THE REVIEW OF SYMBOLIC LOGIC
Volume 11, Number 3, September 2018

REFERENCE IN ARITHMETIC

LAVINIA PICOLLO
Ludwig-Maximilian University of Munich

Abstract. Self-reference has played a prominent role in the development of metamathematics in the past century, starting with Gödel's first incompleteness theorem. Given the nature of this and other results in the area, the informal understanding of self-reference in arithmetic has sufficed so far. Recently, however, it has been argued that for other related issues in metamathematics and philosophical logic a precise notion of self-reference and, more generally, reference is actually required. These notions have been so far elusive and are surrounded by an aura of scepticism that has kept most philosophers away. In this paper I suggest we shouldn't give up all hope. First, I introduce the reader to these issues. Second, I discuss the conditions a good notion of reference in arithmetic must satisfy. Accordingly, I then introduce adequate notions of reference for the language of first-order arithmetic, which I show to be fruitful for addressing the aforementioned issues in metamathematics.

§1. To prove his famous first incompleteness result for arithmetic, Gödel [5] developed a technique called "arithmetization" or "gödelization". It consists in codifying the expressions of the language of arithmetic with numbers so that the language can 'talk' about its own formulae. Then, he constructed a sentence in the language that he described as stating its own unprovability in a system satisfying certain conditions[1] and showed this sentence to be undecidable in the system. His method led to enormous progress in metamathematics and computer science and also in philosophical logic and other areas of philosophy where formal methods became popular. Let's take a closer look.

Let $\mathscr{L}$ be the language of first-order Peano arithmetic (PA). $\mathscr{L}$ contains $=$, $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\forall$, and $\exists$ as logical constants, 0 as the only individual constant, S as a monadic function symbol, $+$ and $\times$ as dyadic function symbols, and a stock of extra function symbols for recursive functions to be specified. All other logical and non-logical symbols are taken to be the usual abbreviations. We assume that PA contains definitions for each extra function symbol. $\mathbb{N}$ is the standard model for $\mathscr{L}$, with ω as its domain. $\mathscr{L}$ is the main formal language we will work with in this paper. Unless otherwise indicated, all symbols and formulae we use or mention belong to $\mathscr{L}$.

The individual term consisting of n occurrences of S followed by 0 is called the "numeral" of n. We denote it by $\overline{n}$. If σ is a string of symbols, $\#\sigma$ is its code or Gödel number and $\ulcorner\sigma\urcorner$ the numeral of its code. To avoid certain difficulties nonstandard codings can lead to (cf. Heck [12], Halbach and Visser [10, 11]), I assume a *fixed* effective and

Received: January 4, 2017.
2010 *Mathematics Subject Classification*: 03-XX.
Key words and phrases: reference, arithmetic, self-reference, diagonalization.

[1] More specifically, the system must be 1-consistent. A theory formulated in the language of arithmetic is ω-consistent if and only if it doesn't entail every numerical instance of a formula and, at the same time, its universal closure. The notion of 1-consistency is that of ω-consistency restricted to Σ_1-formulae (cf. §2).

doi:10.1017/S1755020317000351

monotonic coding. By "effective" I mean that given a number n there is an algorithm to determine which expression it codifies and, vice versa, given an expression σ there is an algorithm that delivers the code of σ. By "monotonic" I imply that if σ is a subexpression of σ', then $\#\sigma \leq \#\sigma'$. For perspicuity, when there's no room for confusion I will talk about expressions of $\mathscr{L}$ when what is really meant is their codes under our fixed coding.

Let $\mathsf{Bew}(x)$ be a formula defining and weakly representing provability in PA in PA:[2] for any sentence φ, $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ is provable in PA iff φ is a theorem as well. Gödel showed there is a sentence γ of $\mathscr{L}$ such that the following equivalence is a theorem of PA:

$$\gamma \leftrightarrow \neg\mathsf{Bew}(\ulcorner\gamma\urcorner). \tag{1}$$

γ, known nowadays as PA's "Gödel sentence", was characterized by Gödel himself in the following terms:[3] "We thus have a sentence before us that states its own unprovability."

Carnap [2] generalized Gödel's construction to any formula with one free variable and proved what today is known as the "diagonalization" or "diagonal lemma".[4] This result can be obtained already in Robinson arithmetic, Q—i.e., PA without induction, plus the axiom $\forall x(x \neq 0 \rightarrow \exists y(x = Sy))$. For all recursive functions are strongly representable in Q.

THEOREM 1.1 (Diagonalization). *For every formula $\varphi(x)$ there is a sentence ψ such that the following equivalence is a theorem of* Q:

$$\psi \leftrightarrow \varphi(\ulcorner\psi\urcorner). \tag{A}$$

Proof. Let $\mathsf{Diag}(x, y)$ strongly represent the primitive recursive (p.r.) function "diagonalization" that takes the code x of a formula $\varphi(x)$ and returns the code y of $\forall x(x = \ulcorner\varphi\urcorner \rightarrow \varphi)$ in Q.

$$\forall x(x = \ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner \rightarrow \forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))) \tag{B}$$

is the result of applying the diagonalization function to $\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))$. Notice that (B) is the ψ we were looking for. Let n be the Gödel code of (B). By the laws of identity, (B) is logically equivalent to

$$\forall y(\mathsf{Diag}(\ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner, y) \rightarrow \varphi(y)), \tag{2}$$

which is equivalent in Q to $\varphi(\overline{n})$. Thus,

$$\mathsf{Q} \vdash \forall x(x = \ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner \rightarrow \forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))) \leftrightarrow \varphi(\overline{n}). \quad \square$$

[2] Recall that a formula $\varphi(x_1, \ldots, x_n)$ defines the relation $R \subseteq \omega^n$ if and only if $\varphi(\overline{k}_1, \ldots, \overline{k}_n)$ is true in $\mathbb{N}$ iff $\langle k_1, \ldots, k_n\rangle \in R$. If, additionally, $\varphi(\overline{k}_1, \ldots, \overline{k}_n)$ is provable in $\mathsf{Th} \subseteq \mathscr{L}$ iff $\langle k_1, \ldots, k_n\rangle \in R$, we say that φ weakly represents R in Th. Finally, if it's also the case that $\neg\varphi(\overline{k}_1, \ldots, \overline{k}_n)$ is provable in Th iff $\langle k_1, \ldots, k_n\rangle \notin R$, we say that φ (strongly) represents R in Th.

[3] The original, in German, reads: "Wir haben also einen Satz vor uns, der seine eigene Unbeweisbarkeit behauptet." (Gödel [5, p. 175]) The English translation is borrowed from Halbach and Visser [10, p. 671].

[4] There is a more general version of this result due to Montague [23], for formulae containing an arbitrary number of free variables. For the purposes of this paper Carnap's version is general enough. A stronger version of diagonalization will be introduced later in §3 (cf. Theorem 3.1), in which any number of free variables is allowed to occur in φ. For more details on the history of diagonalization, see Smoryński [27].

This is the 'universal proof' of the diagonal lemma. A similar proof I call "existential" can be given in terms of an alternative diagonalization function strongly represented by $\mathsf{Diag}^{\exists}(x, y)$, mapping φ to $\exists x(x = \ulcorner\varphi\urcorner \wedge \varphi)$, and then diagonalizing the predicate $\exists y(\mathsf{Diag}^{\exists}(x, y) \wedge \varphi(y))$ to obtain ψ. This will become relevant later.

Equivalences of the form (A) are known as "diagonal" sentences. Following Gödel, every sentence ψ provably satisfying (A), also known as a provable "fixed point" of φ, is commonly regarded as saying of itself that it has the property expressed by φ (whatever that is). *A fortiori*, all fixed points ψ are considered to be self-referential, and the diagonalization lemma is seen as the paradigmatic mechan ism for self-reference in arithmetic. I call this the "naïve view of self-reference".

Naïve view of self-reference: A sentence ψ refers to itself and says of itself that it has the property expressed by the formula $\varphi(x)$, just in case $\psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$ is provable in Q.

This view involves *extensional* conceptions both of what it means for a sentence ψ to say of itself that it has the property expressed by φ and of self-reference *simpliciter*.

Like most naïve notions in philosophical logic, naïve self-reference is trivial. As noted by Leitgeb [19], every sentence ψ is provably equivalent to a sentence of the form $\varphi(\ulcorner\psi\urcorner)$ by logic alone.[5] Take $\varphi(\ulcorner\psi\urcorner)$ to be, for instance, $\ulcorner\psi\urcorner = \ulcorner\psi\urcorner \wedge \psi$. However, the triviality of naïve self-reference *simpliciter* does not carry over the naïve conception of what it means for a sentence ψ to say of itself that it has the property expressed by $\varphi(x)$, which is at the heart of Gödel's construction. To give an example, not every sentence ψ is provably equivalent to $\mathsf{Bew}(\ulcorner\psi\urcorner)$.

Gödel's construction inspired Kleene's [17] recursion theorem, which led to enormous progress in computability. It also had a great influence on investigations on truth and related notions in philosophical logic, prominently on the work of Tarski [30, 31], but also in other areas of philosophy that work with sentential predicates (e.g., knowledge in epistemology, grounding in metaphysics), and, of course, in metamathematics. A salient case of the latter is Löb's [21] theorem, which establishes that only trivial instances of soundness (i.e., $\mathsf{Bew}(\ulcorner\varphi\urcorner) \rightarrow \varphi$) are available in arithmetical theories (if $\mathsf{Bew}(x)$ satisfies certain conditions; cf. Theorem 2.1).

Despite the triviality of naïve self-reference, to prove Gödel's and Löb's results the naïve conception of what it means for a sentence ψ to say of itself that it has the property expressed by $\varphi(x)$, that is, the availability of equivalences of the form (A), suffices. The same can be said about other related phenomena in metamathematics. As Smoryński [28] suggests, this, together with the triviality of naïve self-reference, appears to be the main reason why not many philosophers and almost no mathematicians have been really interested in the notion of self-reference in arithmetic *per se*, with the exception of Kreisel. However, Halbach and Visser [10, 11] have recently shown that there are other issues and questions in metamathematics that, unlike Gödel's and Löb's, cannot even be properly formulated in terms of the naïve conception but call for a rather intensional understanding of self-reference. Moreover, Leitgeb [19] has argued that the debate about whether all semantic paradoxes involve self-reference of some sort goes adrift unless we have a proper, nontrivial notion of self-reference *simpliciter* for the language of arithmetic extended with a truth predicate. This debate originated in the Visser-Yablo paradox, an infinitary semantic paradox in which there is *prima facie* no self-reference involved (cf. Visser [33], Yablo [34, 35]).

[5] Cook [3] and Heck [12] make similar points.

The main purpose of this paper is, nonetheless, to provide a sound and precise definition of reference or aboutness *just for the language of first-order arithmetic*. Other languages or extensions of $\mathscr{L}$ with new primitive predicate symbols such as the truth predicate are the subject of further work (cf. [author]). The resulting notion will help us define salient reference patterns like self-reference, non-well-foundedness, loops, etc. Furthermore, it will give the intuitively right verdict on diagonal sentences obtained via diagonalization and will overcome the difficulties of the naïve notion and other previous attempts to define reference and self-reference for formal languages. As a consequence, the notions I introduce will prove themselves useful to properly formulate the metamathematical problems that Halbach and Visser mention. They will also serve as a blue print for future work on underlying reference patterns of sentences in languages with a truth predicate. This notion, in turn, could help us give a definite answer to the questions whether the Visser-Yablo paradox involves some kind of self-reference and whether all semantic paradoxes do so as well.

This paper is organized as follows. I first present the examples of Halbach and Visser and explain why the naïve understanding of what it means for a sentence to say of itself that it has a certain property cannot account for them. In §3 I list some desiderata for every notion of reference and, therefore, self-reference for the language of arithmetic. §4 gives new definitions of reference, self-reference, and well-foundedness, evaluates their pros and cons, and proves several results that show the notions are adequate from a material point of view. Finally, I indicate how the new notions could be used to provide exact formulations of the examples given by Halbach and Visser.

§2. In this section I show that certain problems in metamathematics require a more fine-grain view on self-reference than the naïve conception introduced in the previous section, even to be properly formulated. Both examples are taken from the studies of Halbach and Visser [10, 11]. The first one is the question over the provability, refutability, or undecidability of 'Henkin' sentences formulated with Rosser's provability predicate. The second example is the question over the status of truth tellers.

