Journal of Philosophical Logic
https://doi.org/10.1007/s10992-019-09524-w

Alethic Reference

Lavinia Picollo[1]

Received: 28 May 2018 / Accepted: 24 July 2019 / Published online: 16 August 2019

Abstract
I put forward precise and appealing notions of reference, self-reference, and well-foundedness for sentences of the language of first-order Peano arithmetic extended with a truth predicate. These notions are intended to play a central role in the study of the reference patterns that underlie expressions leading to semantic paradox and, thus, in the construction of philosophically well-motivated semantic theories of truth.

Keywords Semantic paradoxes · Reference · Self-reference · Well-foundedness

1 Introduction

What is the root of semantic paradox? According to a widely believed hypothesis, famously championed by Russell, Poincaré, and Tarski, the blame can be attributed to circularity. In the case of paradoxes involving truth, such as the paradox of the liar and Curry's paradox, a natural development of this hypothesis diagnoses their paradoxicality as arising from *self-reference.*

Although this view has been around for more than a century, the self-reference diagnosis has never been properly elaborated. The main reason is that the notions of reference and self-reference have proven to be very elusive, despite their frequent use in heuristic contexts in mathematics and mathematical logic.[1]

An additional challenge facing the self-reference diagnosis arises from the Visser-Yablo paradox, involving an infinite list of sentences each of which says that the

[1]For instance, by Gödel [5]. See Smoryński [26] for a pessimistic study of the development of self-reference in logic and mathematics so far and Picollo [21] for more optimistic prospects for the future.

✉ Lavinia Picollo
l.picollo@ucl.ac.uk

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

subsequent sentences are untrue.[2] Appealing to plausible truth principles, it can be shown, paradoxically, that none of these sentences can be true but at least one of them must be at the same time. Many authors have questioned whether the Visser-Yablo paradox exhibits genuine self-reference: since every sentence on the list apparently refers only to the ones that show up later, none obviously refers back to itself.

While the Visser-Yablo paradox triggered an extensive discussion of reference, self-reference, and paradoxicality, one of the main upshots has proven to be that the notions of reference (and self-reference) in the literature are not fit for purpose: the primary contenders are often imprecise, defective, and in some cases even trivial.[3] On this note, the debate ended with a serious scepticism.

My main aim in this paper is to present a systematic and rigorous account of reference *in the context of truth*, which I call "alethic reference", in order to show this scepticism is unfounded. I provide a formally well-defined account of reference and self-reference which, I argue, coheres with and generalizes our guiding intuitive judgements of referentiality. My account is then used to assess the self-reference diagnosis of semantic paradoxicality; I argue that the diagnosis fails, since the Visser-Yablo paradox does not involve self-reference in the relevant sense. However, my theory of reference can nevertheless be used to show that there is a grain of truth in the self-reference hypothesis: in my companion paper [22] I show how restricting the T-schema to sentences which do not exhibit certain problematic patterns of reference – including but not exhausted by self-reference – allows us to construct classically consistent, sound, and attractive theories of truth.

The plan of the paper is as follows. Section 2 provides a technical introduction, followed by an overview of the state of the issue regarding the inadequacy of extent notions of reference and self-reference deployed in the literature on the semantic paradoxes. Section 3 then puts forward precise definitions of reference, self-reference, well-foundedness, and other related concepts, which, I argue, overcome the main obstacles faced by the pre-existing notions. Finally, in Section 4 I extract some conclusions.

2 Technical and Philosophical 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

[2] See Herzberger [12], Visser [28], and Yablo [29, 30].

[3] As stressed by Leitgeb [18] and Cook [4]. A pessimistic conclusion is also reached in Halbach & Visser [7, 8].

standard interpretation of $\mathscr{L}$, with ω as its domain. Then, for each $n \in \omega$, $\mathscr{L}$ has a term $\overline{n}$ (the numeral of n) denoting n, consisting 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. If σ is a string of symbols of $\mathscr{L}_{\mathrm{T}}$, we write $\ulcorner\sigma\urcorner$ for the numeral of its code.[4] 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 $\mathscr{L}_{\mathrm{T}}$ speaks in the first instance about natural numbers, the arithmetization of syntax allows it to express many syntactic properties, relations, and functions. Thus truth theories can be formulated in $\mathscr{L}_{\mathrm{T}}$, with background syntactic principles formulated in $\mathscr{L}$.

Theories can have either a semantic or an axiomatic presentation. A semantic truth theory consists of a model or family of models $\langle \mathbb{N}, \Gamma \rangle$ expanding $\mathbb{N}$ to $\mathscr{L}_{\mathrm{T}}$, where Γ is the extension of T in the model. By contrast, axiomatic truth theories result from adding truth-theoretic axioms to a syntax theory – usually PA. 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.

2.2 Diagonalization and Tarski's Theorem

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:

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

This is often called a "disquotational" principle, and its instances "T-biconditionals". Unfortunately, it cannot be implemented unrestrictedly, as the language is 'expressive enough' to allow for paradoxical expressions such as liar sentences.

Let $\mathbf{v}$ abbreviate a string of variables $v_1, \ldots, v_n$ different from x and y.[5]

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*:*

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

[4] We require that the coding is effective and monotonic to avoid certain issues brought up by Heck [10] and Halbach & Visser [7, 8]. An effective coding is such that, given a number n, there is an algorithm to determine which expression it codifies (if any) and, vice versa, given an expression σ there is an algorithm that delivers the code of σ. A coding is monotonic if, for every two expressions σ and σ', if σ occurs in σ', then the code of σ is smaller than the code of σ'.

[5] The following result due to Montague [20] is a generalization to formulae with an arbitrary number of free variables of a theorem of Carnap [3], which in turn generalizes Gödel's construction of a self-referential statement for the proof of his first incompleteness theorem.

Proof Let $\mathrm{Diag}(x, y)$ represent the p.r. diagonalization function that takes a formula $\varphi(x, \mathbf{v})$ and returns $\forall x\,(x = \ulcorner\varphi(x, \mathbf{v})\urcorner \rightarrow \varphi(x, \mathbf{v}))$. Then,

$$\forall x\,(x = \ulcorner\forall y\,(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))\urcorner \rightarrow \forall y\,(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))) \tag{2}$$

is the result of diagonalizing $\forall y\,(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))$. Notice that (2) is the $\psi(\mathbf{v})$ we are looking for. Let n be the code of (2). (2) is logically equivalent to

$$\forall y\,(\mathrm{Diag}(\ulcorner\forall y(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))\urcorner, y) \rightarrow \varphi(y, \mathbf{v}))$$

which is equivalent in PAT to $\varphi(\overline{n}, \mathbf{v})$. Thus, the following is a theorem of PAT:

$$\forall x\,(x = \ulcorner\forall y\,(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))\urcorner \rightarrow \forall y\,(\mathrm{Diag}(x, y) \rightarrow \varphi(y, \mathbf{v}))) \leftrightarrow \varphi(\overline{n}, \mathbf{v})$$

□

This is the 'universal' proof of the diagonal lemma. An analogous 'existential' proof can be given in terms of an alternative diagonalization function represented by $\mathrm{Diag}^{\exists}(x, y)$ that maps each formula $\varphi(x, \mathbf{v})$ to $\exists x\,(x = \ulcorner\varphi(x, \mathbf{v})\urcorner \wedge \varphi(x, \mathbf{v}))$. Applying 'existential' diagonalization to the predicate $\exists y\,(\mathrm{Diag}^{\exists}(x, y) \wedge \varphi(y, \mathbf{v}))$ we also obtain a suitable $\psi(\mathbf{v})$. This and the proof of Theorem 1 will become relevant later.

In equivalences of the form (1), $\psi(\mathbf{v})$ is said to be a fixed point of $\varphi(x, \mathbf{v})$. If x is the only free variable in φ, ψ is a sentence, commonly regarded as saying of itself that it has the property expressed by $\varphi(x)$, whatever exactly that is. Thus, fixed-point sentences are considered to be self-referential, and diagonalization is seen as the paradigmatic mechanism for obtaining such self-referential sentences.

