$

for each sentence $\varphi \in \mathscr{L}_{\mathrm{T}}$, to obtain the desired truth system. Unfortunately, such a theory would be too weak and, what is worse, unsound.

Although many sentences turn out to be well-founded relative to PA, only very few are *provably* so in PA itself. Consider, for instance, the following:

$$\forall x\,(x = \ulcorner 0 = 0\urcorner \rightarrow \mathrm{T}x) \tag{9}$$

Despite being obviously harmless, as it only refers to $0 = 0$, PA cannot prove that (9) isn't self-referential, for that would mean that PA proves

$$\neg\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\forall x\,(x = \ulcorner 0 = 0\urcorner \rightarrow \mathrm{T}x)\urcorner = \ulcorner 0 = 0\urcorner)$$

In other words, PA would prove its own consistency, which is not possible by Gödel's second incompleteness theorem. In general, PA is not able to prove that a given sentence doesn't q-refer to an arbitrary expression, for it cannot prove that a given number or sequence of numbers provably fails to satisfy a given formula. Therefore, PA can only establish positive cases of q-reference plus those negative ones in which q-reference is nowhere restricted by a conditional – e.g. in $\forall x\,\mathrm{T}\dot{\neg}x$ – or no (non-dummy) quantifiers occur in the formula.

An idea that suggests itself is to inform PA that it doesn't prove false statements, i.e. that it is sound. This can be done by extending PA with all instances of PA*'s uniform reflection principle*, that is,

$$(\mathrm{URfn}_{\mathsf{PA}}) \qquad \forall t_1 \ldots t_n\,(\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\varphi(\dot{t}_1, \ldots, \dot{t}_n)\urcorner) \rightarrow \varphi(t_1^\circ, \ldots, t_n^\circ))$$

where $\varphi \in \mathscr{L}$. Let URfn(PA) be the resulting theory. URfn(PA) is trivially sound, as all instances of $\mathrm{URfn}_{\mathsf{PA}}$ are true in $\mathbb{N}$. Nonetheless, if we extend URfn(PA) with all instances of (8), we obtain a trivial system.

Observation 4 $\mathsf{URfn(PA)} \cup \{\mathrm{Wf}_{\mathsf{PA}}(\ulcorner\varphi\urcorner) \rightarrow (\mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi) : \varphi \in \mathscr{L}_{\mathrm{T}}\}$ *is inconsistent.*

Proof 'Weakly' diagonalizing (cf. Theorem 1) the predicate $\neg\mathrm{Bew}_{\mathsf{PA}}(x)$ we obtain a sentence γ that is provably equivalent in PA to $\neg\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\gamma\urcorner)$ – i.e. PA's Gödel sentence. In turn, Strong Diagonalization (cf. Theorem 2) applied to $\forall x\,(x = y \wedge \gamma \rightarrow \neg\mathrm{T}x)$ delivers a term l^* s.t.

$$\mathsf{PA} \vdash \mathrm{l}^* = \ulcorner\forall x\,(x = \mathrm{l}^* \wedge \gamma \rightarrow \neg\mathrm{T}x)\urcorner$$

Note that $\forall x\,(x = \mathrm{l}^* \wedge \gamma \rightarrow \neg \mathrm{T}x)$ doesn't refer to anything in PA, for PA doesn't prove $\mathrm{Bew}_{\mathsf{PA}}(\ulcorner \gamma \urcorner)$. Moreover, URfn(PA) knows this, as it famously implies γ and, therefore, that PA doesn't prove γ. As a consequence, $\forall x \neg \mathrm{Bew}_{\mathsf{PA}}(\ulcorner \dot{x} = \mathrm{l}^* \wedge \gamma \urcorner)$ is a theorem of URfn(PA), and so is $\mathrm{Wf}_{\mathsf{PA}}(\mathrm{l}^*)$. By an argument similar to that employed in the liar paradox, one can show that the corresponding instance of the T-schema for $\forall x\,(x = \mathrm{l}^* \wedge \gamma \rightarrow \neg \mathrm{T}x)$ entails a contradiction in URfn(PA). □

An immediate consequence of Observation 4 is that the theory that extends PAT with all instances of (8) is unsound. The reason behind this perplexing result is that being well-founded relative to PA does not mean that a sentence is actually well-founded, but rather that PA has no 'evidence' for its unfoundedness, i.e. it doesn't know that $\mathrm{l}^* = \mathrm{l}^* \wedge \gamma$. The proof of Observation 4 shows that, if more evidence is available, such as γ, a sentence might turn out to be unfounded relative to the extended system, URfn(PA). This is a consequence of the incompleteness of PA.

Relativizing q-reference and all the other reference concepts that depend on it to URfn(PA) or a stronger system instead of PA will obviously not solve the problem, as any recursively axiomatizable theory will also be incomplete. The issue would emerge all the same, only at a higher level. A better way of bypassing the obstacle is to focus on sentences for which new evidence does not make a difference in the expressions they refer to. I call these sentences "reference-stable" or "r-stable", for short.[23] R-*un*stable expressions bear a certain analogy with blind truth ascriptions and contingent liars: in both cases we don't know what they express and, a fortiori, if they are paradoxical or not. Only for r-stable sentences we can be sure that their reference patterns are safe.

To properly define this class of sentences, we first need to introduce the idea of a directly-reference-stable sentence, or "dr-stable" for short. These are sentences that only contain reference-restricting conditionals with 'simple' antecedents, that is, antecedents that, if satisfied by a sequence of numbers, PA knows so. Thus, these antecedents must be provably equivalent in PA to Σ^0_1 formulae.

Definition 9 (Dr-stability) Let φ be a sentence. φ is dr-stable iff in φ^* all subformulae of the form $\forall v\,(\chi_1 \vee \chi_2)$ where T occurs in χ_2 but not in χ_1 are s.t. χ_1 is provably equivalent in PA to a Π^0_1 formula.

For instance, whereas $\forall x\,(x = \mathrm{l}^* \rightarrow \neg \mathrm{T}x)$ and $\mathrm{GRfn}_{\mathsf{PA}}$ – i.e. $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$ – are both dr-stable, $\forall x\,(\neg \mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$ and $\forall x\,(x = \mathrm{l}^* \wedge \gamma \rightarrow \neg \mathrm{T}x)$ aren't. Since truth-in-$\mathbb{N}$ and provability-in-PA coincide for sentences provably equivalent in

[23] A complementary strategy would be to replace PA as our base theory with an extension of it by means of a hierarchy of iterated reflection principles structurally identical to $\mathrm{URfn}_{\mathsf{PA}}$ into the transfinite. This would increase the number of expressions that are rightly declared well-founded and, consequently, the number of instances of disquotation that hold in the ultimate truth system. However, this would introduce unnecessary complications, as we will see that URfn(PA) can already serve as base theory of considerably powerful truth systems.

PA to a Σ^0_1 statement, q-reference in PA and q-reference *simpliciter* coincide in the case of dr-stable sentences. Thus, so do d-reference in PA and d-reference *simpliciter*, so we can drop the qualifier in these cases. The same cannot be said of reference itself. For a dr-stable sentence φ might refer to an expression ψ that is not itself dr-stable, so that the latter d-refers to a sentence χ but doesn't d-refer to it in PA. Our original sentence φ, then, also refers to χ but not in PA. Thus, we need the following definition.

Definition 10 (R-stability) Let φ be a sentence. φ is r-stable iff it's dr-stable and only refers to dr-stable sentences.