Gödel has shown that under normal circumstances a sentence asserting its own unprovability in PA is undecidable in this system. So one might also wonder about a sentence that states its own provability instead. Is it provable, refutable, or undecidable? This question is usually known as "Henkin's problem", and sentences asserting their own provability are known nowadays as "Henkin sentences". In Henkin's [13, p. 160] own words:

> If Σ is any standard formal system adequate for recursive number theory, a formula (having a certain integer q as its Gödel number) can be constructed which expresses the proposition that the formula with Gödel number q is provable in Σ. Is this formula provable or independent in Σ?

I take Σ to be PA. Let $\mathsf{Bew}(x)$ be as before. According to the naïve view of self-reference, a Henkin sentence would be any sentence η satisfying the equivalence

$$\eta \leftrightarrow \mathsf{Bew}(\ulcorner \eta \urcorner). \tag{3}$$

Note however that, by the weak representability requirement, any theorem of PA satisfies this equivalence. For instance, since $0 = 0$ is a theorem, $\mathsf{Bew}(\ulcorner 0 = 0 \urcorner)$ is so too and, therefore, $0 = 0 \leftrightarrow \mathsf{Bew}(\ulcorner 0 = 0 \urcorner)$ is provable as well. Thus, this 'Henkin sentence' is decidable and provable.

Henkin was most likely very much aware of this fact and didn't consider it as an answer to his question. As Smoryński [28, p. 114] puts it, "Henkin did not want to know if some

sentence accidentally equivalent to the assertion of its own provability was provable"; he did not mean to inquire about just any fixed point of the provability predicate. In a very clear sense, $0 = 0$ doesn't say of itself that it is provable. It is neither self-referential nor a Henkin sentence.

Nonetheless, in [21] Löb put forward a solution to Henkin's problem that shows that no matter what fixed point of the form (3) we consider, η will always be provable. This is the renowned Löb's theorem.

THEOREM 2.1 (Löb). *Let φ, ψ be sentences and $\mathsf{Bew}(x)$ satisfy Löb's derivability conditions in PA, that is,*

$$\mathsf{PA} \vdash \varphi \;\Rightarrow\; \mathsf{PA} \vdash \mathsf{Bew}(\ulcorner\varphi\urcorner),$$
$$\mathsf{PA} \vdash \mathsf{Bew}(\ulcorner\varphi\urcorner) \wedge \mathsf{Bew}(\ulcorner\varphi \rightarrow \psi\urcorner) \rightarrow \mathsf{Bew}(\ulcorner\psi\urcorner),$$
$$\mathsf{PA} \vdash \mathsf{Bew}(\ulcorner\varphi\urcorner) \rightarrow \mathsf{Bew}(\ulcorner\mathsf{Bew}(\ulcorner\varphi\urcorner)\urcorner).$$

If $\mathsf{PA} \vdash \mathsf{Bew}(\ulcorner\varphi\urcorner) \rightarrow \varphi$, then $\mathsf{PA} \vdash \varphi$ as well.

Let $\mathsf{Bew}(x)$ in (3) satisfy Löb's derivability conditions. If we can prove (3) in PA for any sentence η, we have *a fortiori* that $\mathsf{PA} \vdash \mathsf{Bew}(\ulcorner\eta\urcorner) \rightarrow \eta$ and, by Löb's result, that η is a theorem of PA. As a consequence, even if a more sophisticated view on self-reference is needed to do justice to Henkin's formulation of his problem, the answer can be perfectly given without such notion, if (3) is provable in PA and $\mathsf{Bew}(x)$ satisfies Löb's derivability conditions.

But what if (3) were true in $\mathbb{N}$ yet unprovable in PA? Certainly what matters here is not the provability of a fixed point in this or that system but that the equivalence between η and $\mathsf{Bew}(\ulcorner\eta\urcorner)$ actually holds. In that case it would seem η is intuitively self-referential. Moreover, Henkin's formulation of the problem doesn't exclude this possibility. If a highly complex mechanism for self-reference is used, Löb's theorem wouldn't be able to give an answer to Henkin's question.[6]

Löb's derivability conditions seem to be natural principles for provability, and so they are often considered as meaning postulates. In fact, the first one is one direction of weak representability, which can be seen as another criterion for the expressibility of provability (and other notions), due to Kreisel [18]. However, other provability predicates in the latter sense—that is, those that weakly represent provability in PA—that do not satisfy Löb's conditions have also played a role in the literature. One important case is Rosser's.

Let $\mathsf{Prf}(x, y)$ represent the recursive relation between a sequence of sentences x and a sentence y such that x constitutes a proof of y in PA in a natural way (cf. Halbach and Visser [10]). The standard provability predicate is usually defined in $\mathscr{L}$ as $\exists y\mathsf{Prf}(y, x)$. This predicate satisfies Löb's derivability conditions. Rosser-provability, on the other hand, is defined as follows:

$$\mathsf{Bew}^{\mathsf{R}}(x) := \exists y(\mathsf{Prf}(y, x) \wedge \forall z < y \neg\mathsf{Prf}(z, \dot{\neg}x)),$$

where $\dot{\neg}$ is a function symbol of $\mathscr{L}$ representing the recursive function that maps sentences into their negations. Intuitively, a sentence φ is Rosser-provable if there is a proof of it in

[6] For instance, one could define an alternative diagonalization procedure based on an alternative diagonalization function defined by $\mathsf{Diag}'(x, y)$, that maps sentences φ to $\forall x(x = \ulcorner\varphi\urcorner \wedge \gamma \rightarrow \varphi)$, where γ is PA's Gödel sentence, as before. $\mathsf{Diag}'(x, y)$ is satisfied exactly by the same ordered pairs of natural numbers than $\mathsf{Diag}(x, y)$ in the standard model. But while the latter predicate strongly represents the function it defines, the former doesn't even weakly represent its corresponding function.

PA and there is no proof of $\neg\varphi$ with a smaller code. $\mathsf{Bew}^{\mathsf{R}}(x)$ does not satisfy Löb's conditions for, as Halbach and Visser [11, obs. 7.1] point out, it has both provable and refutable fixed points. For instance, both $0 = 0$ and $0 \neq 0$ are fixed points of $\mathsf{Bew}^{\mathsf{R}}(x)$. Thus, unlike the case for standard provability, if there was a sentence that truly asserted its own Rosser-provability, there would be no trivial answer to the question over its status. Note that, as long as Löb's conditions constitute meaning postulates for provability, these 'Henkin' sentences formulated in terms of Rosser's provability predicate do not really say of themselves that they are provable, but something else. They are not Henkin sentences in the original sense. Call them "Henkin-Rosser" sentences.

One might feel inclined to believe that the naïve view on self-reference is the only kind of view on self-reference we can have in arithmetic, as Cook [3] seems to suggest. Since "the notion of stating one's own provability in the original question cannot be eliminated by the notion of being a fixed point" (Halbach and Visser [10, p. 672]), the question about the status of Henkin-Rosser sentences would be ill-posed. There would not be such thing as a sentence that asserts its own Rosser-provability. On the other hand, one can think, with Henkin, that a better understanding of self-reference for the language of arithmetic is possible, a notion that would make sense of Henkin's problem for Rosser's provability predicate. What Henkin probably had in mind was a sentence that is obtained by a procedure like the one we followed in the proof of Theorem 1.1, but certainly not just that particular one. In this paper I show that a better notion of self-reference for the language of arithmetic is in fact possible. If this notion rather than the naïve one is employed, one can actually make sense of the idea of a Henkin-Rosser sentence, and the question about the status of these expressions becomes a sensible one to be asked.

I now turn to the status of truth tellers in arithmetic. A truth teller is a sentence that states its own truth. Although arithmetic cannot contain its own truth predicate on pain of triviality, as Tarski's theorem on the undefinability of truth shows (cf. Tarski [30]), it does contain partial truth predicates for sentences with limited quantifier complexity.

Formulae in $\mathscr{L}$ can be classified according to their quantifier complexity into sets Σ_n and Π_n as follows. If φ is logically equivalent to a formula where all quantifiers are bounded, φ is both Σ_0 and Π_0. If φ is logically equivalent to a formula consisting of a block of universal quantifiers (possibly of length 1) followed by a Σ_n-expression, then $\varphi \in \Pi_{n+1}$. And if φ is logically equivalent to the negation of a Π_n-formula, then $\varphi \in \Sigma_n$. Note that the sets in the hierarchies Π_n and Σ_n are cumulative, for it's always possible to add superfluous quantifiers in front of a formula.

For every n, $\mathscr{L}$ contains predicates $\mathsf{T}_{\Pi_n}(x)$ and $\mathsf{T}_{\Sigma_n}(x)$ defining the sets of Π_n and Σ_n true sentences.[7] Moreover, we can choose $\mathsf{T}_{\Pi_n}(x)$ for $n \neq 1$ and $\mathsf{T}_{\Sigma_n}(x)$ such that they belong to Π_n and Σ_n, respectively. This means that the sentences that say of themselves that they are Π_n- ($n \neq 1$) and Σ_n-true, however they are obtained, are themselves Π_n and Σ_n, respectively. Thus, we have Π_n- and Σ_n-truth tellers in the language. In most cases partial truth predicates cannot weakly represent the set of corresponding truths, for this set is often too complex. Besides defining their corresponding sets of Π_n- and Σ_n-truth-in-$\mathbb{N}$, the reason why they are called "truth predicates" is that they satisfy the relevant meaning postulates, namely, the T-schema (the equivalence between a sentence and its

[7] See, Kaye [16] or Hájek and Pudlák [7] for details on how to obtain partial truth predicates in PA. I follow Kaye for the most part, except I allow sets Π_n and Σ_n in the truth definitions to be closed under logical equivalence, so every formula in $\mathscr{L}$ belongs to some set in the hierarchy. This implies that, unlike all other partial truth predicates, $T_{\Pi_1}(x)$ is not in Π_1 but only in Π_2.

truth ascription) in PA,

$$\mathsf{PA} \vdash \mathsf{T}_{\Pi_n}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

for each sentence $\varphi \in \Pi_n$ and

$$\mathsf{PA} \vdash \mathsf{T}_{\Sigma_n}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

for each sentence $\varphi \in \Sigma_n$.

As a consequence, each truth predicate has both provable and refutable fixed points. For instance, $0 = 0$ and $0 \neq 0$ are fixed points of every partial truth predicate. In some cases, the predicates can also have undecidable fixed points, like $T_{\Pi_1}(x)$ has γ, PA's Gödel sentence. Therefore, the answer to the question whether Π_n- and Σ_n-truth tellers are provable, refutable, or undecidable doesn't make sense if we turn to the naïve view on self-reference. Every sentence in Π_n (Σ_n) would say of itself that it is Π_n-(Σ_n-)true according to this view. The question over the status of truth tellers in PA requires a more precise definition of what it means for a sentence to say of itself that it has a determinate property, a proper understanding of self-reference that improves on the naïve one.

Setting aside the issues in metamathematics, there are also philosophical reasons why it would be good to have a better notion of self-reference for the language of arithmetic. Formal theories of truth are often formulated in an extension of $\mathscr{L}$ with a truth predicate. As Tarski shows, if instances of the T-schema for certain sentences containing this truth predicate are provable in a system, the system turns out to be unsound. Such sentences are considered paradoxical. The goal of most classical truth-theorists is to identify these sentences, so to exclude their corresponding instances of the T-schema from their theories.[8] Until recently the idea that every paradoxical expression involves some (relevant) kind of self-reference was widely accepted. If true, self-referentiality could be used as a restriction on instances of the T-schema, to avoid triviality.

The Visser-Yablo paradox challenged this view. Roughly, it consists of an infinite list of sentences, each of which says of all the ones coming later on the list that they are untrue. From the assumption that any of these sentences gets a classical truth value, a contradiction can be informally obtained.[9] This antinomy gave rise to a lively debate that evidenced the lack of proper notions of reference and self-reference for the language of arithmetic and its extensions, as Leitgeb [19] pointed out.[10] Without these notions, neither the referential status of sentences in the Visser-Yablo paradox nor the thesis that all paradoxes are self-referential can be adequately assessed. This represents an obstacle to the development of formal (classical) truth systems.