Let φ in Theorem 1 be $\neg\mathrm{T}x$. Then we know there is a sentence λ such that the following is a theorem of PAT:

$$\lambda \leftrightarrow \neg\mathrm{T}\ulcorner\lambda\urcorner \tag{3}$$

λ is normally understood as a liar sentence, i.e. a sentence saying of itself that it is untrue. Given (3), no classical consistent extension of PAT can contain an instance of the T-schema for λ and, *a fortiori*, unrestricted disquotation is untenable. This, in essence, is Tarski's undefinability result. Similar results arise for semantic theories of truth. No model $\langle\mathbb{N}, \Gamma\rangle$ of $\mathscr{L}_{\mathrm{T}}$ can validate the T-biconditional for λ, as all theorems of PAT are true in $\langle\mathbb{N}, \Gamma\rangle$, including (3). Thus, we say λ is paradoxical.

Another example of a paradox is given by the following:

$$\kappa \leftrightarrow (\mathrm{T}\ulcorner\kappa\urcorner \rightarrow 0 \neq 0) \tag{4}$$

This biconditional obtains in PAT by Theorem 1, diagonalizing $\mathrm{T}x \rightarrow 0 \neq 0$. κ is taken to be a Curry sentence, i.e. a sentence that says of itself that it entails something false. Just like in the case of the liar, the T-biconditional for κ is inconsistent with (4), so we say κ is paradoxical.

As a final example of paradoxicality consider the following 2-liar cycle:

$$\begin{aligned} \lambda_1 &\leftrightarrow \neg\mathrm{T}\ulcorner\lambda_2\urcorner \\ \lambda_2 &\leftrightarrow \mathrm{T}\ulcorner\lambda_1\urcorner \end{aligned} \tag{5}$$

Theorem 1 guarantees that both biconditionals are provable in PAT by diagonalizing $\neg\mathrm{T}\ulcorner\mathrm{T}\dot{x}\urcorner$. 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, and $x(y/z)$ for the (p.r.) 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. Since v is free in $\ulcorner\varphi(\dot{v})\urcorner$, we can quantify over it. Clearly, the equivalences in (5) are inconsistent with the T-biconditionals for λ_1 and λ_2; so we say the latter are paradoxical.

2.3 The Visser-Yablo Paradox and (Self-)Reference

According to the self-reference diagnosis, all paradoxical expressions involving truth share a common reference pattern, namely, self-reference. This is intuitively clear in the case of the liar, Curry sentences, and of liar cycles, but not so much of the sentences that comprise the Visser-Yablo paradox, each of which say only of the ones that follow that they are untrue. The existence of the Visser-Yablo sentences can be proved in PAT, diagonalizing the formula $\forall z\,(z > w \rightarrow \neg \mathrm{T}x(\dot{z}/\ulcorner w\urcorner))$. By Theorem 1, there is a predicate $\mathrm{Y}(w)$ such that

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

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

$$\begin{aligned}
\mathrm{Y}(0) &\leftrightarrow \forall z\,(z > 0 \rightarrow \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner) \\
\mathrm{Y}(\overline{1}) &\leftrightarrow \forall z\,(z > \overline{1} \rightarrow \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner) \\
&\ldots \\
\mathrm{Y}(\overline{n}) &\leftrightarrow \forall z\,(z > \overline{n} \rightarrow \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner) \\
&\ldots
\end{aligned} \tag{6}$$

By reductio ad absurdum, from the T-biconditionals for $\mathrm{Y}(\overline{n})$ we can easily derive $\neg \mathrm{T}\ulcorner \mathrm{Y}(\overline{n})\urcorner$ for each $n \in \omega$, as well as $\neg\forall z\, \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner$. On the other hand, $\forall z\, \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner$ does not follow in PAT plus the T-biconditionals, which means that the theory is consistent.[6] However, note 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 paradox in the strict sense, but an ω-paradox. Despite not directly leading to a contradiction in our axiomatic theories, it is still problematic, as no semantic truth theory can validate all T-biconditionals for the sentences in the list.

The presence or absence of self-reference in the Visser-Yablo list has been extensively discussed in the literature.[7] It ultimately transpired that the notions of reference and self-reference deployed in the discussion were incomplete, defective, and in some cases even trivial. As Leitgeb points out, there are at least two notions of reference at play in the debate, a 'naïve' and an 'incomplete' one. According to the naïve account,

[6] See Hardy [9] and Ketland [14, 15]. Note that inferring $\forall z\, \neg \mathrm{T}\ulcorner \mathrm{Y}(\dot{z})\urcorner$ from the set of all its instances $\neg \mathrm{T}\ulcorner \mathrm{Y}(\overline{n})\urcorner$ would require an infinitary rule, not admissible in finitary systems such as the ones we are working with.

[7] See, for instance, Priest [23], Beall [1], and Cook [4].

self-referential sentences ψ are fixed points of some predicate $\varphi(v)$, as in (1). This accounts for the liar, liar cycles, and other paradoxical expressions such as Curry sentences. Underlying this notion of self-reference is the idea that a sentence refers to every object that is mentioned in an equivalent expression:

Naïve reference: a sentence φ refers to an object o (e.g. a sentence) if there is a sentence $\psi(t)$ that is (e.g. arithmetically) equivalent to φ and t denotes o.

This notion is naïve in the sense of being trivial. For instance, every sentence φ is (logically) equivalent to $\varphi \wedge (\mathrm{T}\ulcorner\varphi\urcorner \rightarrow \mathrm{T}\ulcorner\varphi\urcorner)$ and, more generally, to $\varphi \wedge (\mathrm{T}\ulcorner\psi\urcorner \rightarrow \mathrm{T}\ulcorner\psi\urcorner)$, where ψ can be any sentence. As a consequence, every sentence refers to every sentence, including itself. This is unacceptable; a good notion of reference must impose more restrictive criteria.

According to the second notion Leitgeb traces in the literature, sentences can refer to an object in two ways. They can either contain a term denoting the object or state something about it by means of a description:

Incomplete reference:
1. Reference by mention: a sentence φ refers to an object o if it contains a term denoting o.
2. Reference by description:[8] a sentence of the form $\forall x\,(\varphi(x) \rightarrow \psi(x))$ refers to an object o if the latter satisfies $\varphi(x)$.

This notion nicely reflects pre-theoretical intuitions. For instance, the sentence $\neg\mathrm{T}\ulcorner 0 = 0\urcorner$ surely refers to $0 = 0$; and the sentence $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$, stating that all theorems of PA are true, surely refers to the theorems of PA.

The problem with incomplete reference, however, is that it gives no information as to whether (and how) quantified sentences of a different logical form may refer: for instance, sentences of the form $\forall x\,\varphi$ where φ is not a conditional and existential claims. It is by no means clear how to fill in this gap to account for, e.g. the *prima facie* self-referential status of the liar in (3) or the cycle of liars in (5), given that we cannot hold that reference is closed under *logical* equivalence, on pain of trivializing the notion. In Milne's words,

> Provable material equivalence in a theory is not normally a criterion of synonymy so we must suppose that it is something particular to gödel biconditionals that is at issue. For a number of reasons the case is hard to make. (Milne [19, p. 212])

The deficiency of both available concepts led many to adopt a rather pessimistic attitude towards the notions of reference and self-reference, in particular in the context of arithmetic. Leitgeb [18, p. 13] writes: "we either suspect that much philosophical work lies ahead of us before the question is finally settled, or that otherwise the question is ill-posed, i.e. that the talk of self-referentiality is to be banished from scientific contexts". It's now time to abandon this pessimistic attitude, at least when

[8] I'm following Heck's [10] terminology here. The tacit underlying idea is that claims of the form $\forall x\,(\varphi(x) \rightarrow \psi(x))$ are restricted quantificational claims, saying of the φs that they are ψ.

it comes to reference in the context of theories of truth.[9] In the next section I advance a natural way of completing the incomplete notion of reference *just in the context of truth*, giving the expected verdict for all (normally considered to be) clear cases, including the liar and its variants.