Since $\mathrm{GRfn}_{\mathsf{PA}}$ only refers to T-free sentences, it must be r-stable. On the other hand, $\forall x\,(x = \mathrm{l}^* \rightarrow \neg \mathrm{T}x)$ refers to $\forall x\,(x = \mathrm{l}^* \wedge \gamma \rightarrow \neg \mathrm{T}x)$ and, therefore, isn't r-stable. Since r-stable sentences are dr-stable and only refer to dr-stable sentences, it follows from our previous considerations that Definition 10 can be equivalently restated by replacing "refers" with "refers in PA". Thus, reference in PA and reference *simpliciter* concur in the case of r-stable sentences. Moreover, note that dr-stability is expressible in $\mathscr{L}$ by a Σ^0_1 formula, $\mathrm{DRSt}(x)$, as it's defined in terms of p.r. syntactic properties plus the property of being provably equivalent in PA to a Π^0_1 formula. Therefore, by the usual complexity considerations, r-stability is expressible in $\mathscr{L}$ by the following Π_2 formula:

$$\mathrm{RSt}(x) := \mathrm{DRSt}(x) \wedge \forall y\,(\mathrm{Ref}_{\mathsf{PA}}(x, y) \rightarrow \mathrm{DRSt}(y))$$

4.3 Restricting Disquotation

We now have all the resources needed to formulate sound axiomatic theories of disquotational truth. Let URfn(PAT) be PAT+ ($\mathrm{URfn}_{\mathsf{PA}}$) and let WFUTB be the theory extending URfn(PAT) with the following restricted version of the Uniform T-schema, where $\mathbf{t}$ abbreviates $t_1, \ldots, t_n$:

$$\forall \mathbf{t}\,(\mathrm{RSt}(\ulcorner \varphi(\dot{\mathbf{t}}) \urcorner) \wedge \mathrm{Wf}_{\mathsf{PA}}(\ulcorner \varphi(\dot{\mathbf{t}}) \urcorner) \rightarrow (\mathrm{T}\ulcorner \varphi(\dot{\mathbf{t}}) \urcorner \leftrightarrow \varphi(\mathbf{t}^\circ)))$$

WFUTB – for "Well-founded Uniform Tarski Biconditionals" – uniformly entails all instances of the T-schema for sentences that are provably r-stable and well-founded in PA.

Proposition 6 WFUTB *is* ω*-consistent.*

Proof Since every r-stable sentence that is well-founded in PA is also well-founded *simpliciter*, by Proposition 3 the axioms of WFUTB are all true in $\langle \mathbb{N}, \Phi_{\mathrm{wf}} \rangle$. □

Despite being a disquotational theory, WFUTB is proof-theoretically strong, as it can relatively interpret the theory of ramified truth up to Γ_0, $\mathsf{RT}_{<\Gamma_0}$. This consists of iterations of the theory of typed compositional truth over PAT up to the ordinal Γ_0, the

Feferman-Schütte ordinal.[24] Each of these iterations requires its own truth predicate, bringing about a hierarchy of corresponding languages.

As is well known, natural numbers can effectively codify ordinals up to Γ_0.[25] Assuming a fixed (effective) coding, if $\alpha < \Gamma_0$, we write $\overline{\alpha}$ for the numeral of the code of α. For clarity, I often identify ordinals with their codes if there is no room for confusion. Based on this coding, PA can talk about ordinals and represent some of their properties. For instance, the set of ordinals below Γ_0 can be represented by a formula of $\mathscr{L}$, $\mathrm{Ord}(x)$, as well as their ordering, say, by the formula $x \prec y$. Let $\forall\alpha\,\varphi$ and $\exists\alpha\,\varphi$ be short for $\forall v\,(\mathrm{Ord}(v) \to \varphi)$ and $\exists v\,(\mathrm{Ord}(v) \wedge \varphi)$, respectively, where v is a suitable variable. As is also well known, PA can prove all instances of transfinite induction up to ϵ_0 ($< \Gamma_0$), i.e.

$$(\mathrm{TI}_\xi) \qquad \forall\alpha\,(\forall\beta \prec \alpha\,\varphi(\beta) \to \varphi(\alpha)) \to \forall\alpha \prec \overline{\xi}\,\varphi(\alpha)$$

for $\varphi \in \mathscr{L}$ and $\xi < \epsilon_0$.[26] This means PA can prove that all ordinals below ϵ_0 are well ordered.

Let $\mathscr{L}_{<0}$ be $\mathscr{L}$ and, for every α such that $0 < \alpha \leq \Gamma_0$, let $\mathscr{L}_{<\alpha}$ be the result of extending $\mathscr{L}$ with monadic predicates T_β, for each $\beta < \alpha$. $\mathsf{PAT}_{<\alpha}$ consists of the axioms of PA formulated in $\mathscr{L}_{<\alpha}$, with the induction schema extended to the whole language. For each $\alpha < \Gamma_0$, let $\mathrm{Sent}(x, \alpha) \in \mathscr{L}$ – or $\mathrm{Sent}_\alpha(x)$, for short – represent the set of sentences of $\mathscr{L}_{<\alpha}$. We define a cumulative hierarchy of compositional truth theories formulated in these languages, as follows.

Let $\mathsf{RT}_{<0}$ be PA and, for every α such that $0 < \alpha \leq \Gamma_0$, let $\mathsf{RT}_{<\alpha}$ be the theory formulated in $\mathscr{L}_{<\alpha}$ that extends $\mathsf{PAT}_{<\alpha}$ with every instance of following axiom-schemata, for each $\xi < \beta < \alpha$:

$$(\mathrm{RT}_\beta 1) \qquad \forall s\forall t\,(\mathrm{T}_\beta(s \doteq t) \leftrightarrow s^\circ = t^\circ)$$

$$(\mathrm{RT}_\beta 2) \qquad \forall x\,(\mathrm{Sent}_{\overline{\beta}}(x) \to (\mathrm{T}_\beta \dot{\neg} x \leftrightarrow \neg \mathrm{T}_\beta x))$$

$$(\mathrm{RT}_\beta 3) \qquad \forall x\forall y\,(\mathrm{Sent}_{\overline{\beta}}(x \mathbin{\dot{\wedge}} y) \to (\mathrm{T}_\beta(x \mathbin{\dot{\wedge}} y) \leftrightarrow (\mathrm{T}_\beta x \wedge \mathrm{T}_\beta y)))$$

$$(\mathrm{RT}_\beta 4) \qquad \forall x\forall y\,(\mathrm{Sent}_{\overline{\beta}}(x \mathbin{\dot{\vee}} y) \to (\mathrm{T}_\beta(x \mathbin{\dot{\vee}} y) \leftrightarrow (\mathrm{T}_\beta x \vee \mathrm{T}_\beta y)))$$

$$(\mathrm{RT}_\beta 5) \qquad \forall x\forall y\,(\mathrm{Sent}_{\overline{\beta}}(x \mathbin{\dot{\to}} y) \to (\mathrm{T}_\beta(x \mathbin{\dot{\to}} y) \leftrightarrow (\mathrm{T}_\beta x \to \mathrm{T}_\beta y)))$$

$$(\mathrm{RT}_\beta 6) \qquad \forall x\forall y\,(\mathrm{Sent}_{\overline{\beta}}(\dot{\forall} y\, x) \to (\mathrm{T}_\beta(\dot{\forall} y\, x) \leftrightarrow \forall t\,\mathrm{T}_\beta x(t/y)))$$

$$(\mathrm{RT}_\beta 7) \qquad \forall x\forall y\,(\mathrm{Sent}_{\overline{\beta}}(\dot{\exists} y\, x) \to (\mathrm{T}_\beta(\dot{\exists} y\, x) \leftrightarrow \exists t\,\mathrm{T}_\beta x(t/y)))$$

$$(\mathrm{RT}_\beta 8) \qquad \forall t\,(\mathrm{Sent}_{\overline{\xi}}(t^\circ) \to (\mathrm{T}_\beta(\dot{\mathrm{T}}_\xi t) \leftrightarrow \mathrm{T}_\xi t^\circ))$$

$$(\mathrm{RT}_\beta 9) \qquad \forall t\forall\zeta \prec \overline{\beta}\,(\mathrm{Sent}_\zeta(t^\circ) \to (\mathrm{T}_\beta(\dot{\mathrm{T}}_\zeta t) \leftrightarrow \mathrm{T}_\beta t^\circ))$$

where $\dot{\forall}$ and $\dot{\exists}$ are function symbols for the p.r. functions mapping a variable v and a formula φ to $\forall v\,\varphi$ and $\exists v\,\varphi$, respectively. Note that the axioms never quantify over the subindex β of T_β, on pain of triviality. But they do quantify over the predicates themselves up to a given ordinal β in $\mathrm{RT}_\beta 9$. This set is p.r. for each $\beta < \epsilon_0$ and

[24] See Halbach [10, chap. 8].

[25] More comprehensive ordinal notation systems are also possible, but this is enough for our purposes.

[26] See Pohlers [28, chap. 3].

representable also for $\beta \geq \epsilon_0$ insofar as all arithmetical instances of TI_β hold in the theory.