In this paper I focus only on notions of reference and self-reference for $\mathscr{L}$. They will help us, for instance, giving proper formulations of the metamathematical problems introduced in this section, that is, the questions about the status of Henkin-Rosser sentences and

[8] See, e.g., Horwich [15] and Halbach [8].

[9] The antinomy was introduced by Yablo [34] in 1985 and, independently, by Visser [33] in 1989, though early drafts of the latter circulated in the early 1980s (cf. Halbach [9]). While Yablo formulates the sentences in the sequence with one single untyped truth predicate, Visser's version of the paradox is formulated in an illfounded linearly ordered hierarchy of typed truth predicates. Each sentence on the list uses a truth predicate that applies only to sentences containing truth predicates that belong to strictly lower stages in the hierarchy, as Tarski required. Thus, unlike Yablo's version, Visser's is not only intended to show that paradox is possible in the absence of self-reference, but also in the presence of typed truth.

[10] See, for instance, Priest [24], Sorensen [29], and Cook [3].

of Π_n- and Σ_n-truth tellers. However, the new definitions will also serve as blueprints for defining similar concepts for the language expanded with a truth predicate. I hope the new notions can shed some light on the reference and self-reference of expressions in natural language as well.

§3. Every sensible notion of self-reference should be definable in terms of a notion of reference as follows: a sentence is self-referential if and only if it refers to itself. Otherwise it wouldn't be clear why it is self-reference what we are talking about. For instance, if according to the naïve concept of self-reference a sentence φ self-refers in case it's provably equivalent to another sentence $\chi(\ulcorner\varphi\urcorner)$ that mentions φ, that is because the following notion of 'naïve' reference is operating in the background: sentence φ refers to sentence ψ in case there's a sentence of the form $\chi(\ulcorner\psi\urcorner)$ φ is provably equivalent to (which is not very plausible, since whether a sentence refers to a given object has nothing to do with provable equivalents of that sentence). Therefore, to arrive at a successful definition of self-reference the most natural way to go is to devise a good notion of reference first. The purpose of this section is to evaluate what conditions such a notion should satisfy and to understand what kind of concept we are after. §4 will provide a precise definition of reference and other kin notions, in accordance with the results obtained in this section.

To begin with, note that the concept of reference we are after is a relation between sentences. It isn't a relation between terms and objects, or sentences and truth values, as reference has been traditionally understood. It neither relates sentences to numbers. One of the main goals of this paper is to obtain a precise definition of self-reference. Thus, even though the language of arithmetic was originally designed to talk about numbers, and so we can take numbers to be its primary objects, I will focus on the sentences these numbers codify instead. If, for other purposes, the reader is interested in reference to numbers, it shouldn't be hard to obtain a definition of this closely related concept by performing straightforward modifications on the definitions of reference I provide in the next section.

A word of caution is needed about reference to sentences via their codes. The choice of coding, even effective and monotonic ones, is always arbitrary to some extent. As a consequence, what sentences an expression refers to is also very often an arbitrary matter. If we change the coding at play, most sentences will refer to different expressions. Thus, it only makes sense to talk about reference as a relation between sentences once a particular coding has been fixed, as I have done here.

Hopefully, it's now clear that the naïve notion of self-reference must be abandoned. Diagonal sentences, that is, sentences of the form

$$\psi \leftrightarrow \varphi(\ulcorner\psi\urcorner), \tag{A}$$

whether provable in PA or just true in $\mathbb{N}$, should not be enough to conclude that ψ is self-referential, on pain of triviality. As stated in §1, for every sentence ψ there is a predicate φ such that (A) is provable in Q. *A fortiori*, the provable material equivalence

$$\varphi \leftrightarrow \chi(\ulcorner\psi\urcorner) \tag{4}$$

cannot suffice to infer that φ refers to ψ. To conclude this, additional or perhaps just *other conditions* should be met. Good notions of reference and self-reference must have certain *intensional* aspects, as Halbach and Visser [10, 11] maintain.

According to Leitgeb this is problematic. From his point of view, if diagonal sentences are not enough for self-reference,

> [...] no philosopher may any longer argue in the following way: "By Gödel's diagonalization lemma, we know that there is a sentence φ such that φ is equivalent to '$\neg T(\ulcorner\varphi\urcorner)$' in arithmetic. Thus there is a self-referential sentence, that is, φ." (Leitgeb [19, p. 9])

In other words, we wouldn't be able to capture the paradigmatic cases of self-reference the diagonal lemma delivers. However, this remains to be seen. Given a formula $\varphi(x)$, the diagonal lemma does not deliver just *any* sentence ψ satisfying (A), but one of a special kind. It is possible that, due to the particular features ψ exhibits, the conditions a sentence should meet to truly refer to another sentence allow us to infer that ψ is self-referential. The question is what these conditions are. Milne [22, p. 212] shares this concern:

> 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.

In this section I make this case, that is, I specify the conditions a good notion of reference for $\mathscr{L}$ should satisfy and show how, if these conditions were met, diagonal sentences that are obtained via diagonalization turn out to be self-referential. For the most part I follow Leitgeb [19]. He as well considers several desiderata for a notion of reference for formal languages. However, he arrives at a pessimistic conclusion, for his desiderata are mutually incompatible. Later in this section I argue against one of them, the so-called "equivalence condition". In the next section I show that the remaining conditions are simultaneously satisfiable and, therefore, compatible with each other.

As Leitgeb points out, a sentence can refer to an object—in our case another sentence—in two different ways. On the one hand, it can mention this object, that is, contain a term that denotes the object. On the other hand, the sentence can quantify over that object.[11] According to the first way in which sentences could refer,

(C1): a sentence φ refers to a sentence ψ in case a term denoting the code of ψ occurs in φ.

I call this kind of reference "reference by mention", or "m-reference" for short. It is more demanding than the naïve view on reference: for φ to refer to ψ it's not merely required that φ is (provably) equivalent to a sentence of the form $\chi(\ulcorner\psi\urcorner)$, but it should be *identical* to $\chi(\ulcorner\psi\urcorner)$. If φ simply *is* $\chi(\ulcorner\psi\urcorner)$, then (4) follows trivially, for it's just an instance of the tautological schema $\delta \leftrightarrow \delta$. However, (C1) is just a *sufficient* condition for reference; it doesn't exclude other ways in which sentences might refer to other sentences.

We can use (C1) to give a sufficient condition on self-reference, along the following lines:

(C2): a sentence φ is self-referential if it contains a term that denotes φ.

This kind of self-reference, which I call "self-reference by mention" or "m-self-reference" for short, is certainly possible if $\mathscr{L}$ contains a term $\dot{d}(x)$ defining the p.r. function d called "strong diagonalization",[12] which I assume it does. The result is known as the "strong diagonalization lemma" or "strong diagonal lemma". Given a formula φ with x

[11] Preliminary versions of this distinction can be traced back to Ryle [26].

[12] Other similar functions could also do the job. What follows is indifferent to the choice of the function we make.

free, d returns the formula that results from replacing x in φ with $\ulcorner\varphi\urcorner$. Since Q contains definitions for each function symbol in $\mathscr{L}$ other than S, $+$, and $\times$, $\dot{d}(x)$ represents d in Q. Let $\vec{v}$ abbreviate $v_1, \ldots, v_n$, a (possibly empty) sequence of individual variables.

THEOREM 3.1 (Strong diagonalization). *For every formula $\varphi(x, \vec{v})$, where x is different from each $v_1, \ldots, v_n$, there is a term t such that $\mathsf{Q} \vdash t = \ulcorner\varphi(t, \vec{v})\urcorner$.*[13]

Proof. We can prove in Q that $\dot{d}(\ulcorner\varphi(\dot{d}(x), \vec{v})\urcorner) = \ulcorner\varphi(\dot{d}(\ulcorner\varphi(\dot{d}(x), \vec{v})\urcorner), \vec{v})\urcorner$. Let t be $\dot{d}(\ulcorner\varphi(\dot{d}(x), \vec{v})\urcorner)$. □

For each $\varphi(x, \vec{v})$, the identity statement $t = \ulcorner\varphi(t, \vec{v})\urcorner$ delivered by the strong diagonal lemma is often called a "strong diagonal sentence", and $\varphi(t, \vec{v})$, standing on the right-hand side, is often called a "strong fixed point" of φ, as opposed to the 'weak' diagonal sentences (*equivalences* of the form (A)) and the 'weak' fixed points 'weak' diagonalization (Theorem 1.1) delivers. Since identities are stronger requirements than equivalences, (C2) is more demanding than the naïve understanding of self-reference. For instance, (C2) does not allow us to conclude that $0 = 0$ is self-referential, despite the fact that $0 = 0$ is equivalent to $\mathsf{Bew}(\ulcorner 0 = 0\urcorner)$ in PA, for $0 = 0$ does not contain a term denoting itself (because under monotonic codings, 0 isn't the code of a sentence). Thus, we cannot conclude it's a true Henkin sentence.

Note that if a different coding had been chosen, the strong diagonal lemma would still hold, except that the terms it delivers would be different ones. If we vary the code of $\varphi(x, \vec{v})$, then the numeral $\ulcorner\varphi(x, \vec{v})\urcorner$ of its code varies along. As a consequence $\dot{d}(\ulcorner\varphi(x, \vec{v})\urcorner)$ will be a different term. Thus, whether or not a sentence is self-referential also depends on the particular coding that is being used and not only on the structure of the sentence. Since we are working with a fixed (effective and monotonic) but arbitrary coding, it is often the case that we cannot pin down the exact expressions a sentence refers to or the exact terms that occur in it. However, we are often in a position to make true *structural* claims about the reference patterns of formulae, as in Theorem 3.1. Actually, most of the claims about reference in this paper are of this sort.

(C2) is at the base of what Henkin and Kreisel seemed to have in mind in their paper exchange on the status of Henkin sentences: a sentence $\varphi(t)$ says of itself that it has the property expressed by $\varphi(x)$ if and only if t is a closed term denoting $\varphi(t)$ (cf. Henkin [13, 14] and Kreisel [18]). Thus, Halbach and Visser [10] call it the "Kreisel-Henkin criterion for self-reference". It can be seen as a more demanding version of the naïve understanding of what it means for a sentence to say of itself that it has the property expressed by $\varphi(x)$.

Despite its intuitive charm, (C1) could be seen as an over-generating condition on reference. Consider, for instance, the sentence $\mathsf{Bew}(\ulcorner\varphi\urcorner)$. It m-refers to φ, for the numeral $\ulcorner\varphi\urcorner$ $(= \overline{n} = S \ldots S0)$ occurs in it. But it also m-refers to every sentence whose code is smaller than φ's (whatever those are) because the numerals $\overline{m}$ (with $m < n$) of these codes are all subterms of $\ulcorner\varphi\urcorner$. This is a byproduct of not having an individual constant in the language to name each number in ω or, what amounts to the same, each expression of $\mathscr{L}$. Furthermore, if x occurs free in $\mathsf{Bew}(x)$ (recall $\mathsf{Bew}(x)$ is a complex formula) in the context of the open term $t(x)$ at least once, then $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ would also refer to the sentence denoted by $t(\ulcorner\varphi\urcorner)$, if any. Of course, which sentences these numerals and subterms refer to depends entirely on the coding. Under two different codings, sentence like $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ will 'over-generate' in different ways.

[13] As anticipated in footnote 4, this is a more general version of what is normally understood by "strong diagonalization". φ here may (and may not) contain free variables other than x. This possibility will become useful later in this section, in the proof of Proposition 3.3.

Unfortunately, this 'over-generation' is unavoidable. Discriminating between terms that play a role in reference and those that don't can lead to worse situations. For instance, to avoid that $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ m-refers to every sentence whose code is smaller than φ's, one could suggest we shouldn't look into numerals, but only consider terms in sentences that aren't proper subterms of numerals. But what if x in $\mathsf{Bew}(x)$ only occurs in the context of the function symbol S? In that case, we wouldn't be allowed to conclude that $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ refers to φ but only to the sentence codified by the successor of the code of φ, if such sentence exists. An analogous case can be made if other subterms were ignored for determining m-reference. If x occurs free in $\mathsf{Bew}(x)$ in the context of $t(x)$, to avoid that $\mathsf{Bew}(\ulcorner\varphi\urcorner)$ m-refers to the sentence denoted by $t(\ulcorner\varphi\urcorner)$, if any, one could require that only numerals are considered for m-reference. But in that case, we wouldn't be able to account for the self-referential character of sentences delivered by strong diagonalization, such as the 'strong' Gödel sentence of PA, $\neg\mathsf{Bew}(g)$, given by

$$g = \ulcorner\neg\mathsf{Bew}(g)\urcorner. \tag{5}$$

The term g the strong diagonal lemma provides is of the form $\dot{d}(\ulcorner\neg\mathsf{Bew}(\dot{d}(x))\urcorner)$. It's not a numeral.