With the right notions of alethic reference and self-reference in place, we will be in a position to properly evaluate the orthodox view on semantic paradoxes. Moreover, the new notions also allow us to formulate sensible restrictive criteria for instances of disquotation resulting in philosophically and technically appealing truth theories, as shown in my companion paper [22].

3 Alethic Reference

3.1 Four Features of Reference

The main purpose of this section is to give an adequate and precise definition of reference in the context of truth, inspired by and extending the incomplete notion outlined in Section 2.3. But before we start, four remarks are in order.

First, one should be careful to confuse the notion of reference we are after neither with the Fregean notion – the truth value of a sentence – nor with the notion of aboutness.[10] Although they have a strong family resemblance, reference is more tied to the syntactic structure of sentences than aboutness. While tautologies are sometimes considered to be about nothing in particular because they convey no information, we would intuitively say that expressions such as $\mathrm{T}\ulcorner\varphi\urcorner \rightarrow \mathrm{T}\ulcorner\varphi\urcorner$ still refer to φ, as Leitgeb's incomplete notion predicts.[11]

Second, and related to the previous point, closure under logical equivalence should not be required from a definition of reference, as indicated in Section 2.3. On pain of triviality, reference cannot be extensional. It must be *hyperintensional*.

Third, throughout the paper we will understand reference exclusively as a binary relation between *sentences* of $\mathscr{L}_{\mathrm{T}}$, and not between sentences and numbers. The reason is simple. We are concerned with arithmetic not as a theory of numbers but of syntax, for the study of formal truth theories and the semantic paradoxes and related phenomena that might affect them.

Note that, since reference to sentences is achieved via a coding, whether a sentence refers to another will inevitably depend on the coding we choose. This choice, even if restricted to effective and monotonic codings, is always fairly arbitrary. As a consequence, what sentences an expression refers to is also very often an arbitrary matter. This is to be expected. The coding we work with fixes the denotation of the terms of

[9] I hope to have dissipated some doubts already in Picollo [21], as I have shown how to define reference, self-reference, and other referential patterns in the pure language of arithmetic. Unfortunately, it is not possible to simply extend those notions to $\mathscr{L}_{\mathrm{T}}$, as what we are after in this paper is not reference simpliciter but a special kind of reference that only concerns the truth predicate. I come back to this in footnote 15.

[10] See Putnam [24] and Goodman [6] for a historical overview and Urbaniak [27] and Yablo [31] for modern takes on aboutness.

[11] See as well Leitgeb [18], Cook [4], Halbach & Visser [7, 8], and footnote 17.

$\mathscr{L}$ to sentences of $\mathscr{L}_{\mathrm{T}}$ from 'outside'.[12] It is only natural that what a sentence refers to depends on the denotation of the terms that occur in it. Moreover, the arbitrariness of the coding will not cause any inconvenience in the formulation of truth systems, as what counts as an instance of disquotation also depends on the coding and varies accordingly.[13]

Finally, we will only focus on *alethic* reference, that is, reference in the context of truth. Let $\mathrm{Bew}_{\mathsf{PA}}(x)$ weakly represent provability in PA in a natural way.[14] If we diagonalize $\neg\mathrm{Bew}_{\mathsf{PA}}(x)$ we obtain a sentence γ such that

$$\gamma \leftrightarrow \neg\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\gamma\urcorner) \tag{7}$$

is provable in PA. γ is also known as the "Gödel sentence of PA". Since γ has been obtained exactly in the same way as the liar sentence λ, one might conclude it is as self-referential as the latter. Nonetheless, γ is completely unparadoxical, as is every other expression belonging to the pure language of arithmetic. Self-reference and other pathological patterns only become dangerous when combined with the truth predicate (and other semantic or logical notions such as satisfaction, property instantiation, and class membership). Thus, we are only interested in the sentences an expression φ refers to insofar as they fall in the scope of T in φ.[15] For instance, we would like to say that $\ulcorner\psi\urcorner = \ulcorner\psi\urcorner \wedge \neg\mathrm{T}\ulcorner\varphi\urcorner$ or $\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\psi\urcorner) \rightarrow \mathrm{T}\ulcorner\varphi\urcorner$ alethically refer – or just "refer", for short – to φ but not to ψ. Making this idea precise is the aim of the rest of this section.

3.2 Reference by Mention

The incomplete notion of reference identifies two ways in which sentences may refer: by mention and by description. Let us first spell out the alethic version of reference by mention (m-reference) in precise terms.

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 φ m-refers only to those sentences denoted by closed terms that follow T in φ. This means, on the one hand, that terms occurring in the scope of other predicates do not play a role, as anticipated. For instance, while $\neg\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\gamma\urcorner)$ doesn't m-refer to any expression, $\mathrm{Bew}_{\mathsf{PA}}(\ulcorner 0 = 0\urcorner) \wedge \mathrm{T}\ulcorner 0 \neq 0\urcorner$ m-refers to $0 \neq 0$ but not to $0 = 0$. On the other hand, it means that proper subterms of terms occurring after T are to be ignored. Let $\dot{\neg}$ be a function symbol of $\mathscr{L}$ representing the p.r. function that

[12] One might wonder, in the light of Putnam's [25] model-theoretic argument, if there is an alternative way of fixing the reference of the terms of a language at all.

[13] See Heck [10].

[14] See Halbach & Visser [7, 8].

[15] This is why extending the notions of reference for the pure language of arithmetic I put forward in Picollo [21] to $\mathscr{L}_{\mathrm{T}}$ is not viable, as anticipated in footnote 9. However, they can serve a heuristic role in the formulation of the alethic notions.

maps each formula to its negation (and similarly for other logical connectives). Then $\mathrm{T}\dot{\neg}\ulcorner\varphi\urcorner$ m-refers not to φ but to $\neg\varphi$.

This may seem too restrictive. Strictly speaking, only the sentence denoted by t (if any) falls in the scope of the truth predicate in $\mathrm{T}t$. Nonetheless, one might think that by ignoring t's subterms we could be overlooking dangerous reference patterns. I show this is not the case in my companion paper [22]. Moreover, note that allowing that $\mathrm{T}t$ m-refers to all denotations of subterms of t would result in a grossly over-generating notion. For instance, $\mathrm{T}\dot{\neg}\ulcorner\varphi\urcorner$ would m-refer not only to φ but also to every sentence ψ whose code is smaller than φ's; and the same goes for $\mathrm{T}\ulcorner\varphi\urcorner$. For $\ulcorner\varphi\urcorner$ is a complex term of the form $\mathrm{S}\ldots\mathrm{S0}$, of which $\ulcorner\psi\urcorner$ is a subterm. As a consequence, many expressions would be wrongfully classified as self-referential or unfounded, prompting unnecessarily restrictive truth theories.

Although the equivalences delivered by Theorem 1 do not suffice to establish that the fixed points obtained by Diagonalization m-refer to themselves even in the presence of T, there is a stronger version of this result that does so to a certain extent.

Theorem 2 (**Strong Diagonalization**) *For every formula $\varphi(x, \mathbf{v})$ there is a term t s.t. the following is a theorem of* PA*:*

$$t = \ulcorner\varphi(t, \mathbf{v})\urcorner$$

We say that $\varphi(t, \mathbf{v})$ is a strong fixed point of $\varphi(x, \mathbf{v})$. A proof of this result can be found in Jeroslow [13].[16] All that is required is that the language contains a function symbol for the substitution function, as is the case of $\mathscr{L}$. Thus, 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.[17] $\neg\mathrm{Tl}$ is a 'strong' liar sentence. In general, strong fixed points of formulae $\varphi(x)$ with subformulae of the form $\mathrm{T}x$ in which x occurs free are self-m-referential according to Definition 1.