$\mathsf{RT}_{<1}$ is just the theory of typed compositional truth CT where all occurrences of T have been replaced with T_0 (also within corner quotes), for $\mathrm{RT}_0 8$ and $\mathrm{RT}_0 9$ are vacuously true. Similarly, for $1 < \alpha < \Gamma_0$, axioms $\mathrm{RT}_\beta 1$, with $\beta < \alpha$, establish all instances of disquotation for identity statements for T_β, and $\mathrm{RT}_\beta 2$-$\mathrm{RT}_\beta 7$ lay down the compositional character of these truth predicates. $\mathrm{RT}_\beta 8$, in turn, provides instances of disquotation for each truth ascription that belongs to lower levels of the hierarchy, supplementing $\mathrm{RT}_\beta 1$. Finally, $\mathrm{RT}_\beta 9$ establishes that the hierarchy is cumulative.

Proposition 7 *The theory of ramified truth up to* Γ_0, $\mathsf{RT}_{<\Gamma_0}$, *is relatively interpretable in* WFUTB.

Proof I first show that $\mathscr{L}_\mathrm{T}$ contains a formula $\vartheta_{\overline{\beta}}(x)$ that behaves in WFUTB like $\mathrm{T}_\beta(x)$ in $\mathsf{RT}_{<\epsilon_0}$, for each $\beta < \epsilon_0$, following Halbach's [10] demonstration of his Lemma 15.24. I then extend this result to every $\beta < \Gamma_0$.

If $\varphi(x)$ is a formula of $\mathscr{L}_\mathrm{T}$ containing T, let $\mathscr{L}_\varphi \subseteq \mathscr{L}_\mathrm{T}$ be the language that extends $\mathscr{L}$ with $\varphi(x)$ as if it were just a predicate symbol, that is, T occurs in formulae of $\mathscr{L}_\varphi$ only within $\varphi(t)$, where t is a term.[27] The relation that holds between a sentence ψ and a formula φ just in case ψ is a sentence of $\mathscr{L}_\varphi$ is p.r. Therefore, it can be represented by a Δ_0^0 expression, $\mathrm{Sent}\restriction(x, y)$. Strongly diagonalizing the formula

$$
\begin{aligned}
&\exists s \exists t\, (x = (s \doteq t) \wedge s^\circ = t^\circ) \;\vee\\
&\exists z\, ((\mathrm{Sent}_{\overline{0}}(z) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(z, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = \dot{\neg} z \wedge \neg \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{z}/\ulcorner x \urcorner)) \;\vee\\
&\exists z \exists w\, ((\mathrm{Sent}_{\overline{0}}(z \dot{\wedge} w) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(z \dot{\wedge} w, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = (z \dot{\wedge} w) \;\wedge\\
&\qquad (\mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{z}/\ulcorner x \urcorner) \wedge \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{w}/\ulcorner x \urcorner))) \;\vee\\
&\exists z \exists w\, ((\mathrm{Sent}_{\overline{0}}(z \dot{\wedge} w) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(z \dot{\wedge} w, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = (z \dot{\vee} w) \;\wedge\\
&\qquad (\mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{z}/\ulcorner x \urcorner) \vee \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{w}/\ulcorner x \urcorner))) \;\vee\\
&\exists z \exists w\, ((\mathrm{Sent}_{\overline{0}}(z \dot{\rightarrow} w) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(z \dot{\rightarrow} w, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = (z \dot{\rightarrow} w) \;\wedge\\
&\qquad (\mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{z}/\ulcorner x \urcorner) \rightarrow \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\dot{w}/\ulcorner x \urcorner))) \;\vee\\
&\exists z \exists w\, ((\mathrm{Sent}_{\overline{0}}(\dot{\forall} z\, w) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(\dot{\forall} z\, w, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = \dot{\forall} z\, w \;\wedge\\
&\qquad \forall t\, \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\ulcorner \dot{w}(\dot{t}/\dot{z}) \urcorner/\ulcorner x \urcorner)) \;\vee\\
&\exists z \exists w\, ((\mathrm{Sent}_{\overline{0}}(\dot{\exists} z\, w) \vee \exists \zeta \prec y\, \mathrm{Sent}\restriction(\dot{\exists} z\, w, k(\dot{\zeta}/\ulcorner y \urcorner))) \wedge x = \dot{\exists} z\, w \;\wedge\\
&\qquad \exists t\, \mathrm{T}k(\dot{y}/\ulcorner y \urcorner)(\ulcorner \dot{w}(\dot{t}/\dot{z}) \urcorner/\ulcorner x \urcorner)) \;\vee\\
&\exists t \exists \zeta \prec y\, ((\mathrm{Sent}_{\overline{0}}(t^\circ) \vee \exists \xi \prec \zeta\, \mathrm{Sent}\restriction(t^\circ, k(\dot{\xi}/\ulcorner y \urcorner))) \wedge x = k(\dot{\zeta}/\ulcorner y \urcorner)(t/\ulcorner x \urcorner) \;\wedge\\
&\qquad \mathrm{T}k(\dot{\zeta}/\ulcorner y \urcorner)(t/\ulcorner x \urcorner))
\end{aligned}
$$

[27] More precisely, we add the following recursion clause to the definition of well-formed formula: if t is a term, then $\varphi(t)$ is a well-formed formula.

over the free variable k, we obtain a predicate $\vartheta(x, y)$ – $\vartheta_y(x)$ – s.t.[28]

$$\begin{aligned}
\mathsf{PAT} \vdash \vartheta_y(x) \leftrightarrow \quad & \exists s \exists t\, (x = (s \doteq t) \wedge s^\circ = t^\circ) \ \vee \\
& \exists z\, (\mathrm{Sent}_{<y}(z) \wedge x = \dot{\neg} z \wedge \neg \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{z})^\urcorner) \ \vee \\
& \exists z \exists w\, (\mathrm{Sent}_{<y}(z \mathbin{\dot{\wedge}} w) \wedge x = (z \mathbin{\dot{\wedge}} w) \wedge (\mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{z})^\urcorner \wedge \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{w})^\urcorner)) \ \vee \\
& \exists z \exists w\, (\mathrm{Sent}_{<y}(z \mathbin{\dot{\wedge}} w) \wedge x = (z \mathbin{\dot{\vee}} w) \wedge (\mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{z})^\urcorner \vee \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{w})^\urcorner)) \ \vee \\
& \exists z \exists w\, (\mathrm{Sent}_{<y}(z \mathbin{\dot{\rightarrow}} w) \wedge x = (z \mathbin{\dot{\rightarrow}} w) \wedge (\mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{z})^\urcorner \rightarrow \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{w})^\urcorner)) \ \vee \\
& \exists z \exists w\, (\mathrm{Sent}_{<y}(\dot{\forall} z\, w) \wedge x = \dot{\forall} z\, w \wedge \forall t\, \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{w}(\dot{t}/\dot{z}))^\urcorner) \ \vee \\
& \exists z \exists w\, (\mathrm{Sent}_{<y}(\dot{\exists} z\, w) \wedge x = \dot{\exists} z\, w \wedge \exists t\, \mathrm{T}^\ulcorner \vartheta_{\dot{y}}(\dot{w}(\dot{t}/\dot{z}))^\urcorner) \ \vee \\
& \exists t \exists \zeta \prec y\, (\mathrm{Sent}_{<\zeta}(t^\circ) \wedge x = {}^\ulcorner \vartheta_{\dot{\zeta}}(\dot{t})^\urcorner \wedge \mathrm{T}^\ulcorner \vartheta_{\dot{\zeta}}(\dot{t})^\urcorner)
\end{aligned} \tag{10}$$

where $\mathrm{Sent}_{<s}(t) := \mathrm{Sent}_{\overline{0}}(t) \vee \exists \zeta \prec s\, \mathrm{Sent}\restriction(t, {}^\ulcorner \vartheta_{\dot{\zeta}}(x)^\urcorner) \in \Delta^0_0$, s and t are terms and ζ is a suitable ordinal variable.[29] Let $\mathscr{L}_{<\vartheta_\alpha}$ be the language that results from merging together every $\mathscr{L}_{\vartheta_\beta}$, with $\beta < \alpha$. Thus, for every $\alpha < \epsilon_0$, $\mathrm{Sent}_{<\alpha}(x)$ represents the set of sentences of $\mathscr{L}_{<\vartheta_\alpha}$.

Given that (10) was obtained by Strong Diagonalization, we can *identify* the predicate $\vartheta_y(x)$ with the right-hand side of the biconditional, except all occurrences of ${}^\ulcorner \vartheta_y(x)^\urcorner$ are actually occurrences of a more complex term that denotes $\vartheta_y(x)$. Since alethic reference is closed under Leibniz's Law, this difference won't actually make a difference.