Moreover, it's not clear we want to ignore terms like $t(\ulcorner\varphi\urcorner)$ or $\overline{m}$, with $m < \#\varphi$ in every sentence of the form $\psi(\ulcorner\varphi\urcorner)$ whatsoever. For instance, if we ignore the term $\dot{\neg}\ulcorner\varphi\urcorner$ in $\mathsf{Bew}(\dot{\neg}\ulcorner\varphi\urcorner)$ and only let $\ulcorner\varphi\urcorner$ to contribute to the reference of this sentence, we would not be able to say that $\mathsf{Bew}(\dot{\neg}\ulcorner\varphi\urcorner)$ refers to $\neg\varphi$, even though it intuitively says that $\neg\varphi$ is provable. $\psi(\ulcorner\varphi\urcorner)$ can be seen as saying of φ that it has the property expressed by ψ; or of the sentence denoted by $t(\ulcorner\varphi\urcorner)$ that it has the property expressed by $\psi(x)[x/t(x)]$, the result of replacing all occurrences of $t(x)$ in $\psi(x)$ with x; or of the sentence denoted by, e.g., $\ulcorner\varphi\urcorner - 1$, that it has the property expressed by $\psi(Sx)$. Each formula of $\mathscr{L}$ is as legitimate as any other. A similar phenomenon occurs in natural language. For instance, the sentence "The earth's circumference is smaller than Saturn's" not only refers to the earth's circumference but also to the earth; e.g., this sentence could be seen as part an answer to the question which planets have circumferences smaller than Saturn's.

Heck [12] regards m-self-reference to be 'true' self-reference, as opposed to the mere presence of equivalences of the form (A) and also to other *prima facie* possible ways of achieving self-reference. If he's right, m-reference is the only legitimate kind of reference, and (C1) and (C2) are not just sufficient but also necessary conditions for reference and self-reference, respectively. As a consequence, Heck's view is susceptible to Leitgeb's criticisms, since he cannot account for the self-referential character of sentences obtained via the weak diagonal lemma. In the equivalences of the form (A) the diagonal lemma delivers, ψ doesn't m-self-refer; it doesn't contain a term that denotes ψ but is only equivalent to a sentence (i.e., $\varphi(\ulcorner\psi\urcorner)$) that contains such term. But these are part of the paradigmatic cases we wish to account for. As Leitgeb points out, we want to make sense of the usual claim that Gödel sentences are self-referential. This issue is also pressing in the case of semantic paradoxes. If T is a truth predicate, the (weak) diagonalization lemma applied to the formula $\neg Tx$ delivers what is called a "liar" sentence.[14] Together with certain naïve truth principles a paradox can be derived from it. If the diagonal lemma didn't

[14] If we extend $\mathscr{L}$ with a predicate symbol T for truth and formulate Q in the extended language, we can diagonalize $\neg Tx$ to obtain a *liar* sentence λ such that $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$ is provable in this theory.

deliver self-referential expressions, then we would have a non-self-referential semantic paradox on the cheap.

Contra Heck, I suggest we don't limit 'true' reference to m-reference. Indeed, reference can also be achieved using quantifiers. For instance,

$$\forall x(\mathsf{T}_{\Sigma_1}(x) \rightarrow \mathsf{Bew}(x))$$

says that all Σ_1-truths are theorems of PA; it intuitively refers to all Σ_1-truths. Similarly,

$$\exists x(\mathsf{T}_{\Sigma_2}(x) \wedge \neg\mathsf{Bew}(x))$$

says some Σ_2-truths are not theorems of PA; so it intuitively refers to Σ_2-truths. $\forall x\mathsf{Bew}(x)$, instead, seems to refer to everything, for it states that everything is provable in PA. More generally,

(C3): sentences of the form

$$\forall x(\varphi(x) \rightarrow \psi(x)) \tag{C}$$

or

$$\exists x(\varphi(x) \wedge \psi(x)) \tag{D}$$

refer to all sentences satisfying φ, and

(C4): sentences of the form $\forall x\varphi(x)$, where φ is neither a conditional expression nor is equivalent to a conditional expression in a sense to be specified (although $\mathsf{Bew}(x)$ might be), refer to all sentences.

I call this kind of reference "reference by quantification" or "q-reference" for short. It is what Heck calls "reference by description". Unlike him, I do not consider it a second-class kind of reference.

As in the case of m-reference, subterms should not be ignored. Sentences such as

$$\forall x(x = \ulcorner\varphi\urcorner \rightarrow \mathsf{Bew}(\dot{\neg}x))$$

or

$$\exists x(x = \ulcorner\varphi\urcorner \wedge \mathsf{Bew}(\dot{\neg}x))$$

seem to refer not only to φ but also to $\neg\varphi$, for they say of the latter that it is provable. In a similar fashion, nested quantifiers also play a role in reference. For instance,

$$\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow \forall y(y = \dot{\neg}x \rightarrow \mathsf{Bew}(y)))$$

and

$$\exists x(x = \ulcorner 0 \neq 0\urcorner \wedge \exists y(y = \dot{\neg}x \wedge \mathsf{Bew}(y)))$$

appear to be asserting of $\neg 0 \neq 0$ that is provable and so referring to this sentence as well as to $0 \neq 0$. In general,

(C5): sentences of the form (C) or (D) refer, in addition to the φs, to whatever $\psi(\overline{n})$ refers to, provided that n satisfies $\varphi(x)$.

This condition covers the intuitions regarding q-reference behind the last two examples.

It appears to be sound to say that reference is closed under logical connectives, that is,

(C6): a sentence and its negation refer to the same expressions; the conjunction of two sentences refers to every sentence any conjunct refers to; etc.

This holds trivially of m-reference. In the case of q-reference, it implies, for instance, that

$$\neg\forall x(\varphi(x) \rightarrow \psi(x))$$

q-refers to the the same sentences as (C) does, and

$$\forall x(\varphi(x) \rightarrow \psi(x)) \wedge \exists x(\psi(x) \wedge \varphi(x))$$

q-refers to whatever $\forall x(\varphi(x) \rightarrow \psi(x))$ or $\exists x(\psi(x) \wedge \varphi(x))$ q-refer to.

(C3)–(C6) cover formulae of many different forms. Nonetheless, we haven't considered all cases. What sentences do expressions such as

$$\forall x\neg(\mathsf{T}_{\Sigma_1}(x) \rightarrow \mathsf{Bew}(x))$$

q-refer to? What about

$$\exists x\exists y(\mathsf{Bew}(x \dot{\rightarrow} y) \wedge \neg\mathsf{Bew}(y \dot{\rightarrow} x)),$$

where $x \dot{\rightarrow} y$ represents the function that maps formulae x and y to the conditional from x to y? There seems to be no straightforward way to precisely define q-reference. Rather arbitrary decisions will have to be made to give a complete definition. I will come back to this in the next section.

We can see now how sentences resulting from an application of (weak) diagonalization intuitively q-refer to themselves, so the desired condition can be met:

(C7): sentences resulting from an application of (weak) diagonalization are self-referential.

Recall the sentence ψ diagonalization delivers is actually of the form

$$\forall x(x = \ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner \rightarrow \forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))). \tag{B}$$

Although (B) does not satisfy its own antecedent, $\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))$ does. Furthermore, the sentence that results from replacing the free variable x in (B)'s consequent, $\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))$, with $\ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner$, i.e.,

$$\forall y(\mathsf{Diag}(\ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner, y) \rightarrow \varphi(y))),$$

q-refers to (B), for (B) satisfies the antecedent, $\mathsf{Diag}(\ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner, y)$. Therefore, (B), or in other words, ψ, q-refers to itself, by conditions (C3) and (C5). Analogously, these conditions guarantee that ψ is also self-referential when obtained by the existential proof of the diagonal lemma. For in that case ψ is of the form

$$\exists x(x = \ulcorner\exists y(\mathsf{Diag}^{\exists}(x, y) \wedge \varphi(y))\urcorner \wedge \exists y(\mathsf{Diag}^{\exists}(x, y) \wedge \varphi(y))),$$

so it refers to whatever

$$\exists y(\mathsf{Diag}^{\exists}(\ulcorner\exists y(\mathsf{Diag}^{\exists}(x, y) \wedge \varphi(y))\urcorner, y) \wedge \varphi(y))$$

refers to, i.e., to the diagonalization of $\exists y(\mathsf{Diag}^{\exists}(x, y) \wedge \varphi(y))$ or, what is the same, to itself.

In both cases, ψ's self-referential character is not a consequence of it being equivalent to $\varphi(\ulcorner\psi\urcorner)$, but of it somehow quantifying over itself. Dropping the naïve view on self-reference does not prevent us from classifying sentences delivered by (weak) diagonalization—e.g., Gödel sentences like γ in (1)—as self-referential, contrary to what Leitgeb suggests.

Leitgeb also notices that reference *simpliciter* cannot be defined as the disjunction of m- and q-reference. This is just a direct form of reference. In some occasions it also makes

sense to talk about *indirect* reference in the language of arithmetic. Let g_1, g_2 be terms such that we can prove in Q that

$$g_1 = \ulcorner \mathsf{Bew}(g_2) \urcorner \ \wedge \ g_2 = \ulcorner \neg \mathsf{Bew}(g_1) \urcorner.$$

$\mathsf{Bew}(g_2)$ says of $\neg\mathsf{Bew}(g_1)$ that it's provable in PA, while the latter says of the former that it is unprovable. These sentences form a reference cycle. If $\mathsf{Bew}(g_2)$ is true, then $\neg\mathsf{Bew}(g_1)$ is so too, which means that $\mathsf{Bew}(g_2)$ is not provable. Thus, it seems $\mathsf{Bew}(g_2)$ is indirectly saying something about itself. A similar point can me made for $\neg\mathsf{Bew}(g_1)$. Actually, Gödel's proof of the undecidability of PA's Gödel sentence, which relies on the self-referentiality of this sentence, can be easily adapted to show that both $\mathsf{Bew}(g_2)$ and $\neg\mathsf{Bew}(g_1)$ are undecidable as well.

On the other hand, there seem to be cases where reference is intuitively not closed under transitivity. Let $\mathsf{Sent}(x) \in \mathscr{L}$ define the set of sentences of $\mathscr{L}$. Then, $\mathsf{Sent}(\ulcorner \mathsf{Bew}(\ulcorner 0 = 0 \urcorner) \urcorner)$ m-refers to $\mathsf{Bew}(\ulcorner 0 = 0 \urcorner)$, which m-refers in turn to $0 = 0$. However, it's not clear we want to say that $\mathsf{Sent}(\ulcorner \mathsf{Bew}(\ulcorner 0 = 0 \urcorner) \urcorner)$ refers to $0 = 0$, not even indirectly. While $\mathsf{Sent}(\ulcorner \mathsf{Bew}(\ulcorner 0 = 0 \urcorner) \urcorner)$ says of $\mathsf{Bew}(\ulcorner 0 = 0 \urcorner)$ that it's a sentence, it says nothing in principle about $0 = 0$.

Cycles are perfectly possible in the language of arithmetic. Making small adjustments to the strong diagonal lemma, and provided the language contains the relevant function symbols, it is easy to prove the existence of cycles of any length in Q. Let $\dot{n}(x) \in \mathscr{L}$ represent in Q the p.r. function "numeral" that maps a number x to the code of its numeral, and let $\dot{s}(x, y) \in \mathscr{L}$ represent in Q the p.r. function s called "substitution" that maps the codes of a formula x and a term y to the code of the sentence that results from replacing the only free variable in x with y.