However, self-m-reference might not obtain if, instead, T is followed by an open term $s(x)$ in $\varphi(x)$. For instance, suppose $s(x)$ represents a function mapping each sentence to $0 = 0$. In that case, every strong fixed point of $\neg\mathrm{T}s(x)$ will refer to $0 = 0$ and not to itself. As another example, consider the formula $\mathrm{T}\dot{\neg}x$. Strong Diagonalization delivers a term l' such that

$$\mathrm{l}' = \ulcorner\mathrm{T}\dot{\neg}\mathrm{l}'\urcorner \tag{8}$$

is a theorem of PA. $\mathrm{T}\dot{\neg}\mathrm{l}'$ is a sentence that says not of itself but of *its negation* that it is true. However, note that, according to Definition 1, whilst $\mathrm{T}\dot{\neg}\mathrm{l}'$ does not m-refer to itself, its negation does. Thus, despite not being self-referential $\mathrm{T}\dot{\neg}\mathrm{l}'$ will be deemed unfounded (cf. Definition 9), as it m-refers to a self-referential expression.

[16] It is not clear who formulated the result first, but applications of it can be found already in the 1950s, e.g. by Henkin [11] and Kreisel [16].

[17] The condition that a sentence m-refers to itself is also known as the "Kreisel-Henkin" criterion for self-reference. See Halbach & Visser [7].

All that matters for m-reference is the occurrence of subsentences of the form Tt. Any two sentences with the same atomic subsentences of this kind m-refer to the same expressions. This means the notion is trivially compositional: $\neg\varphi$, $\forall v\,\varphi$, and $\exists v\,\varphi$ m-refer to whatever φ m-refers to, and $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, and $(\varphi \rightarrow \psi)$ m-refer to everything either φ or ψ m-refer to. Thus, m-reference is closed under negation, but also renaming of variables and the introduction and elimination of dummy quantifiers – i.e. that don't bind any variable. As a consequence, it is also closed under a kind of equivalence – more fine grained than logical equivalence – that allows for all valid transformations that do not add or remove *types* of atoms. Furthermore, m-reference is closed under Leibniz's Law: if $s = t$, then φ and $\varphi[s/t]$ m-refer to the same expressions.

3.3 Reference by Quantification

Things are somewhat more complicated in the case of reference by description and, in general, in cases where reference is achieved via the presence of quantifiers, which I call "reference by quantification" (or "q-reference", for short). Let's start by focusing on a paradigmatic example, PA's global reflection principle:

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

According to clause 2 of the incomplete notion of reference, ($\mathrm{GRfn}_{\mathsf{PA}}$) refers *just* to all theorems of PA, for it says only of the latter that they are true. The conditional following a universal quantifier *restricts* the referenced sentences to those satisfying the antecedent of the conditional. Thus, in the absence of a conditional, e.g. in $\forall x\,\mathrm{T}x$ and $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \wedge \mathrm{T}x)$, it appears that the sentence must refer to everything.

However, we should not endorse these conditions unrestrictedly in the case of alethic reference.

For one thing, in order to keep reference by mention and reference by quantification apart, it seems reasonable to require that a free variable occurs in the scope of the truth predicate in a sentence for it to refer to other sentences by quantification. For another, we only say that in a statement of the form $\forall v\,(\varphi(v) \rightarrow \psi(v))$ q-reference is restricted to the φs as long as it is possible to determine which sentences satisfy $\varphi(v)$: i.e. if T doesn't occur in φ. Only then do we regard the conditional as restricting reference by means of the antecedent. Recall the aim of this paper is to offer an appropriate account of reference for the study of the reference patterns underlying paradoxical expressions. However, the ultimate hope (realized in the companion paper [22]) is to find shared patterns in terms of which we can formulate sensible restricting criteria on the instances of disquotation. In turn, the truth principles we embrace will then determine the extension of the truth predicate. On this approach, there is no 'standard' interpretation of this predicate at our disposal yet. Thus, if T occurs both in φ and ψ, e.g. in $\forall x\,(\mathrm{T}x \rightarrow \mathrm{T}x)$, we will say the expression refers to every sentence, for the conditional is not reference restricting.

Moreover, complex terms occurring in the scope of T should play a role in the q-reference of expressions just as they do for m-reference. Let φ be a sentence. In Section 3.2 I stipulated that, e.g. $\mathrm{T}\dot{\neg}\ulcorner\varphi\urcorner$ m-refers to $\neg\varphi$ and not to φ, because the

former and not the latter occurs in the scope of T. I argued that considering subterms is neither necessary nor desirable. For analogous reasons, I would like to say that

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

q-refers, not to φ, but to $\neg\varphi$. Similarly, it seems reasonable to say that $\forall x\,\mathrm{T}\dot{\neg}x$ q-refers not to every sentence but just to all negations, and $\forall x \forall y\,\mathrm{T}(x \dot{\rightarrow} y)$ to all sentences of conditional form.

A *prima facie* sensible way of capturing this idea is to say that universal claims without a reference-restricting conditional q-refer to what their instances m-refer to, whereas for those with a reference-restricting conditional only the instances with a true antecedent matter. But in order to keep m- and q-reference apart, we should add that the occurrences of the terms in the scope of T in the instances are 'new', that is, they are not present in the sentence itself but are a product of the instantiation of the quantifiers at issue. For instance, $\forall x\,(x = \ulcorner\varphi\urcorner \rightarrow \mathrm{T}x \wedge \mathrm{T}\ulcorner\psi\urcorner)$ would q-refer to φ but refer to ψ only by mention.

However, this proposal is not general enough. It could be that, intuitively, the relevant instances of a universal statement don't m- but q-refer themselves to other sentences, in the presence of *embedded quantifiers*. Consider the following sentence:

$$\forall x\,(x = \ulcorner\varphi\urcorner \rightarrow \forall y\,(y = \dot{\neg}x \rightarrow \mathrm{T}y)) \tag{10}$$

By analogy with (9), (10) seems to q-refer to what $\forall y\,(y = \dot{\neg}\ulcorner\varphi\urcorner \rightarrow \mathrm{T}y)$ refers to. But the latter does not m-refer to any expression. Intuitively, it q-refers to $\neg\varphi$. Likewise, the numerical instances of

$$\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \wedge \forall y\,\mathrm{T}(y \dot{\rightarrow} x)) \tag{11}$$

are of the form $\mathrm{Bew}_{\mathsf{PA}}(\overline{n}) \wedge \forall y\,\mathrm{T}(y \dot{\rightarrow} \overline{n})$. Although they do not m-refer to any expression, it looks like they do q-refer, provided that n codes a sentence φ. In each such case, we would like to say that $\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\varphi\urcorner) \wedge \forall y\,\mathrm{T}(y \dot{\rightarrow} \ulcorner\varphi\urcorner)$ q-refers to all conditional sentences that have φ as their consequent, due to the right conjunct. Q-reference calls for a *recursive* definition whereby q-referents are specified by considering the m- and q-referents of the quantifier instances of its subsentences. This can be done roughly along the following lines: a universally quantified expression q-refers to what its instances newly m-refer or q-refer to, unless it has a reference-restricting conditional, in which case only the instances with true antecedents are to be considered.

Yet, we should be careful when dealing with *strings of quantifiers*. Consider, for example, the following expression:

$$\forall x \forall y\,(\mathrm{Bew}_{\mathsf{PA}}(x \dot{\wedge} y) \rightarrow \mathrm{T}x) \tag{12}$$

Intuitively, (12) q-refers just to theorems of PA, as only these fall in the scope of T in the instances satisfying the antecedent. But if we applied our recursive clause unrestrictedly, we would be forced to conclude that (12) q-refers to every sentence φ of the language, as each $\forall y\,(\mathrm{Bew}_{\mathsf{PA}}(\ulcorner\varphi\urcorner \dot{\wedge} y) \rightarrow \mathrm{T}\ulcorner\varphi\urcorner)$ m-refers to φ. The recursive clause should be modified to take into account consecutive quantifiers 'all at once'.