To show that, for each $\beta < \epsilon_0$, $\vartheta_{\overline{\beta}}(x)$ behaves like $\mathrm{T}_\beta(x)$, we first need to remove the occurrences of T in (10), that is, we need to show that WFUTB contains an instance of the Uniform T-schema for each $\vartheta_\beta(x)$, with $\beta < \epsilon_0$. In other words, I give a uniform proof that every instance of each of these formulae is r-stable and well-founded in PA.

Note first that the normalization of $\vartheta_y(x)$ is a disjunction of negated universally quantified statements followed by just one reference-restricting conditional, except for the first disjunct that doesn't contain T. Moreover, the antecedents of these conditionals are all Δ^0_0. This means, on the one hand, that each $\vartheta_{\overline{\beta}}(t)$ is dr-stable, so

$$\mathsf{PA} \vdash \forall \beta \forall t\, \mathrm{DRSt}({}^\ulcorner \vartheta_{\dot{\beta}}(\dot{t})^\urcorner) \tag{11}$$

that is, if $\vartheta_{\overline{\beta}}(t)$ q-refers to a sentence in PA, then PA knows about it. But also, the fact that all the antecedents are Δ^0_0 implies that URfn(PA) knows about all *negative* cases of q-reference for each $\vartheta_{\overline{\beta}}(t)$ as well. For instance, if $\mathrm{Sent}_{<\overline{\beta}}(z \mathbin{\dot{\wedge}} w) \wedge t = (z \mathbin{\dot{\wedge}} w)$ isn't true of $m, n \in \omega$, then PA proves the negation of $\mathrm{Sent}_{<\overline{\beta}}(\overline{m} \mathbin{\dot{\wedge}} \overline{n}) \wedge t = (\overline{m} \mathbin{\dot{\wedge}} \overline{n})$, which means that URfn(PA) proves that PA doesn't prove $\mathrm{Sent}_{<\overline{\beta}}(\overline{m} \mathbin{\dot{\wedge}} \overline{n}) \wedge t = (\overline{m} \mathbin{\dot{\wedge}} \overline{n})$.

[28] Note that the fourth disjunct on the right-hand side of the following biconditional requires that the conjunction of z and w is a sentence of $\mathscr{L}_{<y}$, instead of their disjunction. This guarantees that z and w are sentences of the language themselves, and not two disjuncts conforming an expression of the form $\vartheta_k(t)$, with $k < y$, which is always a disjunction. This way, only one disjunct in $\vartheta_y(x)$ can be satisfied at a time.

[29] Although strictly speaking this formula is not Δ^0_0, as it quantifies over the possibly infinitely many ordinals smaller than ζ, it's easy to see that only finitely many cases need to be checked for each t and each s, for every sentence of $\mathscr{L}_\mathrm{T}$ is of finite length. Thus, $\mathrm{Sent}_{<s}(t)$ is provably equivalent in PA to a Δ^0_0 expression.

Furthermore, since each of the antecedents holds exactly of one natural number or ordered pair of natural numbers, URfn(PA) can also prove general facts about the sentences each $\vartheta_{\overline{\beta}}(t)$ q-refers to. For example, it follows in URfn(PA) that, if t denotes an identity statement or doesn't denote a sentence of $\mathscr{L}_{<\vartheta_\beta}$, then $\vartheta_{\overline{\beta}}(t)$ doesn't refer to any expression.

By (10) it also follows easily in URfn(PA) that each $\vartheta_{\overline{\beta}}(t)$ d-refers just to sentences of the form $\vartheta_{\overline{\zeta}}(s)$, where $\zeta \prec \beta$ or s denotes either a sentence in $\mathscr{L}_{<\vartheta_\beta}$ of lower complexity than t's or a sentence in $\mathscr{L}_{<\vartheta_\zeta}$, for some $\zeta \prec \beta$. Thus, if $\alpha < \epsilon_0$, we can prove that

$$\mathsf{URfn(PA)} \vdash \forall \beta \prec \overline{\alpha} \forall t \, (\mathrm{RSt}(\ulcorner \vartheta_{\dot{\beta}}(\dot{t}) \urcorner) \wedge \mathrm{Wf}_{\mathsf{PA}}(\ulcorner \vartheta_{\dot{\beta}}(\dot{t}) \urcorner)) \tag{12}$$

by internal transfinite induction up to α on β. We reason informally in URfn(PA). Assume $\forall \zeta \prec \beta \forall t \, (\mathrm{RSt}(\ulcorner \vartheta_{\dot{\zeta}}(\dot{t}) \urcorner) \wedge \mathrm{Wf}_{\mathsf{PA}}(\ulcorner \vartheta_{\dot{\zeta}}(\dot{t}) \urcorner))$. We derive $\forall t \, (\mathrm{RSt}(\ulcorner \vartheta_{\dot{\beta}}(\dot{t}) \urcorner) \wedge \mathrm{Wf}_{\mathsf{PA}}(\ulcorner \vartheta_{\dot{\beta}}(\dot{t}) \urcorner))$ by an internal complete induction on the complexity of the sentence denoted by t°.

If t° doesn't denote a sentence of $\mathscr{L}_{<\vartheta_\beta}$ or denotes an identity statement, the result follows trivially, given (11). Assume $\mathrm{RSt}(\ulcorner \vartheta_{\dot{\beta}}(\dot{s}) \urcorner) \wedge \mathrm{Wf}_{\mathsf{PA}}(\ulcorner \vartheta_{\dot{\beta}}(\dot{s}) \urcorner)$ for every s s.t. s° denotes a sentence of lower complexity than t°'s. If t° denotes a negation $\neg\varphi \in \mathscr{L}_{<\vartheta_\beta}$, by (10) $\vartheta_\beta(t^\circ)$ d-refers to $\vartheta_\beta(\ulcorner \varphi \urcorner)$ that, by i.h., is both r-stable and well-founded in PA. By (11), so is $\vartheta_\beta(t^\circ)$. The cases for the other connectives and the quantifiers are similar.[30] Let t° denote $\vartheta_\zeta(s)$, for some $\zeta \prec \beta$. By the last line of (10), $\vartheta_\beta(t^\circ)$ d-refers to $\vartheta_\zeta(s)$ which, by our main i.h. and (11), is r-stable and well-founded in PA. This completes the proof of (12).

As a consequence, for every $\alpha < \epsilon_0$ the following principle of disquotation holds in WFUTB:

$$\forall \beta \prec \bar{\alpha} \forall t \, (\mathrm{T}^\ulcorner \vartheta_{\dot{\beta}}(\dot{t}) \urcorner \leftrightarrow \vartheta_\beta(t^\circ)) \tag{13}$$

We are now in a position to show that $\mathsf{RT}_{<\epsilon_0}$ is relatively interpretable in WFUTB.

Consider the following translation function $\eta : \mathscr{L}_{<\Gamma_0} \to \mathscr{L}_{\mathrm{T}}$:

$$\eta(\varphi) := \begin{cases} \varphi & \text{if } \varphi := s = t \\ \neg\eta(\psi) & \text{if } \varphi := \neg\psi \\ \eta(\psi) \wedge \eta(\chi) & \text{if } \varphi := (\psi \wedge \chi) \\ \eta(\psi) \vee \eta(\chi) & \text{if } \varphi := (\psi \vee \chi) \\ \eta(\psi) \to \eta(\chi) & \text{if } \varphi := (\psi \to \chi) \\ \forall v \, \eta(\psi) & \text{if } \varphi := \forall v \, \psi \\ \exists v \, \eta(\psi) & \text{if } \varphi := \exists v \, \psi \\ \vartheta_{\overline{\alpha}}(\dot{\eta}(t)) & \text{if } \varphi := \mathrm{T}_\alpha t \end{cases}$$

where $\dot{\eta}$ is a term of $\mathscr{L}$ representing η. We know such a function exists and is p.r. by Kleene's Recursion Theorem. This way, η not only translates truth predicates occurring in a formula but also those that occur inside corner quotes, as well as those inside corner quotes inside corner quotes, etc. For instance, the translation of $\mathrm{T}_{\overline{\omega}}(\ulcorner \mathrm{T}_0(\ulcorner 0 = 0 \urcorner) \urcorner)$ is not $\vartheta_{\overline{\omega}}(\ulcorner \mathrm{T}_0(\ulcorner 0 = 0 \urcorner) \urcorner)$ but $\vartheta_{\overline{\omega}}(\ulcorner \vartheta_{\overline{0}}(\ulcorner 0 = 0 \urcorner) \urcorner)$.