PROPOSITION 3.2 (n-cycles). *For any formulae $\varphi_1(x), \ldots, \varphi_n(x)$ there are terms $t_1, \ldots, t_n$ such that the following are provable in* Q:

$$\begin{aligned} t_1 &= \ulcorner \varphi_1(t_2) \urcorner \\ &\ldots \\ t_{n-1} &= \ulcorner \varphi_{n-1}(t_n) \urcorner \\ t_n &= \ulcorner \varphi_n(t_1) \urcorner. \end{aligned}$$

Proof. I prove it just for $n = 2$. We can show in Q that

$$\underbrace{\dot{d}(\ulcorner \varphi_2(\dot{s}(\ulcorner \varphi_1(\dot{d}(x)) \urcorner, \dot{n}(x))) \urcorner)}_{t_2} = \ulcorner \varphi_2(\underbrace{\dot{s}(\ulcorner \varphi_1(\dot{d}(x)) \urcorner, \dot{n}(\ulcorner \varphi_2(\dot{s}(\ulcorner \varphi_1(\dot{d}(x)) \urcorner, \dot{n}(x))) \urcorner))}_{t_1}) \urcorner.$$

Thus,

$$\mathsf{Q} \vdash t_2 = \ulcorner \varphi_2(t_1) \urcorner$$

and

$$\mathsf{Q} \vdash t_1 = \ulcorner \varphi_1(\dot{d}(\ulcorner \varphi_2(\dot{s}(\ulcorner \varphi_1(\dot{d}(x)) \urcorner, \dot{n}(x))) \urcorner)) \urcorner = \ulcorner \varphi_1(t_2) \urcorner.$$

This proof can be extended to cycles of any length in a recursive way as follows: If for an n-cycle for $\varphi_1(x), \ldots, \varphi_n(x)$ we start by strongly diagonalizing $\varphi(x)$ (e.g., for cycles of length 2 we strongly diagonalize $\varphi_2(\dot{s}(\ulcorner \varphi_1(\dot{d}(x)) \urcorner, \dot{n}(x)))$), and for an $n + 1$-cycle for $\varphi_1(x), \ldots, \varphi_{n+1}(x)$ we strongly diagonalize $\varphi_{n+1}(\dot{s}(\ulcorner \varphi(x) \urcorner, \dot{n}(x)))$. In other words, for

$\varphi_1(x), \ldots, \varphi_n(x)$ we begin by 'unwinding'

$$\underbrace{\dot{d}(\ulcorner\varphi_n(\dot{s}(\ulcorner\varphi_{n-1}(\ldots\dot{s}(\ulcorner\varphi_1(\dot{d}(x))\urcorner, \dot{n}(x))\ldots)\urcorner, \dot{n}(x)))\urcorner)}_{t_n}.$$

□

Moreover, we can prove the existence of ω-sequences of sentences, each of which m-refers to the expression coming next.

PROPOSITION 3.3 (ω-chains). *For every formula $\varphi(x)$ there is an infinite sequence of distinct terms $t_0, t_1, \ldots, t_n, \ldots$ such that, for every $n \in \omega$,*

$$\mathsf{Q} \vdash t_n = \ulcorner\varphi(t_{n+1})\urcorner.$$

Proof. Applying the strong diagonalization lemma to $\varphi(\dot{s}(x, \dot{n}(Sy)))$ we obtain a term t such that

$$\mathsf{Q} \vdash t = \ulcorner\varphi(\dot{s}(t, \dot{n}(Sy)))\urcorner.$$

Thus, applying $\dot{s}(x, \dot{n}(y))$ to both sides of the equation, we can prove in Q that

$$\forall y\ \dot{s}(t, \dot{n}(y)) = \dot{s}(\ulcorner\varphi(\dot{s}(t, \dot{n}(Sy)))\urcorner, \dot{n}(y)). \tag{6}$$

For each $n \in \omega$, let $t_n := \dot{s}(t, \dot{n}(\overline{n}))$. It follows from (6) that

$$\begin{aligned} t_n &= \dot{s}(\ulcorner\varphi(\dot{s}(t, \dot{n}(Sy)))\urcorner, \dot{n}(\overline{n})) \\ &= \ulcorner\varphi(\dot{s}(t, \dot{n}(S\overline{n})))\urcorner \\ &= \ulcorner\varphi(\dot{s}(t, \dot{n}(\overline{n+1})))\urcorner \\ &= \ulcorner\varphi(t_{n+1})\urcorner. \end{aligned}$$

Moreover, we can show that not only each term in the sequence is distinct from the others but also that they denote different sentences, as follows:

$$\begin{aligned} t_n = t_m &\Rightarrow \dot{s}(t, \dot{n}(\overline{n})) = \dot{s}(t, \dot{n}(\overline{m})) \\ &\Rightarrow \dot{s}(\ulcorner\varphi(\dot{s}(t, \dot{n}(Sy)))\urcorner, \dot{n}(\overline{n})) = \dot{s}(\ulcorner\varphi(\dot{s}(t, \dot{n}(Sy)))\urcorner, \dot{n}(\overline{m})) \\ &\Rightarrow \ulcorner\varphi(\dot{s}(t, \dot{n}(S\overline{n})))\urcorner = \ulcorner\varphi(\dot{s}(t, \dot{n}(S\overline{m})))\urcorner \\ &\Rightarrow \overline{n} = \overline{m}, \end{aligned}$$

(for the coding is injective) which means that $n = m$. □

Thus, for instance, we can have an ω-chain of Henkin sentences, that is,

$$\begin{aligned} h_1 &= \ulcorner\mathsf{Bew}(h_2)\urcorner \\ h_2 &= \ulcorner\mathsf{Bew}(h_3)\urcorner \\ &\ldots \\ h_n &= \ulcorner\mathsf{Bew}(h_{n+1})\urcorner \\ &\ldots \end{aligned}$$

As in the case of standard Henkin sentences, if a member of this chain is provable, and thus true, then all the ones coming after it in the chain are true and provable as well; while if a sentence is unprovable and thus false, so are the following ones. This invites the thought that each sentence in the sequence refers, albeit indirectly, to all the ones that come after and not only to the one that follows immediately.

Furthermore, if a truth predicate were available in the language, turning to Proposition 3.2 we could formulate cycles of liars, from which we could derive a paradox pretty much in the same way we do with the standard liar sentence. If sentences in the cycle weren't

indirectly self-referential, we would get non-self-referential semantic paradoxes on the cheap, as before. I take all this to show that, although sometimes it might not be necessary or adequate to go beyond direct reference,

(C8): reference *simpliciter* is a transitive relation.

The last condition Leitgeb imposes on reference is the equivalence condition:

(EC): *logically* (not merely arithmetically) equivalent sentences refer to the same things.[15]

He gives the following supporting argument:

> (EC) is plausible because logically equivalent sentences are not only extensionally equivalent in the actual world, but indeed in every logically possible world, and thus indistinguishable in terms of the semantics of first-order predicate logic. If self-reference is to be defined by extending the usual reference relation for terms, i.e., a semantical relation, it is certainly strange if (EC) is invalidated. If (EC) is not true, the self-referentiality or circularity of a sentence does not only depend on what the sentence says, but also in which way its content is being expressed. (Leitgeb [19, p. 10])

This argument consists of two premises: (i) that semantic notions of first-order logic cannot distinguish between logically equivalent sentences and (ii) that (self-)reference is a semantic relation. The question we need to ask is what is meant here by "semantic". If "semantic" means a predicate of sentences that is definable purely in terms of models or possible worlds, such as "being a logical truth", then premise (i) is true, but premise (ii) is false. According to the previous conditions Leitgeb himself suggested for reference, a definition of this concept will inevitably mention the syntactic components of the referring sentences. Indeed, the self-referentiality of a sentence does depend on the way its content is being expressed.[16] On the other hand, if "semantic" is to designate also definitions that mention concepts other than models or possible worlds, then premise (i) fails to be true because we can certainly distinguish between logically equivalent sentences such as $0 = 0$ and $\forall x(\varphi(x) \rightarrow \varphi(x))$ attending to their syntactic structure. Leitgeb's argument fails to support (EC) as a condition a notion of reference should satisfy.

Moreover, as Leitgeb himself notices, (EC) is incompatible with some of the other conditions for reference that have been discussed in this section. On pain of trivializing reference, if the equivalence condition held we would have to drop (C1). Given any two sentences φ and ψ, there is always a sentence that is logically equivalent to φ and mentions ψ, e.g., $\varphi \wedge \ulcorner\psi\urcorner = \ulcorner\psi\urcorner$. Thus, (C1) would imply that every sentence refers to every other

[15] Actually, this is not Leitgeb's original formulation. His reads: "if A is self-referential/circular, and if B is logically equivalent to A, then also B is self-referential/circular." (Leitgeb [19, p. 10]) However, as Urbaniak [32] points out, this condition is far stronger and less convincing. Let $t = \mathsf{Bew}(t)$ and consider its logical equivalent $\forall x(x = t \rightarrow \mathsf{Bew}(x))$. While it might be intuitively appealing to assert that $\mathsf{Bew}(t)$ and $\forall x(x = t \rightarrow \mathsf{Bew}(x))$ refer to the same things, is far less clear that we want to commit ourselves to the idea that $\forall x(x = t \rightarrow \mathsf{Bew}(x))$ is self-referential just because $\mathsf{Bew}(t)$ is. Intuitively, $\forall x(x = t \rightarrow \mathsf{Bew}(x))$ refers to $\mathsf{Bew}(t)$, but not to itself. Given that the weaker version of the equivalence condition I suggest is already problematic, I stick to it. Moreover, Leitgeb's arguments support my version of the equivalence condition rather than his own.

[16] In [20] Leitgeb changes his mind: he acknowledges this fact and rejects (EC).

sentence. For similar reasons, (C3) should be dropped in the presence of (EC): (C) is logically equivalent to

$$\forall x((\varphi(x) \to \varphi(x)) \to (\varphi(x) \to \psi(x))),$$

whose antecedent is true of every sentence in the language.

If one still finds (EC) to be an appealing condition, a way of resolving the conflict with the other conditions could *prima facie* consist in imposing restrictions on which terms and predicates can be sources of reference, as has been done by Putnam [25], Goodman [6], and Urbaniak [32]. According to these accounts of reference or *aboutness* sentences refer more or less as expected, as long as the relevant terms or predicates occur *informatively*. Roughly, a sentence of the form $\varphi(t)$ is about the object t denotes only if $\varphi(t)$ doesn't logically imply $\varphi(s)$ for every other term s. In the same fashion, sentences of the form $\forall x(\varphi(x) \to \psi(x))$ are about the φs (or the class of φs) just in case they don't logically imply $\forall x(\chi(x) \to \psi(x))$ for every other formula $\chi(x)$ (or φ is a logical predicate). Thus, if the equivalence condition was at play, for instance, φ and $\varphi \wedge \ulcorner\psi\urcorner = \ulcorner\psi\urcorner$ would refer to the same sentences, but the latter would not refer to ψ, for $\ulcorner\psi\urcorner$ does not occur informatively in $\varphi \wedge \ulcorner\psi\urcorner = \ulcorner\psi\urcorner$. In this way, triviality can be avoided.

However, other counterintuitive cases and incompatibilities emerge. Most saliently, the very idea of informativity prevents the identification of strong diagonalization as a mechanism for self-reference. If we strongly diagonalize the (logical) predicate $x = x$, we obtain a term t such that $t = \ulcorner t = t\urcorner$. Since $t = t$ logically implies $s = s$ for every term s, $t = t$ isn't about itself and, therefore, isn't self-referential. An analogous claim can be made about weak diagonalization.

I conclude that the equivalence condition is not a reasonable requirement we should impose on reference. Thus, it appears we can still hope to find an adequate notion of reference in $\mathscr{L}$, contrary to what Leitgeb [20] seems to suggest. As Smoryński [28] and Halbach and Visser [10, 11] indicate, the fruitless attempts to make these conditions work together, the resistance of reference to be treated as an extensional concept, could be what lead mathematicians and philosophers away from the formal study of reference and self-reference in arithmetic.

Indeed, rejecting (EC) implies reference should not only be intensional but also *hyperintensional* in the following sense:[17]

(C9): some logically equivalent sentences fail to refer to the same objects.