What about expressions of a different logical form, such as $\exists v\,\varphi(v)$, $\neg\forall v\,\varphi(v)$, and $\exists v\,(\varphi(v) \wedge \psi(v))$? A compositional notion of reference by quantification seems desirable, as in the case of m-reference. This can be achieved by tying q-reference

to the presence of quantified expressions as *subformulae*. In this way, for instance, sentences of the form $\neg\forall v\,\varphi(v)$ and $\forall v\,\varphi(v)$ would q-refer to the same expressions. It would also be desirable to close q-reference under renaming of variables, addition and removal of dummy quantifiers, and most importantly, under certain valid propositional transformations that do not add or remove atoms. For example, although φ and $\varphi \wedge \forall x\,(\mathrm{T}x \rightarrow \mathrm{T}x)$ cannot always be said to q-refer to the same sentences, the following pairs of expressions intuitively do: $\forall v\,(\varphi \rightarrow \psi)$ and $\forall v\,(\neg\psi \rightarrow \neg\varphi)$, $\forall v\,(\neg\varphi \vee \psi)$ and $\forall v\,(\varphi \rightarrow \psi)$, $\exists v\,(\varphi \wedge \psi)$ and $\exists v\,(\psi \wedge \varphi)$, $\forall v\,\varphi$ and $\forall v\,\neg\neg\varphi$, and $\forall v\,\neg\varphi$ and $\neg\exists v\,\varphi$. A definition of q-reference that focuses on universal claims but is closed under these and other propositional transformations could fix the q-reference of expressions of any logical form, including existential statements.

For this reason, to provide a formally precise definition with the outlined characteristics, we first take a detour that consists in 'normalizing' all formulae of $\mathscr{L}_{\mathrm{T}}$. Then, q-reference for sentences is defined in terms of their normalizations. Note that distinct sentences may have the same normalization. The normalization procedure therefore induces an equivalence relation – *having the same normalization* – under which q-reference is closed. The notion, however, is not trivialized, for this equivalence relation is more fine grained than logical equivalence (cf. Section 2.3).

Call "prime" any atomic or universal formulae, or their negation. 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).

Definition 2 (**ADNF**) A formula is in ADNF iff it contains no dummy quantifiers and every subformula of the form $\forall v\,\varphi$ satisfies the following two conditions:

1. φ is a disjunction (of length ≥ 0) of conjunctions (of length ≥ 0) of primes;
2. if φ is of the form $\psi \vee \chi$, then ψ contains all and only the T-free disjuncts of φ, if any (and, consequently, χ contains all of the T-containing disjuncts, if any).

Formulae in ADNF of the form $\forall \mathbf{v}\,\varphi$ are roughly in prenex disjunctive normal form, as normally defined in textbooks,[18] except quantifiers are not rearranged to appear at the beginning of the formula. The point of writing these sentences in ADNF is that one can easily see whether or not they take the form of a restricted quantified claim. Consider a sentence containing T whose normalized form is $\forall \mathbf{v}\,\varphi$. Clause 1 ensures that φ is written in disjunctive form; clause 2 ensures that all of the T-free disjuncts of φ (if any) are pushed to the left. Thus if φ is $\psi \vee \chi$ where the disjuncts of ψ are T-free and the disjuncts of χ are not, the original sentence can be seen to be equivalent to the restricted quantificational claim $\forall \mathbf{v}\,(\neg\psi \rightarrow \chi)$. 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 the quantifiers $\forall \mathbf{v}$. If, in contrast, φ is not a disjunction or ψ is not T-free, the quantifiers $\forall \vec{v}$ in $\forall \vec{v}\,\varphi$ are unrestricted.

Since formulae in ADNF cannot contain conditionals or existential quantifiers, the first step in the normalization of an expression is to replace these connectives

[18] See, for instance, Boolos et al. [2, §19.1].

with negations, conjunctions, disjunctions, and universal quantifiers, making use of the standard definitions. Let $\tau : \mathscr{L}_{\mathrm{T}} \to \mathscr{L}_{\mathrm{T}}$ carry out these replacements, that is, $\tau(\varphi \to \psi) := \neg\tau(\varphi) \vee \tau(\psi)$ and $\tau(\exists v\, \varphi) := \neg\forall v\, \neg\tau(\varphi)$. Once conditionals and existential quantifiers have been removed, normalization proceeds in stages. It consists of successive transformations of each subformula of the form $\forall v\, \varphi$ into ADNF, starting from those with fewer embedded quantifiers, i.e. of lesser *depth*. Let *dep* assign numbers to universally quantified formulae without conditionals or existential quantifiers as follows:

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

For every formula φ without conditionals, existential quantifiers, or dummy quantifiers, its i-normalization 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.

Since every step in the i-normalization of a formula involves only finitely many transformations, the process always terminates. Therefore, we can introduce the following definition:

Definition 3 (**Normalization**) The normalization φ^* of a formula φ is the result of erasing all dummy quantifiers in $\tau(\varphi)$ and, if there are any quantifiers left, performing successive i-normalizations starting with $i = 1$ until a fixed point is reached.[19]

By way of example, consider the following sentence (we assume $\mathrm{Bew}_{\mathsf{PA}}(x)$ is already in ADNF):

$$\forall z \forall x \forall y\, (\exists z \exists w\, (z = w \dot{\to} y \ \wedge\ \neg \mathrm{T}z) \ \to\ \mathrm{Bew}_{\mathsf{PA}}(x) \ \vee\ y \neq \dot{\neg} x) \tag{13}$$

To normalize it, we first apply τ to get rid of conditionals and existential quantifiers:

$$\forall z \forall x \forall y\, (\neg\neg\forall z\, \neg\neg\forall w\, \neg(z = w \dot{\to} y \ \wedge\ \neg \mathrm{T}z) \vee\ \mathrm{Bew}_{\mathsf{PA}}(x) \ \vee\ y \neq \dot{\neg} x)$$

[19] It's guaranteed that we reach a fixed point, for the maximum depth of φ's subformulae is always finite.

Then, we then erase the dummy occurrence of $\forall z$ at the beginning and 1-normalize the resulting expression, that is, we distribute the negation over the disjunction and erase double negations in $\forall w\, \neg(z = w \dot{\rightarrow} y \ \wedge\ \neg \mathrm{T}z)$ – the only subformula of depth 1 – obtaining:

$$\forall x \forall y\, (\neg\neg\forall z\, \neg\neg\forall w\, (z \neq w \dot{\rightarrow} y \ \vee\ \mathrm{T}z) \vee \mathrm{Bew}_{\mathsf{PA}}(x) \ \vee\ y \neq \dot{\neg} x)$$

Next, we 2-normalize the sentence above, erasing the double negation in $\forall z\, \neg\neg\forall w\, (z \neq w \dot{\rightarrow} y \ \vee\ \mathrm{T}z)$, the only subformula of depth 2, obtaining:

$$\forall x \forall y\, (\neg\neg\forall z \forall w\, (z \neq w \dot{\rightarrow} y \ \vee\ \mathrm{T}z) \vee \mathrm{Bew}_{\mathsf{PA}}(x) \ \vee\ y \neq \dot{\neg} x)$$

Finally, we 3-normalize the latter, erasing the double negation and swapping disjuncts so that the T-free ones occur on the left. The result is the normalization of (13):

$$\forall x \forall y\, (\mathrm{Bew}_{\mathsf{PA}}(x) \ \vee\ y \neq \dot{\neg} x \ \vee\ \forall z \forall w\, (z \neq w \dot{\rightarrow} y \ \vee\ \mathrm{T}z))$$

We can see this is a notational variant of

$$\forall x \forall y\, (\neg\mathrm{Bew}_{\mathsf{PA}}(x) \wedge\ y = \dot{\neg} x \ \rightarrow\ \forall z \forall w\, (z = w \dot{\rightarrow} y \ \rightarrow\ \mathrm{T}z))$$

so the first pair of quantifiers is restricted by the formula $\neg\mathrm{Bew}_{\mathsf{PA}}(x) \wedge\ y = \dot{\neg} x$, and the second pair by $z = w \dot{\rightarrow} y$. As another example, note that the normalization of $\forall x\, (\mathrm{T}x \rightarrow \mathrm{T}x)$ is just $\forall x\, (\neg\mathrm{T}x \vee \mathrm{T}x)$, as none of the steps 1-4 can be applied. The quantifier is, thus, unrestricted.