[30] A word of caution about the disjunction case is in order, however. Note that, if t° denotes a disjunction, not only the restricting antecedent on the fourth line of (10) might be true, but also the one on the last line, since each formula of the form $\vartheta_\zeta(s)$ is itself a disjunction.

In what follows, I show that WFUTB proves the translation of every axiom of $\mathsf{RT}_{<\epsilon_0}$. Actually, I prove *general*, *quantified* versions of the translations to facilitate the derivation of $\mathrm{RT}_\beta 9$. I only deal with $\mathrm{RT}_\beta 1$, $\mathrm{RT}_\beta 6$, $\mathrm{RT}_\beta 8$, and $\mathrm{RT}_\beta 9$. The other cases can be proved in a similar fashion. Let $\alpha < \epsilon_0$. We reason informally in WFUTB under the assumption that $\beta \prec \overline{\alpha}$. By (10) we have that

$$\forall \beta \prec \overline{\alpha}\, (\vartheta_\beta(s \doteq t) \leftrightarrow s^\circ = t^\circ)$$

which entails the translation of $\mathrm{RT}_\beta 1$. For the other axioms, note that

$$\mathsf{PA} \vdash \forall \beta \prec \overline{\alpha}\, \forall x\, (\mathrm{Sent}_\beta(x) \leftrightarrow \mathrm{Sent}_{<\beta}(\dot{\eta}(x))) \tag{14}$$

by TI_α. Thus, we have that

$$\begin{aligned}
\mathrm{Sent}_\beta(\dot{\forall} y\, x) &\rightarrow \mathrm{Sent}_{<\beta}(\dot{\eta}(\dot{\forall} y\, x)) && (14)\\
&\rightarrow \mathrm{Sent}_{<\beta}(\dot{\forall} y\, \dot{\eta}(x)) \text{Def. } \eta\\
&\rightarrow (\vartheta_\beta(\dot{\forall} y\, \dot{\eta}(x)) \leftrightarrow \forall t\, \mathrm{T}^\ulcorner \vartheta_{\dot{\beta}}(\dot{\eta}(\dot{x})(\dot{t}/\dot{y}))^\urcorner) && (10)\\
&\rightarrow (\vartheta_\beta(\dot{\forall} y\, \dot{\eta}(x)) \leftrightarrow \forall t\, \vartheta_\beta(\dot{\eta}(x)(t/y))) && (13)
\end{aligned}$$

which implies the translation of $\mathrm{RT}_\beta 6$. For $\mathrm{RT}_\beta 8$'s, assume $\xi \prec \beta$:

$$\begin{aligned}
\mathrm{Sent}_\xi(t^\circ) &\rightarrow \mathrm{Sent}_{<\xi}(\dot{\eta}(t^\circ)) && (14)\\
&\rightarrow (\vartheta_\beta(^\ulcorner \vartheta_{\dot{\xi}}(\dot{\eta}(\dot{t}))^\urcorner) \leftrightarrow \mathrm{T}^\ulcorner \vartheta_{\dot{\xi}}(\dot{\eta}(\dot{t}))^\urcorner) && (10)\\
&\rightarrow (\vartheta_\beta(^\ulcorner \vartheta_{\dot{\xi}}(\dot{\eta}(\dot{t}))^\urcorner) \leftrightarrow \vartheta_\xi(\dot{\eta}(t^\circ))) && (13)
\end{aligned}$$

Finally, we prove the translation of $\mathrm{RT}_\beta 9$, that is,[31]

$$\forall t \forall \zeta \prec \overline{\beta}\, (\mathrm{Sent}_\zeta(t^\circ) \rightarrow (\vartheta_{\overline{\beta}}(^\ulcorner \vartheta_{\dot{\zeta}}(\dot{\eta}(\dot{t}))^\urcorner) \leftrightarrow \vartheta_{\overline{\beta}}(\dot{\eta}(t^\circ))))$$

Assume $\zeta \prec \beta$ and $\mathrm{Sent}_\zeta(t^\circ)$. By (14), we have that $\mathrm{Sent}_{<\zeta}(\dot{\eta}(t^\circ))$. By our proof of the translation of $\mathrm{RT}_\beta 8$, in which ξ is a *variable*, we know that $\vartheta_\beta(^\ulcorner \vartheta_{\dot{\zeta}}(\dot{\eta}(\dot{t}))^\urcorner) \leftrightarrow \vartheta_\zeta(\dot{\eta}(t^\circ))$. Thus, we just need to show that $\vartheta_\zeta(\dot{\eta}(t^\circ)) \leftrightarrow \vartheta_\beta(\dot{\eta}(t^\circ))$, for every term t. I prove it by an internal complete induction on the complexity of t°.

If $t^\circ = (t_1 \doteq t_2)$, we have that

$$\begin{aligned}
\vartheta_\zeta(\dot{\eta}(t^\circ)) &\leftrightarrow \vartheta_\zeta(t_1 \doteq t_2) && \text{Def. } \eta\\
&\leftrightarrow t_1^\circ = t_2^\circ && \text{Proof of } \mathrm{RT}_\beta 1\\
&\leftrightarrow \vartheta_\beta(t_1 \doteq t_2) && \text{Proof of } \mathrm{RT}_\beta 1\\
&\leftrightarrow \vartheta_\beta(\dot{\eta}(t^\circ)) && \text{Def. } \eta
\end{aligned}$$

[31] Note that $\mathrm{RT}_\beta 9$ does not have a 'corresponding' disjunct in (10), as it would clash with that for $\mathrm{RT}_\beta 8$ in the sense that there would be two simultaneously satisfiable disjuncts, blocking the proof of $\mathrm{RT}_\beta 8$.

which completes the base case. Assume the result holds for every term denoting a sentence of lower complexity than t°'s. If $t^\circ = \dot{\neg} s$, we have the following:

$$\begin{aligned}
\vartheta_\zeta(\dot{\eta}(t^\circ)) &\leftrightarrow \vartheta_\zeta(\dot{\neg}\dot{\eta}(s)) && \text{Def. } \eta\\
&\leftrightarrow \neg\vartheta_\zeta(\dot{\eta}(s)) && \text{Proof of RT}_\beta 2\\
&\leftrightarrow \neg\vartheta_\beta(\dot{\eta}(s)) && \text{Ind. hyp.}\\
&\leftrightarrow \vartheta_\beta(\dot{\neg}\dot{\eta}(s)) && \text{Proof of RT}_\beta 2\\
&\leftrightarrow \vartheta_\beta(\dot{\eta}(t^\circ)) && \text{Def. } \eta
\end{aligned}$$

The cases for the other logical operators follow from the i.h. in a similar way. For the last case, assume $t^\circ = \ulcorner \mathrm{T}_\xi(s) \urcorner$, for some $\xi \prec \zeta$. We then have that

$$\begin{aligned}
\vartheta_\zeta(\dot{\eta}(t^\circ)) &\leftrightarrow \vartheta_\zeta(\ulcorner \vartheta_{\dot{\xi}}(\dot{\eta}(\dot{s}))\urcorner) && \text{Def. } \eta\\
&\leftrightarrow \vartheta_\xi(\dot{\eta}(s)) && \text{Proof of RT}_\beta 8\\
&\leftrightarrow \vartheta_\beta(\ulcorner \vartheta_{\dot{\xi}}(\dot{\eta}(\dot{s}))\urcorner) && \text{Proof of RT}_\beta 8\\
&\leftrightarrow \vartheta_\beta(\dot{\eta}(t^\circ)) && \text{Def. } \eta
\end{aligned}$$

This completes the proof of the relative interpretability of $\mathsf{RT}_{<\epsilon_0}$ in WFUTB. Therefore, WFUTB entails every arithmetical theorem of $\mathsf{RT}_{<\epsilon_0}$. This includes all instances of transfinite induction for formulae of $\mathscr{L}$ up to α, for some $\alpha > \epsilon_0$, i.e. WFUTB can prove that larger segments of the ordinal notation system are well-ordered. This implies that our proofs can be extended to show that $\mathsf{RT}_{<\alpha}$ is relatively interpreted in WFUTB, as only arithmetical transfinite induction has been employed so far. In turn, $\mathsf{RT}_{<\alpha}$ allows us to prove further instances of arithmetical transfinite induction, which means WFUTB can actually relatively interpret more iterations of compositionality, and so on. The progression reaches a fixed point at Γ_0, completing the proof.[32] □

Whilst Proposition 6 establishes the soundness of WFUTB, Proposition 7 shows that the theory is well positioned with respect to some of the better-known truth systems in terms of proof-theoretic strength. For instance, the Kripke-Feferman theory KF, that classically axiomatizes the family of Kripke's fixed-point models over the Strong Kleene evaluation scheme, is relatively interpretable already in $\mathsf{RT}_{<\epsilon_0}$. So is Halbach's PUTB. In turn, FS, for "Friedman-Sheard", is even weaker than the latter.[33] Proposition 7 shows that WFUTB is stronger than these four renowned systems.[34]

[32] See Feferman [4], Halbach [10, sec. 22.2].