In particular, pairs of logically equivalent sentences of the form φ and $\varphi \wedge t = t$, or $\forall x(\varphi(x) \to \psi(x))$ and $\forall x((\chi(x) \to \chi(x)) \to (\varphi(x) \to \psi(x)))$, do not necessarily refer to the same things. There are other cases of logically equivalent schemata, however, for which we feel inclined to believe they do refer to the same objects. Take for instance $\forall x(\neg\varphi(x) \to \psi(x))$ and $\forall x(\neg\psi(x) \to \varphi(x))$. Unlike the previous examples, these just seem to be two different ways of expressing exactly the same (trivial or nontrivial) content about the same objects. A similar point can be made concerning $\forall x(\varphi(x) \to \psi(x))$ and $\forall x(\neg\psi(x) \to \neg\varphi(x))$, $\forall x(\varphi(x) \to \psi(x))$ and $\forall x\neg\neg(\varphi(x) \to \psi(x))$, $\exists x(\varphi(x) \wedge \psi(x))$ and $\exists x(\psi(x) \wedge \varphi(x))$, $\forall x\forall y\varphi$ and $\forall y\forall x\varphi$, $\forall x\varphi(x)$ and $\forall y\varphi(y)$, and other transformations of the like.

Thus, in an ideal situation, instead of rejecting the equivalence condition altogether, we keep the good bits without trivializing our definition of reference. How can we identify the

[17] See Cresswell [4]. Since hyperintensionality is a kind of intensionality, sometimes hyperintensional predicates are just referred to as "intensional".

good bits? (EC) trivializes m- and q-reference because, for instance, we can always add $t = t$ as a conjunct to every formula and obtain a logically equivalent expression, or relativize every conditional to a logical truth like $\varphi(x) \rightarrow \varphi(x)$. Unlike the examples I gave at the end of last paragraph, these logical equivalents add *irrelevant* compounds to sentences. Thus, an idea would be to adopt a version of (EC) restricted to logical transformations that do not add new atomic formulae but, roughly, just distribute them in a different way. According to this restricted version of (EC),

(C10): reference is closed, not under classical logic, but under some kind of relevant consequence relation.

In this sense we could say the notion of reference we are after is not absolutely hyperintensional but lies somewhere between intensionality and hyperintensionality.

I have examined the conditions that should hold of a good notion of reference that could help us to properly formulate certain problems in metamathematics, like the ones introduced in §2. These are conditions (C1)–(C10). The resulting notion could serve as a blueprint for a concept of reference applicable to the study of semantic paradoxes and, perhaps, even natural language. I've argued that reference should hold between sentences (via their codes, in a particular fixed coding) and do justice to the intuitions behind reference by mention and by quantification and the possible transitivity of reference. Moreover, it should be hyperintensional but, at the same time, closed under a weaker consequence relation than (classical) logical consequence. In the next section I show such concept is possible by setting an example and proving the pertaining results.

§4. In order to give a precise definition of reference I first need to introduce three main preliminary notions: m-reference, q-reference, and direct reference. Then, a definition of reference *simpliciter* is given, and several results on this notion that establish its adequacy are stated and proved. Finally, I define self-reference and well-foundedness in terms of reference. I show that the diagonalization procedures used in the proofs of Theorems 1.1 and 3.1 and Proposition 3.2 deliver self-referential sentences, whereas sentences in the ω-chains that Proposition 3.3 provides turn out to be non-well-founded (but unfortunately also self-referential).

It's important to highlight that the concepts of reference I put forward are of a semantic nature, for they depend on the standard interpretation $\mathbb{N}$ of the language.[18] For instance, when defining m-reference, I assume the denotation of terms that occur in sentences is given by $\mathbb{N}$;[19] and in defining q-reference, when I say the code of a sentence satisfies the antecedent of $\forall x(\varphi(x) \rightarrow \psi(x))$, I mean satisfaction in $\mathbb{N}$. I believe this is the most natural way of understanding conditions (C1) and (C3) in the previous section. Moreover, it doesn't tie reference to a particular theory such as PA, which would lead to undesired results.

DEFINITION 4.1 (M-reference). *If φ, ψ are sentences, then φ m-refers to ψ if and only if φ contains a closed term t such that $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.*

I require that t is closed in φ to keep apart m- from q-reference, as we will see soon. Definition 4.1 clearly satisfies condition (C1). As a consequence, all sentences denoted by

[18] Notions of reference of a proof-theoretic nature have been explored by the author in [author] for the language of truth, and are the subject of further work.

[19] Although in this case it would suffice to consider the denotation relations that are provable in Q, for Q proves all true identities and inequalities.

terms delivered by the strong diagonal lemma m-refer to themselves. If t is a closed term and $t = \ulcorner\varphi(t)\urcorner$ is provable in Q, then it's true in $\mathbb{N}$, so $\varphi(t)$ contains a closed term t that denotes $\varphi(t)$. Sentences like $t = t$, where $t = \ulcorner t = t\urcorner$, m-refer to themselves just like $\neg\mathsf{Bew}(g)$ in (5) does.

As expected, m-reference is closed under logical connectives. Also, it is naturally closed under the desired kind of relevant equivalence. Valid propositional transformations that do not add any new atoms do not alter m-reference. For instance, φ and $\neg\neg\varphi$, $\varphi \to \psi$ and $\neg\psi \to \neg\varphi$, $\varphi \vee \psi$ and $\neg\varphi \to \psi$, and $\neg(\varphi \wedge \psi)$ and $\neg\varphi \vee \neg\psi$ m-refer to the same sentences, correspondingly. Moreover, we can swap quantifiers of the same kind and rename variables without affecting the sentences an expression m-refers to, since none of this changes the closed terms that occur in a formula.

On the other hand, defining reference by quantification in a way that also enjoys closure under this kind of relevant consequence relation turns out to be a much more complicated task. To begin with, I introduce the notion of a formula being in *postnex disjunctive normal form* (PDNF, cf. Definition 4.3) and describe a procedure that allows us to transform any given formula into an expression in this form, which I call "normalization" (cf. Definition 4.6). Roughly, a formula is in PDNF just in case all its subformulae are disjunctions of conjunctions of atomic, universal, negated atomic, or negated universal expressions, and the normalization of a formula is the result of applying successive transformations to it that preserve logical equivalence and do not add any new atoms, until the resulting formula is in PDNF. This is close to the notion of *prenex disjunctive normal form* and the algorithms to obtain such formulae that can be usually found in textbooks.[20] Then, a direct definition of the q-reference of sentences in PDNF will be given. The q-reference of other sentences will be defined as the q-reference of their corresponding normalizations (cf. Definition 4.9). This will guarantee that all sentences that have the same normalization, which are obviously logically equivalent, refer by quantification to the same things. Moreover, those sentences whose respective normalizations differ only in the order or association of their conjuncts and disjuncts, or in the renaming of the variables, will also q-refer to the same expressions. Thus, q-reference will be closed, not under classical logic but under the kind of relevant transformations mentioned at the end of last section.

To express every formula of the language as a disjunction of conjunctions of atomic, universal, negated atomic, or negated universal expressions, we first need to get rid of the logical connectives that cannot occur in such formulae, that is, $\to$ and $\exists$. In order to do so, we translate each formula of $\mathscr{L}$ into $\mathscr{L}\restriction \subseteq \mathscr{L}$, the language that results from removing from $\mathscr{L}$ all formulae containing $\to$ or $\exists$. We turn to the usual definitions of $\to$ and $\exists$ in terms of $\neg$ and $\vee$ resp. $\neg$ and $\forall$. Let $\tau : \mathscr{L} \mapsto \mathscr{L}\restriction$ be defined as follows:

$$\tau(\varphi) := \begin{cases} \varphi & \text{if } \varphi \text{ is of the form } s = t, \\ \neg\tau(\psi) & \text{if } \varphi \text{ is of the form } \neg\psi, \\ \tau(\psi) \wedge \tau(\chi) & \text{if } \varphi \text{ is of the form } \psi \wedge \chi, \\ \tau(\psi) \vee \tau(\chi) & \text{if } \varphi \text{ is of the form } \psi \vee \chi, \\ \neg\tau(\psi) \vee \tau(\chi) & \text{if } \varphi \text{ is of the form } \psi \to \chi, \\ \forall v\,\tau(\psi) & \text{if } \varphi \text{ is of the form } \forall v\,\psi, \\ \neg\forall v\neg\tau(\psi) & \text{if } \varphi \text{ is of the form } \exists v\,\psi. \end{cases}$$

τ obviously preserves truth-in-a-model and provability-in-a-theory.

[20] See, for instance, Boolos *et al.*, [1, sec. 19.1].

DEFINITION 4.2 (Prime). *A formula φ of $\mathscr{L}\restriction$ is a prime if and only if it is an atomic formula, the negation of an atomic formula, a universally quantified expression, or the negation of a universally quantified expression.*

DEFINITION 4.3 (Postnex disjunctive normal form). *A formula of $\mathscr{L}\restriction$ is in PDNF if and only if*

1. *every subformula is a disjunction of conjunctions of primes*;
2. *it contains no superfluous quantifiers (i.e., that don't bind any variable)*;
3. *every subformula of the form $\forall v_1 \dots v_n(\varphi_1 \vee \dots \vee \varphi_m)$ is such that v_i is free in φ_j for each $1 \leq i \leq n$ and $1 \leq j \leq m$.*

For instance,

$$\neg\forall x(\forall y x = y \vee x \neq Sx)$$

is in PDNF, while

$$\neg\forall x\forall y(x = y \vee x \neq Sx)$$

isn't because y is not free in $x \neq Sx$. In turn,

$$\neg\forall x\neg(\neg\forall y x = y \wedge z = Sx)$$

isn't in PDNF either, since the subformula $\neg(\neg\forall y x = y \wedge x = Sx)$ is not a disjunction of conjunctions of primes.

Next I introduce the notion of normalization, an algorithm for turning each formula φ of $\mathscr{L}\restriction$ into PDNF form. It consists on the step-by-step transformation of each subformula of φ according to the number of nested quantifiers that occur in the subformula. Thus, I first provide the following two definitions.

DEFINITION 4.4 (Depth). *Let dep be an assignment of numbers to universally quantified formulae of $\mathscr{L}\restriction$ such that*

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

Intuitively, the depth of $\forall v\varphi$ is the maximum length of chains of nested quantifiers occurring in the formula. Thus, each universal formula of $\mathscr{L}\restriction$ has finite depth.

DEFINITION 4.5 (i-normalization). *The i-normalization $[\varphi]^i$ of a formula φ of $\mathscr{L}\restriction$ without superfluous quantifiers is the result of successively applying the following transformations to each subformula $\forall v\psi$ of φ of depth i:*

1. *Replace every subformula of the form $\neg(\gamma \vee \delta)$ and $\neg(\gamma \wedge \delta)$ with $(\neg\gamma \wedge \neg\delta)$ and $(\neg\gamma \vee \neg\delta)$ resp. until they don't occur any longer, starting with the innermost.*
2. *Erase all double negations.*
3. *Replace every subformula of the form $\chi \wedge (\gamma \vee \delta)$ and $(\gamma \vee \delta) \wedge \chi$ with $(\chi \wedge \gamma) \vee (\chi \wedge \delta)$ and $(\gamma \wedge \chi) \vee (\delta \wedge \chi)$ resp. until they don't occur any longer, starting with the innermost.*
4. *Replace every subformula of the form $\forall v_1 \dots v_n(\chi_1 \vee \dots \vee \chi_m)$ where v_1 isn't free in some χ_j, $1 \leq j \leq m$, with $\forall v_2 \dots v_n(\gamma \vee \forall v_1\delta)$, where γ is the disjunction of the χ_j in which v_1 is not free, and δ is the disjunction of the χ_j in which it is free; until such subformulae don't occur any longer.*

The i-normalization of a formula turns all its subformulae of depth i into disjunctive normal form, except literals (i.e., atomic or negated atomic expressions) are replaced with primes. Unlike prenex normal forms, quantifiers are pushed inside rather than outside of disjunctions, just next to the disjuncts whose variables they can bind. Consider the formula

$$\forall x(\forall y\neg(\forall z(x = 0 \vee z = 0) \wedge x = y) \vee \forall z\neg(z \neq z)). \tag{7}$$

Its 1-normalization consists in applying transformations 1-4 in the above definition to its subformulae of the form $\forall v\varphi$ of depth 1, i.e., $\forall z(z = 0 \vee x = 0)$ and $\forall z\neg(z \neq z)$. This results in $x = 0 \vee \forall z(z = 0)$ and $\forall z(z = z)$, resp. Thus, $[(7)]^1$ is

$$\forall x(\forall y\neg((x = 0 \vee \forall z(z = 0)) \wedge x = y)) \vee \forall z(z = z). \tag{8}$$

On the other hand, $[(7)]^2$ consists in transforming the universal subformulae of (7) of depth 2, so we just replace $\forall y\neg(\forall z(x = 0 \vee z = 0) \wedge x = y)$ with $\neg\forall z(x = 0 \vee z = 0) \vee \forall y(x \neq y)$. This results in

$$\forall x(\neg\forall z(x = 0 \vee z = 0) \vee \forall y(x \neq y) \vee \forall z\neg(z \neq z)).$$

DEFINITION 4.6 (Normalization). *The normalization* $[\varphi]$ *of a formula* $\varphi \in \mathscr{L}\restriction$ *is the result of erasing all superfluous quantifiers and then performing successive* i*-normalizations starting with* $i = 1$ *and stoping after* $i = max\{dep(\forall v\psi) : \forall v\psi$ *is a subformula of* $\varphi\}$.