Proposition 1 *For every formula φ, φ^* is in ADNF and is logically equivalent to φ.*

A similar result can be found in Picollo [21], together with a proof. We are finally in a position to provide an adequate and precise definition of reference by quantification. Let $\mathbf{n}$ abbreviate $n_1, \ldots, n_m \in \omega$, and $\overline{\mathbf{n}}$ abbreviate $\overline{n_1}, \ldots, \overline{n_m}$.

Definition 4 (**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. χ is not a universal statement, 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, 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, and $\chi[\overline{\mathbf{n}}/\mathbf{v}]$ q-refers or newly m-refers to ψ, for some $\mathbf{n} \in \omega$.

Roughly, the definition of q-reference is intended to capture the idea that the q-referents of universally quantified claims are the sentences their (possibly restricted) instances refer to. Note that it can be easily turned into a (fairly cumbersome)

recursive definition that bottoms out in m-reference.[20] In the simple cases, where there are no embedded quantifiers, q-reference is defined purely in terms of the m-referents of the instances, whereas in the presence of embedded quantifiers, the latter may also q-refer. In what follows I offer a variety of examples to illustrate how the definition works and to show that it gives the correct verdicts in paradigmatic cases.

Definition 4 deals adequately with simple cases of restricted quantification. For instance, it entails that $\mathrm{GRfn}_{\mathsf{PA}}$ q-refers just to all theorems of PA. In its normalization, $\forall x\,(\neg \mathrm{Bew}_{\mathsf{PA}}(x) \vee \mathrm{T}x)$, $\forall x$ is followed by a disjunction. Thus, by clause 1, $\mathrm{GRfn}_{\mathsf{PA}}$ q-refers to every sentence each $\mathrm{T}\overline{n}$ newly m- or q-refers to, provided that $\mathbb{N} \vDash \neg\neg \mathrm{Bew}_{\mathsf{PA}}(\overline{n})$, that is, that n codes a theorem of PA. Thus, Definition 4 closely follows clause 2 of the incomplete notion of reference introduced in Section 2.3. What's more, since every formula of $\mathscr{L}_{\mathrm{T}}$ can be normalized, this definition provides a *complete* account of reference by quantification.

Clause 2 deals with unrestrictedly quantified claims, such as $\forall x\,\mathrm{T}x$ and $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \wedge \mathrm{T}x)$. In both cases, the expressions q-refer to whatever each instance, *without restriction*, newly m-refers to, that is, to every sentence.

Complex terms in the scope of the truth predicate are dealt with as expected. Take, for instance, (9) – $\forall x(x = \ulcorner\varphi\urcorner \rightarrow \mathrm{T}\dot{\neg}x)$ – whose normalization is $\forall x\,(x \neq \ulcorner\varphi\urcorner \vee \mathrm{T}\dot{\neg}x)$. Clause 1 guarantees that (9) q-refers just to $\neg\varphi$, for it entails that the sentence q-refers to what $\mathrm{T}\dot{\neg}\ulcorner\varphi\urcorner$ m-refers to. Likewise, clause 2 entails that $\forall x\,\mathrm{T}(x \dot{\rightarrow} \ulcorner 0 = 0\urcorner)$, which is its own normalization, q-refers to whatever each $\mathrm{T}(\overline{n} \dot{\rightarrow} \ulcorner 0 = 0\urcorner)$ m-refers to. Thus, the formula q-refers just to all sentences of conditional form with $0 = 0$ as consequent.

The definition also gives the intuitively right verdict when embedded quantifiers are involved, as in (10) – $\forall x\,(x = \ulcorner\varphi\urcorner \rightarrow \forall y\,(y = \dot{\neg}x \rightarrow \mathrm{T}y))$. Its normalization is $\forall x\,(x \neq \ulcorner\varphi\urcorner \vee \forall y\,(y \neq \dot{\neg}x \vee \mathrm{T}y))$. By clause 1, (10) q-refers to the q-referents of each $\forall y\,(y \neq \dot{\neg}\overline{n} \vee \mathrm{T}y)$, provided that $\mathbb{N} \vDash \neg\overline{n} \neq \ulcorner\varphi\urcorner$, i.e. that n is the code of φ. $\forall y\,(y \neq \dot{\neg}\ulcorner\varphi\urcorner \vee \mathrm{T}y)$, in turn, q-refers to what $\mathrm{T}\overline{n}$ m-refers to, provided that $\mathbb{N} \vDash \overline{n} = \dot{\neg}\ulcorner\varphi\urcorner$, i.e. that n is the code of $\neg\varphi$. As a consequence, (10) q-refers just to $\neg\varphi$, as desired. Similarly, (11) – $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \wedge \forall y\,\mathrm{T}(y \dot{\rightarrow} x))$ – q-refers just to all conditional sentences.

[20] Explicitly: φ q-refers to ψ iff φ and ψ are both sentences and φ^* has a subsentence of the form $\forall \mathbf{v}\chi$ containing T s.t. χ is not a universal statement and either

(i) χ contains no quantifiers and either

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

(ii) χ contains at least one quantifier and either

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

Finally, strings of quantifiers receive an adequate treatment, as the definition instantiates them all at once. For instance, $\forall x \forall y\, \mathrm{T}(x \dot{\rightarrow} y)$ can be easily seen to refer to all sentences of conditional form, by clause 2, whilst (12) q-refers just to all theorems of PA, by clause 1.

Just as the definition of m-reference accounts for the self-referentiality of certain sentences delivered by the Strong Diagonal Lemma, Definition 4 implies that 'weakly' diagonalizing certain predicates also delivers self-referential sentences.[21] Let's look back at the proof of Theorem 1 in Section 2.2. Diagonalization applied to $\neg \mathrm{T}x$ delivers the following fixed point:

$$\forall x\, (x = \ulcorner \forall y\, (\mathrm{Diag}(x, y) \rightarrow \neg \mathrm{T}y) \urcorner \rightarrow \forall y\, (\mathrm{Diag}(x, y) \rightarrow \neg \mathrm{T}y))$$

– dubbed λ – whose normalization is

$$\forall x\, (x \neq \ulcorner \forall y\, (\mathrm{Diag}(x, y) \rightarrow \neg \mathrm{T}y) \urcorner \vee \forall y\, ((\neg \mathrm{Diag}(x, y))^* \vee \neg \mathrm{T}y))$$

By clause 1 of Definition 4, the q-referents of λ are those of

$$\forall y\, (\mathrm{Diag}(\ulcorner \forall y\, (\mathrm{Diag}(x, y) \rightarrow \neg \mathrm{T}y) \urcorner, y)^* \vee \neg \mathrm{T}y)$$

that is, the m-referents of $\neg \mathrm{T}\overline{n}$, provided that n is the code of the result of applying the diagonalization function to $\forall y\, (\mathrm{Diag}(x, y) \rightarrow \neg \mathrm{T}y)$, i.e. λ. In other words, λ q-refers to itself. Similarly, we can conclude that the Curry sentence κ in (4) is self-q-referential, for it obtains by Diagonalization applied to $\mathrm{T}x \rightarrow 0 \neq 0$.

Heck [10] argues to the contrary. They maintain that, unlike m-reference, q-reference is not a genuine kind of reference, for it cannot account for certain pieces of self-referential reasoning. As an example, they claim that 'weak' Diagonalization is not sufficient to imply the existence of a real liar sentence in Kripke's fixed-point theory of truth over the Strong Kleene evaluation scheme; rather, Strong Diagonalization is needed (cf. footnote 8).[22]

The salient features of Kripke's models are that each sentence φ and its truth ascription $\mathrm{T}\ulcorner\varphi\urcorner$ receive the same truth value, and that paradoxical sentences are neither true nor false. In addition, the biconditional in the Strong Kleene evaluation scheme behaves in such a way that $\varphi \leftrightarrow \psi$ is true only if φ and ψ are both true or both false; if either φ or ψ are neither true nor false, so is $\varphi \leftrightarrow \psi$. Thus, as Heck points out, the equivalence $\lambda \leftrightarrow \neg \mathrm{T}\ulcorner\lambda\urcorner$ in (3) delivered by Theorem 1 cannot be true in Kripke's models, on pain of triviality. If it were so, then λ and $\neg \mathrm{T}\ulcorner\lambda\urcorner$ would be either both true or both false, so λ and $\mathrm{T}\ulcorner\lambda\urcorner$ would receive different truth values in the model.