[33] See Feferman [4] for KF, Halbach [9] for PUTB, and Friedman & Sheard [7] and Halbach [8] for FS.

[34] These results make WFUTB and other (classical disquotational) systems that will be introduced next attractive candidates for minimalist theories of truth. As it's often quoted, according to Horwich [16, p. 42] "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."

Yet, WFUTB can be soundly strengthened in several ways. For instance, we can close WFUTB's truth predicate under provable equivalence in URfn(PAT), that is, we can replace WFUTB's truth-theoretic axiom schema with the following:

$$\forall \mathbf{t} \forall x\, (\mathrm{RSt}(x(\dot{\mathbf{t}})) \wedge \mathrm{Wf}_{\mathsf{PA}}(x(\dot{\mathbf{t}})) \wedge \mathrm{Bew}_{\mathsf{URfn(PAT)}}(\ulcorner \varphi(\dot{\mathbf{t}}) \urcorner \dot{\leftrightarrow} x(\dot{\mathbf{t}})) \rightarrow (\mathrm{T}\ulcorner \varphi(\dot{\mathbf{t}}) \urcorner \leftrightarrow \varphi(\mathbf{t}^{\circ})))$$

where φ is a formula with exactly n free variables $v_1, \ldots, v_n$, $x(\dot{\mathbf{t}})$ is short for $x(t_1/\ulcorner v_1 \urcorner) \ldots (t_n/\ulcorner v_n \urcorner)$, and $\mathrm{Bew}_{\mathsf{URfn(PAT)}}(x)$ weakly represents provability in URfn(PAT). Call the strengthened system WFUTB$^+$. In WFUTB$^+$, not only r-stable well-founded sentences have corresponding instances of disquotation, but also those that are provably equivalent in URfn(PAT) to r-stable well-founded sentences. This includes, among other expressions, unfounded logical truths and falsities such as $\forall x\, (\mathrm{T}x \rightarrow \mathrm{T}x)$ and $\exists x\, (0 \neq 0 \wedge \mathrm{T}x)$, which are obviously harmless but are excluded from WFUTB's truth principles.

The ω-consistency of WFUTB$^+$ follows from Lemma 2, which establishes that every sentence φ depends on Θ_φ, the set of sentences φ d-refers to. This implies that well-founded sentences can be shown to be grounded in the sense of Leitgeb.[35] Since dependence is closed under equivalence in every expansion of $\mathbb{N}$ to $\mathscr{L}_{\mathrm{T}}$ and, a fortiori, in URfn(PAT), so is Leitgeb's truth predicate, which means that his semantic theory $\langle \mathbb{N}, \Phi_{\mathrm{lf}} \rangle$ is a model of WFUTB$^+$.[36]

Alternatively, we could extend WFUTB along the lines suggested by the results established in Section 3.2. By Proposition 4, any theory extending WFUTB with finitely many instances of the T-schema for r-stable non-self-referential non-well-founded sentences is ω-consistent. For example, we could add instances of disquotation for finite subsets of the Visser-Yablo sentences in (5) to WFUTB without stepping into ω-inconsistencies.

In turn, Proposition 5 allows us to ω-consistently extend WFUTB with all instances of the Uniform T-schema for arbitrary r-stable non-self-referential non-well-founded sentences, provided that they d-refer to finitely many expressions. Call the resulting theory FUUTB, for "Fintely Unfounded Uniform Tarski Biconditionals". This theory leaves the Visser-Yablo sentences out but allows for instances of disquotation for, e.g. sentences in ω-chains such as the 'truth-teller' sequence, given by an infinite list of sentences, each of which says only of the one coming after that is true.[37]

Unlike WFUTB's, the truth predicate of the extensions considered in the previous paragraph cannot always be closed under equivalence, on pain of unsoundness. To see it, notice that $\forall x\, (x = 1 \rightarrow \neg \mathrm{T}x)$ is logically equivalent to the strong liar $\neg \mathrm{Tl}$ $(= \mathrm{l})$ but also r-stable, non-self-referential – it only refers to $\neg \mathrm{Tl}$, which just refers to itself – and it d-refers to finitely many expressions.

Allow me to consider one last disquotational theory, NSRTB, for "Non-self-referential Tarski Biconditionals". NSRTB extends URfn(PAT) with all instances of the following schema:

$$\mathrm{RSt}(\ulcorner \varphi \urcorner) \wedge \neg \mathrm{SRef}_{\mathsf{PA}}(\ulcorner \varphi \urcorner) \rightarrow (\mathrm{T}\ulcorner \varphi \urcorner \leftrightarrow \varphi) \tag{15}$$

[35] See Leitgeb [21, Lemma 11].
[36] See Leitgeb [21, Lemma 5].
[37] See Picollo [26, Prop. 3.3].

where φ is a sentence. Every r-stable non-self-referential sentence has a corresponding instance of disquotation in NSRTB, including all sentences on the Visser-Yablo list. As a consequence, the theory is ω-inconsistent (cf. Section 2.2) and, therefore, unsound. However, it can be easily seen that NSRTB is not inconsistent, by Proposition 4 and an application of compactness. Thus, we may extract the general conclusion that sets of non-self-referential sentences of $\mathscr{L}_{\mathrm{T}}$ cannot be paradoxical simpliciter but only ω-paradoxical.

But note also that, if we adopt the uniform version of (15) instead, that is,

$$\forall \mathbf{t}\,(\mathrm{RSt}(\ulcorner\varphi(\dot{\mathbf{t}})\urcorner) \wedge \neg\mathrm{SRef}_{\mathsf{PA}}(\ulcorner\varphi(\dot{\mathbf{t}})\urcorner) \rightarrow (\mathrm{T}\ulcorner\varphi(\dot{\mathbf{t}})\urcorner \leftrightarrow \varphi(\mathbf{t}^{\circ})))$$

the resulting system is downright inconsistent, for we can show in URfn(PA) that all instances of the Visser-Yablo predicate $\mathrm{Y}(v)$ satisfy the antecedent of this principle.[38] Since no new instances of disquotation *for sentences* are allowed by the uniform version, one might wonder what has gone wrong. According to Priest [29], Beall [1], Cook [3], and others, the inconsistency is a consequence of admitting instances of (uniform) disquotation, not for self-referential sentences, but for self-referential or circular *predicates*, e.g. $\mathrm{Y}(v)$. Although the Visser-Yablo sentences are not self-referential, the argument goes, they are formulated in terms of a circular predicate and, therefore, are circular themselves. To evaluate this claim, adequate notions of reference, self-reference, etc. that apply to *formulae*, and not just sentences, are required. I leave the task of extending the notions introduced in Section 2.3 to formulae for another time. Let me just say that, provided a natural extension is possible, the claim seems quite plausible.

5 Conclusions

In this paper I have explored a number of semantic and axiomatic truth theories motivated by the notions of reference I introduced in [27]. I have shown the latter to be proof-theoretically strong compared to most well-known systems in the literature, despite being purely disquotational. However, proof-theoretic strength is not necessarily a sign (or the only sign) of theoretical value. To assess the worth of the systems in themselves and compared to others, in this concluding section I test them against the criteria discussed in Leitgeb [22].

The criteria are the following:

(a) Truth should be expressed by a predicate, and a theory of syntax should be available.
(b) If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.
(c) The truth predicate should not be subject to any type restrictions.
(d) T-biconditionals should be derivable unrestrictedly.
(e) Truth should be compositional.
(f) The theory should allow for standard interpretations.

[38] See Ketland [18, 19] and (4).

(g) The outer logic and the inner logic should coincide.
(h) The outer logic should be classical.

As Leitgeb shows, it is not possible to meet all these criteria at once. Every formal account of truth will necessarily have to give up some of these requirements. It will be instructive to see how the systems we have considered here fare.