$max\{dep(\forall v\psi) : \forall v\psi$ is a subformula of $\varphi\}$ is the maximum of the depths of the universal subformulae of φ. If φ doesn't contain quantifiers, let $max\{dep(\forall v\psi) : \forall v\psi$ is a subformula of $\varphi\} = 0$.

Going back to our previous example, $[(7)]^1$ is (8), so $[[(7)]^1]^2$ (which is not the same as $[(7)]^2$) is the result of replacing $\forall y\neg((x = 0 \vee \forall z(z = 0)) \wedge x = y)$ in (8) with $(x \neq 0 \wedge \neg\forall z(z = 0)) \vee \forall y(x \neq y)$, that is,

$$\forall x((x \neq 0 \wedge \neg\forall z(z = 0)) \vee \forall y(x \neq y) \vee \forall z(z = z)). \tag{9}$$

Finally, $[[[(7)]^1]^2]^3$, that is, $[(7)]$, is the result of pushing $\forall x$ inside in (9), as clause 4 of Definition 4.5 requires, for x is not free in $\forall z(z = z)$:

$$\forall z(z = z) \vee \forall x((x \neq 0 \wedge \neg\forall z(z = 0)) \vee \forall y(x \neq y)).$$

It can be shown that every formula of $\mathscr{L}\restriction$ is logically equivalent to a PDNF formula. Moreover, normalization is an effective procedure to find this expression. For reasons of perspicuity, in what follows I will often talk of normalizations of formulae of $\mathscr{L}$ when what is really meant are normalizations of the translations of these formulae into $\mathscr{L}\restriction$.

PROPOSITION 4.7. *Every formula* $\varphi \in \mathscr{L}\restriction$ *is logically equivalent to a formula in which all subformulae of the form* $\forall v\psi$ *are in PDNF, and normalization is an effective procedure to find one.*

Proof. Note that clauses 1–4 of Definition 4.5 imply only a finite number of transformations. Thus, i-normalizations terminate in finitely many steps. Note as well that erasing superfluous quantifiers and performing the transformation steps in Definition 4.5 to a formula result in a logically equivalent expression.

Let φ^- be the result of erasing all superfluous quantifiers in φ. We now show by induction on n that, for all $n \geq 1$, in the formula that results from successively performing i-normalizations, from $i = 1$ to $i = n$ to φ^- (i.e., $[\ldots[[\varphi^-]^1]\ldots]^n$), all universal subformulae of depth $\leq n$ are in PDNF. Since erasing superfluous quantifiers is a finite operation as well, we get the desired proof.

Let $\forall v\,\psi$ be any subformula of $[\varphi^{-}]^{1}$ of depth 1. By clause 4 of Definition 4.5, if ψ is of the form $\psi_1 \vee \cdots \vee \psi_m$, v is free in ψ_i, for each $1 \leq i \leq m$. Note that pushing the only quantifier inside a disjunction does not generate new formulae of the form $\chi \wedge (\gamma \vee \delta)$ and $(\gamma \vee \delta) \wedge \chi$ inside a subformula *of depth 1*. Then, by clause 3, no conjunctions of disjunctions can occur in $\forall v\,\psi$. By clause 2 and the fact that pushing the quantifier inside a disjunction and distributing conjunctions over disjunctions doesn't create new double negations, there are also no double negations in $\forall v\,\psi$. Finally, pushing the quantifier inside a disjunction, distributing conjunctions over disjunctions, and erasing double negations doesn't create new formulae of the form $\neg(\gamma \vee \delta)$ or $\neg(\gamma \wedge \delta)$ inside a subformula *of depth 1* if there originally weren't any. Thus, by clause 1, the latter don't occur either in $\forall v\,\psi$. As a consequence, $\forall v\,\psi$ is in PDNF.

Assume now that all universal subformulae of $[\ldots[[\varphi^{-}]^{1}]^{\cdots}]^{n}$ of depth $\leq n$ are in PDNF and let $\forall v\,\psi$ by any subformula of $[\ldots[[\varphi^{-}]^{1}]^{\cdots}]^{n+1}$. By clause 4 of Definition 4.5 and the inductive hypothesis, if ψ is of the form $\forall v_1 \ldots v_k(\psi_1 \vee \cdots \vee \psi_m)$, then v_i is free in ψ_p, for each $1 \leq i \leq k$ and $1 \leq p \leq m$. Also, pushing the first quantifier of a sequence inside a disjunction in a subformula of arbitrary depth m does not generate new subformulae of the form $\chi \wedge (\gamma \vee \delta)$ and $(\gamma \vee \delta) \wedge \chi$ inside a subformula *of depth m*. Thus, by clause 3, no conjunctions of disjunctions can occur in $\forall v\,\psi$. By clause 2 and the fact that pushing the first quantifier of a sequence inside a disjunction and distributing conjunctions over disjunctions doesn't create new double negations, there are also no double negations in $\forall v\,\psi$. Finally, pushing the first quantifier of a sequence inside a disjunction, distributing conjunctions over disjunctions, and erasing double negations in a subformula of arbitrary depth m doesn't create new formulae of the form $\neg(\gamma \vee \delta)$ or $\neg(\gamma \wedge \delta)$ inside a subformula *of depth m*, if there originally weren't any. Therefore, by clause 1, the latter don't occur either in $\forall v\,\psi$. As a consequence, $\forall v\,\psi$ is in PDNF. □

Actually, the normalization of a formula does not return an expression in PDNF, but just one in which all quantified subformulae are in PDNF. Although it won't be necessary in what follows, one can easily obtain a formula in PDNF by applying clauses 1–3 of Definition 4.5 to its normalization.

DEFINITION 4.8 (Permutations). *The set of permutations of a formula $\varphi \in \mathscr{L}^{\restriction}$ is the smallest set containing φ that is closed under commutativity and associativity of disjunction.*

Although Leitgeb's equivalence condition should be rejected on pain of triviality, a restricted version of it is desirable, as stated in (C9) and (C10). My proposal here consists in just requiring that sentences whose normalizations have the same set of permutations—which are obviously logically equivalent—refer to the same objects. In other words, reference will be closed under permutations of disjunctive subformulae of normalizations. Note that translations, normalizations, and permutations do not disturb the atomic components of sentences, which seemed to be the problem with Leitgeb's equivalence condition, but just change connectives and redistribute quantifiers. Later in this section I show that the resulting definition of reference does not lead to triviality and, furthermore, it gives the right verdict in several intuitive cases. Let $\vec{k}$ abbreviate $k_1, \ldots, k_n$.

DEFINITION 4.9 (Q-reference). *If φ, ψ are sentences, φ q-refers to ψ if and only if a member of the set of permutations of $[\tau(\varphi)]$ has a subsentence of the form $\forall\vec{v}\,\chi$ satisfying one of the following two conditions*:

1. χ *is atomic, a negated formula, or a conjunction.*
2. χ *is of the form* $\delta \vee \gamma$, *there are* $\vec{k} \in \omega$ *such that* $\mathbb{N} \vDash \neg\delta[\overrightarrow{\overline{k}}/\vec{v}]$, *and*
   (a) $\ulcorner\psi\urcorner = \overline{k_i}$ *for some* $1 \leq i \leq n$, *or*
   (b) $\gamma\,[\overrightarrow{\overline{k}}/\vec{v}]$ *is the normalization of a sentence that q-refers to* ψ *or contains an occurrence of a closed term* t *that isn't in* γ *such that* $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.[21]

Since only quantified subsentences contribute to q-reference, sentences not containing quantifiers, such as $\ulcorner\varphi\urcorner \dot{\rightarrow} \ulcorner\psi\urcorner = \ulcorner\varphi \rightarrow \psi\urcorner$, do not q-refer, as expected. Also, the fact that only normalizations are considered implies that superfluous quantifiers aren't a source of q-reference either, for formulae in PDNF cannot contain them. Thus, $\forall x \ulcorner\varphi\urcorner \dot{\rightarrow} \ulcorner\psi\urcorner = \ulcorner\varphi \rightarrow \psi\urcorner$, for instance, does not q-refer to any sentence.

Moreover, Definition 4.9 satisfies the conditions (C3)–(C5) stated in the previous section. The idea behind clauses 1 and 2(a) is that the only way of restricting the referential power a quantifier carries with it is via conditional expressions, that is, bounded quantification allows for restricted q-reference. Recall conditionals are translated into $\mathscr{L}^{\restriction}$ as disjunctions. If a normalized sentence φ has a subsentence of the form $\forall\vec{v}\chi$ in which χ is not a disjunction, given the normalization process, it means χ cannot be a conditional, and reference by quantification is unrestricted. Thus, $\forall\vec{v}\chi$ refers to everything, and so does φ. As a consequence, condition (C4) is satisfied. For example,

$$\forall x(x \neq \dot{\neg} x),$$

$$\forall x \forall y \neg(x = y \rightarrow x \dot{\rightarrow} y \neq y \dot{\rightarrow} x), \text{ and}$$

$$\forall x \neg \forall y \neg(y = \dot{\neg} x \rightarrow y \neq x \dot{\rightarrow} x)$$

q-refer to every sentence because $x \neq \dot{\neg} x$, $\neg(x = y \rightarrow x \dot{\rightarrow} y \neq y \dot{\rightarrow} x)$, and $\neg\forall y \neg(y = \dot{\neg} x \rightarrow y \neq x \dot{\rightarrow} x)$ cannot be rewritten as conditionals since the normalizations of $\forall x(x \neq \dot{\neg} x)$, $\forall x \forall y \neg(x = y \rightarrow x \dot{\rightarrow} y \neq y \dot{\rightarrow} x)$, and $\forall x \neg \forall y \neg(y = \dot{\neg} x \rightarrow y \neq x \dot{\rightarrow} x)$ are $\forall x(x \neq \dot{\neg} x)$ resp. $\forall x \forall y(x = y \wedge x \dot{\rightarrow} y = y \dot{\rightarrow} x)$ and $\forall x \neg \forall y(y = \dot{\neg} x \vee y = x \dot{\rightarrow} x)$. In the latter, $\forall y$ cannot be pushed inside the disjunction, for y is free in both disjuncts. Also, the q-reference of sentences for the form

$$\forall x \neg(\varphi(x) \rightarrow \psi(x)) \tag{10}$$

is now decided as follows: if $\varphi(x)$ and $\psi(x)$ are atomic, then (10) q-refers to all sentences. Otherwise, it depends on what the normalization of (10) is.