[21] Since our definitions are intended to capture not reference simpliciter but alethic reference, they do not allow us to conclude that every sentence delivered by Theorem 1 is self-referential. For a more general result of this kind, see the definition of reference by quantification in Picollo [21].

[22] Kripke's [17] semantic account consists of a family of partial models $\langle \mathbb{N}, \Gamma^+, \Gamma^- \rangle$ expanding $\mathbb{N}$ to $\mathscr{L}_{\mathrm{T}}$, in which T is assigned not only an extension, Γ^+, but also an anti-extension, Γ^-, and numbers in ω may belong neither to Γ^+ nor to Γ^-. Truth-in-a-model is then defined over the Strong Kleene evaluation scheme that assigns truth values 'true', 'false', and 'neither true nor false' to each sentence of the language in a determinate way. In particular, sentences of the form $\mathrm{T}t$ are true in $\langle \mathbb{N}, \Gamma^+, \Gamma^- \rangle$ if t denotes a number in Γ^+, false if t denotes a number in Γ^-, and neither true nor false otherwise.

From this impossibility Heck concludes that 'real' self-reference cannot be established by means of 'weak' Diagonalization, presumably because they believe that without the equivalence between λ and $\neg \mathrm{T}\ulcorner\lambda\urcorner$ there is no liar sentence. However, it can be easily shown that the way in which, e.g. λ is obtained entails that its truth value must coincide with that of $\neg \mathrm{T}\ulcorner\lambda\urcorner$ in every – classical or non-classical – expansion of $\mathbb{N}$ (and thus, λ is a true liar sentence), forcing λ to be neither true nor false in Kripke's models. Q-reference is a legitimate kind of reference after all.

Together with Definition 1, Definition 4 also allows us to account for the reference pattern that underlies the 2-liar cycle given by λ_1 and λ_2 in (5), that results from diagonalizing the predicate $\neg \mathrm{T}\ulcorner \mathrm{T}\dot{x}\urcorner$. Let λ_1 be the following:

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

Applying clause 1 and the same reasoning as before, λ_1 q-refers to $\mathrm{T}\ulcorner\lambda_1\urcorner$. Let λ_2 be this sentence. Thus, λ_2 m-refers to λ_1 and the latter q-refers to the former.

Another interesting example is the Visser-Yablo list. Diagonalizing the predicate $\forall z\,(z > w \rightarrow \neg \mathrm{T}x(\dot{z}/\ulcorner w\urcorner))$, we obtain the following fixed point:

$$\begin{aligned}&\forall x\,(x = \ulcorner \forall y\,(\mathrm{Diag}(x, y) \rightarrow \forall z\,(z > w \rightarrow \neg \mathrm{T}y(\dot{z}/\ulcorner w\urcorner)))\urcorner \rightarrow \\ &\forall y\,(\mathrm{Diag}(x, y) \rightarrow \forall z\,(z > w \rightarrow \neg \mathrm{T}y(\dot{z}/\ulcorner w\urcorner))))\end{aligned}$$

This predicate is $\mathrm{Y}(w)$. Each instance, $\mathrm{Y}(\overline{n})$, is a sentence that, by clause 1, q-refers to the result of instantiating $\mathrm{Y}(w)$ in each number greater than n, as expected.

Finally, although the fixed point of the predicate $\mathrm{T}\dot{\neg}x$ that Theorem 1 delivers does not q-refer to itself, its negation does, as in the case of the strong fixed point of this predicate in (8).

Unlike m-reference, q-reference is not closed under all valid transformations that preserve atoms or literals (i.e. atomic and negation of atomic subformulae), not even for sentences in ADNF. Consider, for instance, $\forall x\,(x \neq \ulcorner\lambda\urcorner \wedge \mathrm{T}x \wedge \neg \mathrm{T}x)$ and $\forall x\,(x \neq \ulcorner\lambda\urcorner \vee (\mathrm{T}x \wedge \neg \mathrm{T}x))$. These sentences are both logically false, in ADNF, and contain the same literals, but while the former q-refers to every sentence, the latter only q-refers to λ.

However, since q-reference depends exclusively on the presence of certain subformulae of the form $\forall \mathbf{v}\,\varphi$, the notion is compositional with respect to propositional connectives: $\neg\varphi$ q-refers to whatever φ q-refers to, and $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, and $(\varphi \rightarrow \psi)$ q-refer to what either φ or ψ q-refer to. Note also that q-reference is trivially closed under renaming of variables and dummy quantifiers.

More importantly, the fact that q-reference is defined in terms of normalization guarantees that the notion is closed under the following propositional transformations: the addition and elimination of double negations, the commutativity of disjunction and conjunction, contraposition of the conditional, the distributivity of conjunction over disjunction and vice versa, the De Morgan laws, and the interdefinability of connectives and quantifiers. Furthermore, q-reference is closed under Leibniz's Law, and the addition and elimination of tautological antecedents: if ψ is a formula of $\mathscr{L}$ that is true of every n-tuple of natural numbers, $\forall v\,\varphi$ and $\forall v\,(\psi \rightarrow \varphi)$ q-refer to the same sentences.

Closure under these and other transformations allows Definition 4 to give the intuitively correct verdict with respect to existentially quantified statements. For instance,

as $\exists x\ (\mathrm{Bew}_{\mathsf{PA}}(x) \wedge \neg \mathrm{T}x)$ is normalized into $\neg\forall x\ ((\neg\mathrm{Bew}_{\mathsf{PA}}(x))^* \vee \mathrm{T}x)$, it q-refers to the same expressions as $\forall x\ (\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$, i.e. to all theorems of PA, as desired. In general, statements of the form $\exists \mathbf{v}\, \varphi$ q-refer to the same sentences as $\forall \mathbf{v}\, \neg\varphi$. This allows the given definition of q-reference to account for the self-referentiality and other reference patterns of fixed points delivered by the 'existential' proof of Theorem 1, just as in the case of the 'universal' proof (cf. Section 2.2).

3.4 Direct Reference, Reference, and Reference Patterns

In the previous section a completely general account of reference by quantification was provided, extending (an alethic version of) clause 2 of the incomplete notion of reference to all sentences of the language. We are finally in a position to define alethic reference simpliciter, as the disjunction of m- and q-reference. However, I call it "direct reference" or "d-reference", for short, as an indirect reference relation is also possible and relevant to our purposes.

Definition 5 (**D-reference**) Let φ, ψ be sentences. φ d-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* $\psi \vee \varphi$*,*
   - $\varphi \wedge \psi$ *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 5 allows us to characterize many reference patterns, including the self-referentiality of both weak and strong liars, i.e. λ and $\neg\mathrm{Tl}$, and the unfoundedness of the Visser-Yablo sentences, as will be seen later. Nonetheless, there are other cases of intuitive self-reference, such as the circularity of cycles, and other problematic

patterns, such as the unfoundedness of chains, which are not accounted for. Consider, for instance, sentences λ_1 and λ_2 in the 2-liar cycle in (5). They directly refer only to one another. Intuitively, however, they somehow refer to themselves as well, albeit *indirectly*. Otherwise, we would get a semantic paradox without self-reference on the cheap. Furthermore, we can prove the existence of ω-chains, that is, sequences of sentences, each of which directly refers to the expression coming next.