Clearly, all of our systems – WFUTB, WFUTB$^+$, FUUTB, and NSRTB – meet the first criterion, for T is a predicate and arithmetic plays the role of the syntax theory in the background. The same holds of other popular theories considered in the previous section, i.e. PUTB, KF, FS, and systems of ramified truth.

(b), on the other hand, is met only partially. For every theorem φ of our base theory – URfn(PA) – each of our systems can be shown to entail $\mathrm{T}\ulcorner\varphi\urcorner$. This is because φ is T-free and, therefore, well-founded. However, none of the systems can prove what Leitgeb demands, i.e. a general statement asserting that every theorem of URfn(PA) is true. This would amount to showing that the systems entail URfn(PA)'s global reflection principle:

$$(\mathrm{GRfn}_{\mathsf{URfn(PA)}}) \qquad \forall x\,(\mathrm{Bew}_{\mathsf{URfn(PA)}}(x) \rightarrow \mathrm{T}x)$$

KF, FS, and all systems of ramified truth are known to entail similar principles for their respective base theories. PUTB, on the other hand, doesn't, as shown by Halbach [9, §6]. We can adapt his proof to show that $\mathrm{GRfn}_{\mathsf{URfn(PA)}}$ is not provable in any of our systems either. Using Kleene's Recursion Theorem let us define a p.r. function s, represented in PA by the function symbol $\dot{\mathrm{s}}$, as follows:

$$\mathrm{s}(\varphi, n) := \begin{cases} \varphi & \text{if } \varphi := s = t \\ \mathrm{T}\dot{\mathrm{s}}(t, \overline{n}) \wedge \mathrm{lh}(t) \leq \overline{n} & \text{if } \varphi := \mathrm{T}t \\ \neg \mathrm{s}(\psi, n) & \text{if } \varphi := \neg\psi \\ \mathrm{s}(\psi, n) \wedge \mathrm{s}(\chi, n) & \text{if } \varphi := (\psi \wedge \chi) \\ \mathrm{s}(\psi, n) \vee \mathrm{s}(\chi, n) & \text{if } \varphi := (\psi \vee \chi) \\ \mathrm{s}(\psi, n) \rightarrow \mathrm{s}(\chi, n) & \text{if } \varphi := (\psi \rightarrow \chi) \\ \forall v\, \mathrm{s}(\psi, n) & \text{if } \varphi := \forall v\, \psi \\ \exists v\, \mathrm{s}(\psi, n) & \text{if } \varphi := \exists v\, \psi \end{cases}$$

The function symbol $\mathrm{lh}(x)$ represents the function that maps every formula to the number of logical operators occurring in it. Then, we can show the following:

Lemma 4 *If* WFUTB $\vdash \varphi$*, then there is an* $n \in \omega$ *s.t.* WFUTB $\vdash \mathrm{s}(\varphi, n)$*, and similarly for* WFUTB$^+$*,* FUUTB*, and* NSRTB*.*

I omit the proof, as it's roughly the same as the proof of Halbach's Lemma 6.1. An important lesson to draw from this lemma is that none of the systems entails URfn(PA)'s global reflection principle. If they did, then they would also entail

$$\forall x\,(\mathrm{Bew}_{\mathsf{URfn(PA)}}(x) \rightarrow \mathrm{T}\dot{\mathrm{s}}(x, \overline{n}) \wedge \mathrm{lh}(x) \leq \overline{n})$$

for some natural number n, which would fix an upper bound to the number of logical operators occurring in every theorem of URfn(PA).

Moving on to (c), it can be easily seen that all the systems meet this requirement, for in all of them T applies to sentences containing T itself – e.g. $\mathrm{T}\ulcorner\mathrm{T}\ulcorner 0 = 0\urcorner\urcorner$ is

provable in the four systems. So do KF, FS, and PUTB. Systems of ramified truth obviously fail this criterion.

By contrast, (d) obviously doesn't hold of any of the systems we are considering, as the liar sentence λ, for example, doesn't have a corresponding instance of disquotation, on pain of triviality. Nonetheless, some of our systems fare better than others. While WFUTB contains 'the least number' of T-biconditionals, NSRTB is the most encompassing one, and the other two systems lie somewhere in between.

Criterion (e) is not met by our systems either. Suppose one of the systems entailed the following compositional principle:

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

where $\mathrm{Sent}_{\mathscr{L}}(x)$ is a formula of $\mathscr{L}$ representing the property of being a sentence of this language. By Lemma 4, there should be an $n \in \omega$ s.t. the following is also derivable:

$$\forall x \forall y\, (\mathrm{Sent}_{\mathscr{L}}(x \dot{\wedge} y) \rightarrow (\mathrm{T}\dot{\mathrm{s}}(x \dot{\wedge} y, \overline{n}) \wedge \mathrm{lh}(x \dot{\wedge} y) \leq \overline{n} \leftrightarrow \mathrm{T}\dot{\mathrm{s}}(x, \overline{n}) \wedge \mathrm{lh}(x) \leq \overline{n} \wedge \mathrm{T}\dot{\mathrm{s}}(y, \overline{n}) \wedge \mathrm{lh}(y) \leq \overline{n})))$$

Since $\forall x\, (\mathrm{Sent}_{\mathscr{L}}(x) \rightarrow x = \dot{\mathrm{s}}(x, \overline{n}))$ is provable in PA for every $n \in \omega$, we also have that

$$\forall x \forall y\, (\mathrm{Sent}_{\mathscr{L}}(x \dot{\wedge} y) \rightarrow (\mathrm{T}(x \dot{\wedge} y) \wedge \mathrm{lh}(x \dot{\wedge} y) \leq \overline{n} \leftrightarrow \mathrm{T}x \wedge \mathrm{lh}(x) \leq \overline{n} \wedge \mathrm{T}y \wedge \mathrm{lh}(y) \leq \overline{n})))$$

By T$\wedge$, this formula entails the following:

$$\forall x \forall y\, (\mathrm{Sent}_{\mathscr{L}}(x \dot{\wedge} y) \wedge \mathrm{T}(x \dot{\wedge} y) \rightarrow (\mathrm{lh}(x) \leq \overline{n} \wedge \mathrm{lh}(y) \leq \overline{n} \rightarrow \mathrm{lh}(x \dot{\wedge} y) \leq \overline{n}))$$

Since there are theorems of PA with more than n logical operators that take the form of a conjunction, each of whose conjuncts has at most n logical operators, and all our truth systems prove that each single theorem of arithmetic is true, we get a contradiction. Similar arguments can be given for other compositional principles.

Thus, none of the systems puts forward a compositional notion of truth, not even regarding sentences of $\mathscr{L}$. For the same reasons, PUTB's truth predicate is also non-compositional. By contrast, KF, FS, and ramified truth theories are axiomatized by compositional principles. Note, however, that our systems can be consistently extended with compositional principles for $\mathscr{L}$: since every sentence of this language is well-founded, our systems prove all instances of each compositional principle already, which means that these principles are all true in the models we used as witnesses to the consistency of the systems. This also shows that $\mathrm{GRfn}_{\mathsf{URfn(PA)}}$ can be consistently added to our systems, for it's entailed by the compositional principles for $\mathscr{L}$.[39] Moreover, ω-consistency is preserved in every case.

By contrast, not all of our systems can be consistently extended with compositional principles for expressions containing T. Although unrestricted compositionality might be an unreasonable requirement, one would expect that each truth system is compatible at least with compositional principles for the class of sentences that have a corresponding instance of disquotation in the system. This is certainly the case of

[39] See, for instance, Halbach [10, Theorem 8.32].

WFUTB and WFUTB$^+$. This result follows from the fact that the instances of each principle of compositionality for grounded sentences in the sense of Leitgeb are all true in Leitgeb's semantic theory of grounded truth, $\langle \mathbb{N}, \Phi_{\text{lf}} \rangle$, as dependence is closed under logical connectives and equivalence in every expansion of $\mathbb{N}$ to $\mathscr{L}_{\text{T}}$.[40] Thus, the compositional principles for truth relativized to the predicate WFUTB$^+$ deployes in the restriction of disquotation are also true in $\langle \mathbb{N}, \Phi_{\text{lf}} \rangle$ (cf. Lemma 2). Since the latter is a model of WFUTB$^+$, extending this theory with compositionality relativized to this predicate preserves ω-consistency. *A fortiori*, the same can be said of WFUTB and its respective restricting predicate.