Proposition 2 (**ω-chains**) *For every formula $\varphi(x, \mathbf{v})$ there is an infinite sequence of distinct terms $t_0, \ldots, t_n, \ldots$ s.t., for every $n \in \omega$, the following is provable in* PA*:*

$$t_n = \ulcorner \varphi(t_{n+1}, \mathbf{v}) \urcorner$$

See Picollo [21] for a proof. Note that, since the terms $t_0, \ldots, t_n, \ldots$ are different, so are the sentences $\varphi(t_0), \varphi(t_1), \ldots, \varphi(t_n), \ldots$ and, therefore, the numbers those terms denote. For example, we can obtain an ω-chain for the predicate $\mathrm{T}x$ – a 'truth-teller' chain of sentences $\mathrm{Tt}_0, \mathrm{Tt}_1, \ldots, \mathrm{Tt}_n, \ldots$ – such that the following are provable in PA:

$$\begin{aligned} \mathrm{t}_0 &= \ulcorner \mathrm{Tt}_1 \urcorner \\ \mathrm{t}_1 &= \ulcorner \mathrm{Tt}_2 \urcorner \\ &\ldots \\ \mathrm{t}_n &= \ulcorner \mathrm{Tt}_{n+1} \urcorner \\ &\ldots \end{aligned} \tag{14}$$

Each Tt_n d-refers just to Tt_{n+1}, but intuitively each of them also refers, indirectly, to all the ones coming later on the list. A more general notion of reference, that is, the transitive closure of d-reference, is therefore necessary. To define such a notion, we make use of the following definition.

Definition 6 (**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 7 (**Reference**) Let φ, ψ be sentences. φ refers to ψ iff there's a chain of reference starting with φ and ending with ψ.

It easily follows from the definition of reference that λ_1 and λ_2 refer to themselves, and that each sentence in the truth-teller chain in (14) refers to every expression that comes up later on the list. Moreover, we can employ the notions just introduced to define salient reference patterns, such as the following.

Definition 8 (**Self-reference**) A sentence is self-referential iff it refers to itself.

Definition 9 (**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.

Sentences that don't d-refer to any expression – e.g. every theorem of PA – are well-founded. Moreover, if a sentence d-refers only to well-founded expressions, it is also well-founded, for it extends the chains of reference of the latter only by one sentence. For example, $\forall x\,(\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow \mathrm{T}x)$ is well-founded. 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 (6) and the truth-teller chain in (14). In both cases, sentences on the list refer to all the expressions occurring later on, but never to other statements that occur before them, thus never to themselves.

If the given definitions are correct, the Visser-Yablo sequence shows that the self-reference diagnosis of semantic paradoxes is mistaken after all: there are non-self-referential (ω-)paradoxes – at least if self-reference is taken, as we have been assuming, to be a property of sentences. But this doesn't mean there is no use for the new notions of reference in the quest for the root of paradox or in the formulation of interesting truth theories. Although restricting our truth principles to non-self-referential expressions will not always lead to sound truth systems, restricting them to well-founded sentences will, as I show in the companion piece, [22]. This suggests that it is not self-reference but unfoundedness which is behind the semantic paradoxes.

4 Conclusions

Let's recapitulate. We have seen that the two conceptions of reference and self-reference that are commonly deployed in the literature on semantic paradox and incompleteness are defective or incomplete. Whereas the naïve conception is outright trivial, it was not clear how to turn the incomplete notion, which seems to be somehow on the right track, into a full-fledged account. What's more, a great deal of scepticism surrounded this project. In particular, it has so far been unclear how a notion of self-reference along those lines could account for the self-referential character of 'weak' fixed points – e.g. of λ.

Throughout this paper I have provided a precise account of reference and self-reference via truth that supplements the incomplete notion, extending it to all sentences of $\mathscr{L}_{\mathrm{T}}$. It is my hope that this account dispels some of the doubts affecting the notions of reference and self-reference for formal languages, at least in the context of truth. The notion of reference by quantification introduced in Section 3.3 explains the self-referentiality of many expressions obtained by Diagonalization – e.g. λ – in terms of the way the fixed points are constructed. There is subsequently no need to appeal to a naïve and trivializing account of self-reference according to which it is the equivalence this result yields – e.g. between λ and $\neg\mathrm{T}\ulcorner\lambda\urcorner$ – that is responsible for self-reference.

More generally, unlike the naïve conception of reference, the definitions of m- and q-reference given in Sections 3.2 and 3.3 account for the reference patterns underlying many expressions, including the liar and also liar cycles and Visser-Yablo sentences, without invoking any provable or true equivalences, but attending to the

syntactic structure of these expressions. The new notions seem to answer Milne's worries satisfactorily and, hopefully, help dissipate to some extent the scepticism over reference and self-reference aired by Leitgeb and Cook.

Acknowledgements I am deeply indebted to Volker Halbach, with whom I had countless fruitful discussions on reference and self-reference over the last seven 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. Boolos, G., Burgess, J., Jeffrey, R. (2007). *Computability and logic*, 1edn. Cambridge: Cambridge University Press.
3. Carnap, R. (1937). *Logische Syntax der Sprache*. London: Routledge.
4. Cook, R.T. (2006). There are Non-circular Paradoxes (but Yablo's Isn't One of Them!). *The Monist*, *89*(1), 118–149.
5. Gödel, K. (1931). Über formal unentscheidebarre Sätze der Principia Mathematica und verwandler System I. *Monathshefte für Mathematik und Physik*, *38*, 173–198.
6. Goodman, N. (1961). About. *Mind*, *70*, 1–24.
7. Halbach, V., & Visser, A. (2014). Self-reference in Arithmetic I. *Review of Symbolic Logic*, *7*, 671–691.
8. Halbach, V., & Visser, A. (2014). Self-reference in Arithmetic II. *Review of Symbolic Logic*, *7*, 692–712.
9. Hardy, J. (1995). Is Yablo's Paradox liar-like? *Analysis*, *55*(3), 197–198.
10. Heck, R.K. (2007). Self-reference and the languages of Arithmetic. Philosophia Mathematica III (pp. 1–29). Originally published under the name "Richard G. Heck Jr".
11. Henkin, L. (1952). A problem concerning provability. *Journal of Symbolic Logic*, *17*, 160.
12. Herzberger, H. (1970). Paradoxes of grounding in semantics. *Journal of Philosophical Logic*, *67*, 145–167.
13. Jeroslow, R.G. (1973). Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic*, *38*, 359–367.
14. Ketland, J. (2004). Bueno and Colyvan on Yablo's Paradox. *Analysis*, *64*, 165–172.
15. Ketland, J. (2005). Yablo's Paradox and ω-inconsistency. *Synthese*, *145*, 295–307.
16. Kreisel, G. (1953). On a problem of Henkin's. *Indagationes Mathematicae*, *15*, 405–406.
17. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosphy*, *72*, 690–716.
18. Leitgeb, H. (2002). What is a self-referential sentence? CRITICAL remarks on the alleged (non)-circularity of Yablo's Paradox. *Logique et Analyse*, *177–178*, 3–14.
19. Milne, P. (2007). On Gödel sentences and what they say. *Philosophia Mathematica*, *III*(15), 193–226.
20. Montague, R. (1962). Theories incomparable with respect to relative interpretability. *Journal of Symbolic Logic*, *27*, 195–211.
21. Picollo, L. (2018). Reference in Arithmetic. *Review of Symbolic Logic*, *11*, 573–603.

22. Picollo, L. Reference and truth. *Journal of Philosophical Logic* (to appear).
23. Priest, G. (1997). Yablo's Paradox. *Analysis*, *57*, 236–242.
24. Putnam, H. (1958). Formalization of the Concept of "About". *Philosophy of Science*, *25*, 125–130.
25. Putnam, H. (1980). Models and Reality. *Journal of Symbolic Logic*, *45*, 464–482.
26. Smoryński, C. (1991). The development of self-reference: Löb's theorem. In Drucker, T. (Ed.) *Perspectives on the history of mathematical logic* (pp. 110–133). Boston: Birkhäuser.
27. Urbaniak, R. (2009). Leitgeb, "About," Yablo. *Logique et Analyse*, *207*, 239–254.
28. 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.
29. Yablo, S. (1985). Truth and reflexion. *Journal of Philosphical Logic*, *14*, 297–349.
30. Yablo, S. (1993). Paradox without self-reference. *Analysis*, *53*, 251–252.
31. Yablo, S. (2014). *Aboutness*. Princeton: Princeton University Press.