On the other hand, neither FUUTB nor NSRTB can be extended with appropriately relativized compositional axioms. If we strongly diagonalize the predicate $\text{T}\dot{\neg}x$ we obtain a term l' s.t. the identity statement $\text{l}' = \ulcorner \text{T}\dot{\neg}\text{l}' \urcorner$ is provable in PA. Note that $\text{T}\dot{\neg}\text{l}'$ d-refers just to the negation of the sentence denoted by l', i.e. to $\neg\text{T}\dot{\neg}\text{l}'$. $\text{T}\dot{\neg}\text{l}'$ is therefore non-self-referential and d-referes to finitely many expressions. Moreover, it is r-stable. Thus, both FUUTB and NSRTB contain an instance of disquotation for this sentence, namely, $\text{T}\ulcorner \text{T}\dot{\neg}\text{l}' \urcorner \leftrightarrow \text{T}\dot{\neg}\text{l}'$. By Leibniz's Law, this entails $\text{Tl}' \leftrightarrow \text{T}\dot{\neg}\text{l}'$, which is incompatible with the compositional principle that says that truth commutes with negation.

Regarding Leitgeb's criterion (f), all our systems except NSRTB fare well. While the latter is ω-inconsistent, as we saw in the last section, the other three systems all have models extending $\mathbb{N}$, which is what Leitgeb had in mind when he demanded that a formal truth theory based on arithmetic allowed for standard interpretations. While KF, PUTB, and ramified truth theories also meet (f), FS famously doesn't.

Finally, we consider criteria (g) and (h) together. According to Leitgeb, the outer logic of a truth theory is given by the logical laws the theory can prove, whereas its inner logic consists of the logical laws the theory proves to be true. Since the systems we are considering are classical, (h) is met by all of them. (g), instead, is only satisfied by WFUTB$^+$. To see this, note that all logical truths are provably equivalent to each other in first-order logic and, thus, also in URfn(PAT). Since, say, $0 = 0$ is well-founded and r-stable, it has an instance of disquotation in WFUTB$^+$. This implies that all logical truths have instances of disquotation as well in the system. Given that they are all provable, we can conclude that they are also provably true. In contrast, none of the other systems has an instance of disquotation for, e.g. $\forall x\,(\text{T}x \rightarrow \text{T}x)$. This is a logical truth, yet it q-refers to every sentence. Thus, it is self-referential, so it's not declared true by WFUTB, FUUTB, or NSRTB. While FS is known for its matching inner and outer logics, KF, PUTB, and systems of ramified truth don't meet this criterion.

It's time to take stock. As we have seen, all the axiomatic systems we have considered in this paper fare equally well regarding criteria (a), (d), and (h). Due to their pure disquotational character, like PUTB the new systems fail requirements (b) and (e). However, unlike it, both WFUTB and WFUTB$^+$ can be soundly extended with compositional principles for the language with the truth predicate. Moreover, unlike the systems of ramified truth, ours are untyped; unlike FS, WFUTB, WFUTB$^+$, and

[40] See Leitgeb [21, Lemma 5].

FUUTB, ours allow for standard interpetations; and, unlike KF and PUTB, the inner and the outer logics of WFUTB^+ do coincide.

This shows that the axiomatic systems introduced in the previous section, especially WFUTB^+, compare favourably to the best known axiomatic truth theories in the literature against Leitgeb's criteria. Additionally, their proof-theoretic power places them above many popular systems.

Acknowledgements I am deeply indebted to Volker Halbach, with whom I had countless fruitful discussions on reference and self-reference over the last six years. I would also like to particularly thank Dan Waxman for extremely helpful comments on the final drafts, Thomas Schindler, for great suggestions and encouragement, and two anonymous referees for serious improvements in clarity and exposition. I should mention as well Eduardo Barrio, Catrin Campbell-Moore, Luca Castaldo, Roy T. Cook, Benedict Eastaugh, Martin Fischer, Hannes Leitgeb, Øystein Linnebo, Carlo Nicolai, Graham Priest, Johannes Stern, Albert Visser, the Buenos Aires Logic Group, the MCMP logic community, and the Oxford logic group. Finally, I would like to thank the Alexander von Humboldt Foundation and, especially, the Deutsche Forschungsgemeinschaft (DFG) for generously funding the research projects "Reference patterns of paradox" (PI 1294/1-1) and "The Logics of Truth: Operational and Substructural Approaches" (GZ HJ 5/1-1, AOBJ 617612).

References

1. Beall, J.C. (2001). Is Yablo's Paradox non-circular? *Analysis*, *61*, 176–187.
2. Carnap, R. (1937). *Logische Syntax der Sprache*. London: Routledge.
3. Cook, R.T. (2006). There are non-circular paradoxes (but Yablo's isn't one of them!). *The Monist*, *89*(1), 118–149.
4. Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*, 1–49.
5. Field, H. (2008). *Saving truth from paradox*. New York: Oxford University Press.
6. Fischer, M., Halbach, V., Stern, J., Kreiner, J. (2015). Axiomatizing semantic theories of truth? *Review of Symbolic Logic*, *8*, 257–278.
7. Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, *33*, 1–21.
8. Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, *35*, 311–327.
9. Halbach, V. (2009). Reducing compositional to disquotational truth. *Review of Symbolic Logic*, *2*, 786–798.
10. Halbach, V. (2014). *Axiomatic theories of truth*, 2nd edn. Cambridge: Cambridge University Press.
11. Halbach, V., & Horsten, L. (2006). Axiomatizing kripke's theory of truth. *Journal of Symbolic Logic*, *71*, 677–712.
12. Halbach, V., & Visser, A. (2014). Self-reference in arithmetic I. *Review of Symbolic Logic*, *7*, 671–691.
13. Halbach, V., & Visser, A. (2014). Self-reference in arithmetic II. *Review of Symbolic Logic*, *7*, 692–712.
14. Hardy, J. (1995). Is Yablo's paradox liar-like? *Analysis*, *55*(3), 197–198.
15. Herzberger, H. (1970). Paradoxes of grounding in semantics. *Journal of Philosophical Logic*, *67*, 145–167.
16. Horwich, P. (1998). *Truth*, 2nd edn. Oxford: Blackwell.
17. Jeroslow, R.G. (1973). Redundancies in the Hilbert-Bernays derivability conditions for gödel's second incompleteness theorem. *Journal of Symbolic Logic*, *38*, 359–367.
18. Ketland, J. (2004). Bueno and colyvan on yablo's paradox. *Analysis*, *64*, 165–172.

19. Ketland, J. (2005). Yablo's paradox and ω-inconsistency. *Synthese*, *145*, 295–307.
20. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosphy*, *72*, 690–716.
21. Leitgeb, H. (2005). What truth depends on. *Journal of Philosphical Logic*, *34*, 155–192.
22. Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, *2*(2), 276–290.
23. Martin, R.L., & Woodruff, P.W. (1975). On Representing True-in-L in L. *Philosophia*, *5*, 213–217.
24. McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema. *Journal of Philosphical Logic*, *21*, 235–241.
25. Montague, R. (1962). Theories incomparable with respect to relative interpretability. *Journal of Symbolic Logic*, *27*, 195–211.
26. Picollo, L. (2018). Reference in arithmetic. Review of Symbolic Logic, pp. 1–31. https://doi.org/10.1017/S1755020317000351.
27. Picollo, L. Alethic Reference. Journal of Philosophical Logic (to appear).
28. Pohlers, W. (2009). *Proof theory: the first step into impredicativity*. Berlin-Heidelberg: Springer.
29. Priest, G. (1997). Yablo's paradox. *Analysis*, *57*, 236–242.
30. Schindler, T. (2014). Axioms for grounded truth. *Review of Symbolic Logic*, *7*, 73–83.
31. Schindler, T. (2015). A disquotational theory of truth as strong as Z_2^-. *Journal of Philosophical Logic*, *44*, 395–410.
32. Tarski, A. (1935). The concept of truth in formalized languages. In *Logic, semantics, metamathematics* (pp. 152–278). Oxford: Clarendon Press.
33. Tarski, A. (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and Phenomenological Research*, *4*, 341–376.
34. Visser, A. (1989). Semantics and the liar paradox. In Gabbay, D.M., & Günthner, F. (Eds.) *Handbook of philosophical logic*, (Vol. 4 pp. 617-706). Dordrecht: Reidel.
35. Yablo, S. (1985). Truth and reflexion. *Journal of Philosphical Logic*, *14*, 297–349.
36. Yablo, S. (1993). Paradox without self-reference. *Analysis*, *53*, 251–252.