If, on the other hand, φ contains a subsentence of the form $\forall\vec{v}\chi$ in which χ is of the form $\delta \vee \gamma$, we can read the latter as the conditional $\neg\delta \rightarrow \gamma$, which restricts the quantifiers $\forall\vec{v}$ to the codes of sentences $\vec{k} \in \omega$ satisfying $\neg\delta$ in $\mathbb{N}$. Clause 2(a) guarantees that sentences of the form

$$\forall x(\varphi(x) \rightarrow \psi(x)) \tag{C}$$

and

$$\exists x(\varphi(x) \wedge \psi(x)) \tag{D}$$

[21] A way of making precise the idea of an occurrence of a term t in $\gamma\,[\overrightarrow{\overline{k}}/\vec{v}]$ that wasn't in γ is to see whether t occurs in $\gamma\,[u/t][\overrightarrow{\overline{k}}/\vec{v}]$, that is, the formula that results form γ by, first, replacing all occurrences of t with the variable u and then instantiating the variables $\vec{v}$ with $\overrightarrow{\overline{k}}$.

q-refer to sentences satisfying $\varphi(x)$, fulfilling (C3). (C) translates into $\forall x(\neg\tau(\varphi(x)) \vee \tau(\psi(x)))$, whose normalization is

$$\forall x([\neg\tau(\varphi(x))] \vee [\tau(\psi(x))]). \tag{11}$$

Thus, (C) q-refers to every sentence χ such that $\mathbb{N} \vDash \neg[\neg\tau(\varphi)][\ulcorner\chi\urcorner/x]$ or, equivalently, such that $\mathbb{N} \vDash \varphi[\ulcorner\chi\urcorner/x]$. The same can be said of sentences of the form $\forall x\varphi(x)$ where φ can be rewritten as a conditional expression, that is, where the normalization of φ (but not necessarily φ itself) is a disjunction. (D), in turn, translates into $\neg\forall x\neg(\tau(\varphi(x)) \wedge \tau(\psi(x)))$, whose normalization is

$$\neg\forall x([\neg\tau(\varphi(x))] \vee [\neg\tau(\psi(x))]). \tag{12}$$

Furthermore, clause 2(b) in Definition 4.9 guarantees that (C5) holds, for it's there to help us deal with subterms and nested q-reference. Given that the normalizations of (C) and (D) are (11) resp. (12), clause 2(b) guarantees that (C) and (D) q-refer to whatever sentences $\psi(\overline{n})$ m- (on condition that the term involved is a result instantiating x with $\overline{n}$) or q-refers to, provided that $\mathbb{N} \vDash \varphi(\overline{n})$. For instance, it entails that

$$\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow \mathsf{Bew}(\dot{\neg}x))$$

and

$$\exists x(x = \ulcorner 0 \neq 0\urcorner \wedge \exists y(y = \dot{\neg}x \wedge \mathsf{Bew}(y)))$$

q-refer not only to $0 \neq 0$ but also to its negation. Clause 2 also allows us to conclude that the sentences the weak diagonal lemma (Theorem 1.1), both in universal and in existential forms, delivers q-refer to themselves.

Of course, this carries the same 'problems' that affect m-reference: $\psi(\overline{n})$ refers not only to the sentence coded by n, if any, but also to every sentence denoted by a closed term occurring in $\psi(\overline{n})$. In particular, this means that (C) and (D) q-refer to every sentence whose code is equal or smaller that n, for every $n \in \omega$ satisfying $\varphi(x)$ in $\mathbb{N}$. Of course, which sentences these are will depend entirely on the coding.

The requirements that terms are closed in Definition 4.1 and that closed terms are 'new' in Definition 4.9 are to keep m- and q-reference apart. If a closed term t denoting a sentence χ already occurs in $\psi(x)$, and no new occurrence of t is generated by replacing x in $\psi(x)$ with $\overline{n}$, where $\mathbb{N} \vDash \varphi(\overline{n})$, then it doesn't seem right to conclude that (C) q-refers to χ, but only that it m-refers to χ. For the occurrence of t in $\psi(\overline{n})$ is not a product of instantiating the quantifier in (C) but was already there. For instance,

$$\forall x(x < \ulcorner\neg\varphi\urcorner \rightarrow \mathsf{Bew}(x))$$

doesn't q-refer to $\neg\varphi$; it only m-refers to $\neg\varphi$. On the other hand, if an open term $t(x)$ occurs in a sentence φ, it must do so in the scope of a quantifier $\forall x$ (in the normalization of φ). In that case, the occurrence of a closed term $t(\overline{n})$ denoting a sentence ψ is the result of instantiating $\forall x$, even if $\forall x(t(x) = \ulcorner\psi\urcorner)$ is true in $\mathbb{N}$. Thus, we say φ q-refers to ψ.

In addition, our definition of q-reference avoids the difficulties that Milne [22] pointed out. Let $\mathsf{Th} \subseteq \mathscr{L}$ be an unsound theory (with respect to $\mathbb{N}$) extending Q and χ a theorem of Th such that $\mathbb{N} \nvDash \chi$. Then, $\mathsf{Diag}'(x, y) := \mathsf{Diag}(x, y) \wedge \chi$ strongly represents the diagonalization relation *in* Th. However, as Milne notices, it doesn't seem right to claim that

$$\forall x(x = \ulcorner\forall y(\mathsf{Diag}'(x, y) \rightarrow \varphi(y))\urcorner \rightarrow \forall y(\mathsf{Diag}'(x, y) \rightarrow \varphi(y))) \tag{13}$$

refers to itself, for $\mathbb{N} \nvDash \mathsf{Diag}(\ulcorner\forall y(\mathsf{Diag}'(x, y) \rightarrow \varphi(y))\urcorner, (13)) \wedge \chi$, that is, (13) does not satisfy the antecedent of $\mathsf{Diag}'(\ulcorner\forall y(\mathsf{Diag}'(x, y) \rightarrow \varphi(y))\urcorner, y) \rightarrow \varphi(y)$. This is precisely

the reason why Definition 4.9 does not allow us to conclude that (13) q-refers to itself, despite the provability of the equivalence between this sentence and $\varphi(13)$ in Th or other theories. To conclude so, a concept of reference relative to a theory rather than an absolute, semantic notion like the one introduced here would be needed.[22]

It's also worth noting that q-reference is not a trivial notion. Although in many cases q-reference depends on how the formulae involved in the sentence really look like (and of course on the chosen coding), there are sentences of the form (C) that we can be sure do not q-refer to every sentence in the language. Take, for instance,

$$\forall x(x = \ulcorner\varphi\urcorner \rightarrow x \neq \neg\dot{}\ulcorner\varphi\urcorner),$$

where φ is a sentence. This simple expression just q-refers to φ, $\neg\varphi$, every sentence whose code is smaller than φ's, and nothing else.

Definition 4.9 also allows us to assess the q-reference of sentences of the form $\forall\vec{v}\varphi(\vec{v})$ and $\exists\vec{v}\varphi(\vec{v})$, where φ is preceded not just by one but by a string of quantifiers of arbitrary length. In that case, for instance, according to clause 2(a) sentences of the form

$$\forall\vec{x}(\varphi(\vec{x}) \rightarrow \psi(\vec{x}))$$

q-refer to every sentence that is an entry of an n-tuple satisfying φ (and $\neg\psi$). I opt for dismantling tuples satisfying the antecedent of sentences of this form to keep q-reference as a relation between sentences, instead of sentences on the one hand, and tuples of sentences on the other hand. This seems to be the most natural way of making sense of notions such as self-reference and well-foundedness that are introduced later in this section.

Like m-reference, q-reference is also closed under logical connectives. If φ q-refers to ψ, by Definition 4.9 φ must contain a subsentence of the form $\forall\vec{v}\chi$ satisfying clause 1 or 2. Then, so do $\neg\varphi$, $\varphi \wedge \delta$, $\varphi \vee \delta$, and $\varphi \rightarrow \delta$, for any sentence δ of the language. This implies that $\forall x(\varphi(x) \rightarrow \psi(x))$ and $\exists x(\varphi(x) \wedge \neg\psi(x))$ q-refer to the same sentences, as can be reasonably expected, given that the latter translates into $\neg\forall x\neg\neg(\varphi(x) \rightarrow \psi(x))$, whose normalization is the negation of $\forall x(\varphi(x) \rightarrow \psi(x))$'s.

Before I turn to general notions of reference and their derivatives, let me point out that the fact that we look into the set of permutations of the normalizations of sentences to assess q-reference entails that reference is closed under the translation τ, normalization and permutations of normalizations, as expected. This implies, for instance, that q-reference is closed under propositional transformations such as double negation, de Morgan laws, and the distributivity of conjunction over disjunction. It is easy to check that q-reference is also closed under the commutativity and associativity of conjunction, the renaming of variables, and the commutativity of quantifiers of the same kind.

DEFINITION 4.10 (Direct reference). *If φ, ψ are sentences, φ directly refers to ψ if and only if φ m- or q-refers to ψ.*

DEFINITION 4.11 (Chains of reference). *A sequence of sentences $\varphi_1, \ldots, \varphi_n$, with $n \in \omega$, is a chain of reference if and only if, for each $i < n$, φ_i directly refers to φ_{i+1}.*

DEFINITION 4.12 (Reference). *If φ, ψ are sentences, φ refers to ψ if and only if there is a chain of reference starting with φ and ending with ψ.*

Thus, reference is the transitive closure of direct reference, that is, the union of m- and q-reference. Condition (C8) is satisfied. If one does not find the transitivity of reference intuitively appealing, one can stick to direct reference rather than reference *simpliciter*.

[22] See [author].

Since both m- and q-reference are closed under negation, conjunction, disjunction, and implication, so is direct reference and, therefore, also reference, as required by (C6). Also, sentences delivered by weak and strong diagonalization (directly) refer to themselves, for they q- and m-refer to themselves, respectively.

Finally, a word on hyperintensionality. Reference as given by Definition 4.12—and, *a fortiori*, direct reference as well—is not closed under first-order logical equivalence, that is, it is hyperintensional. For instance, if 0 is not the code of a sentence, $0 = 0$ directly refers to no sentence, whereas $0 = 0 \vee \mathsf{Bew}(\ulcorner 0 \neq 0 \urcorner)$ refers to $0 \neq 0$. Similarly,

$$\forall x(x = x \rightarrow (x = \ulcorner 0 = 0 \urcorner \rightarrow x \neq \ulcorner 0 \neq 0 \urcorner))$$

refers to everything, but

$$\forall x(x = \ulcorner 0 = 0 \urcorner \rightarrow x \neq \ulcorner 0 \neq 0 \urcorner)$$

doesn't. Thus, (C9) is satisfied.

Nonetheless, reference and direct reference are closed under many logical transformations, as required by condition (C10). This is a consequence of the closure of m- and q-reference under these transformations, which I pointed out before. The following proposition offers some examples:

PROPOSITION 4.13. *The following pairs of sentences directly refer to the same sentences*:

1. $\forall v(\varphi \rightarrow \psi)$ *and* $\forall v(\neg\psi \rightarrow \neg\varphi)$,
2. $\exists v(\varphi \wedge \psi)$ *and* $\exists v(\psi \wedge \varphi)$,
3. $\forall v\varphi$ *and* $\forall u\neg\neg\varphi[u/v]$, *if* v *is free for* u *in* φ,
4. $\forall v\neg(\varphi \wedge \psi)$ *and* $\forall v(\neg\varphi \vee \neg\psi)$,
5. $\forall v(\varphi \wedge (\psi \vee \chi))$ *and* $\forall v((\varphi \wedge \psi) \vee (\varphi \wedge \chi))$,
6. $\forall v\forall u\varphi$ *and* $\forall u\forall v\varphi$.

We are able to define now three prominent patters of reference.

DEFINITION 4.14 (Direct self-reference). *A sentence* φ *is directly self-referential if and only if it directly refers to itself.*

DEFINITION 4.15 (Self-reference). *A sentence* φ *is self-referential if and only if it refers to itself.*

As expected, sentences delivered by the weak and the strong diagonal lemmata turn out to be directly self-referential according to Definition 4.14. Definition 4.15, on the other hand, can also account for the self-referential character of cycles of any length, such as the ones delivered by Proposition 3.2. Cycles given by pairs of sentences such as $\varphi(t)$ and $\forall x(\psi(x) \rightarrow \chi(x))$, where t denotes $\forall x(\psi(x) \rightarrow \chi(x))$ and $\varphi(t)$ is a ψ, or $\forall x(\varphi(x) \rightarrow \psi(x))$ and $\forall x(\gamma(x) \rightarrow \delta(x))$, where the former is a γ and the latter a φ, are self-referential as well. The notion of reference also allows us to define *a form* of well-foundedness.

DEFINITION 4.16 (Well-foundedness). *A sentence* φ *is well-founded if and only if there is a finite limit to the length chains of reference starting with* φ *can have.*

Obviously, all self-referential expressions are not well-founded. For given a chain of reference $\varphi, \varphi_1, \ldots, \varphi_n, \varphi$, we can extend it indefinitely with the sequence $\varphi_1, \ldots, \varphi_n, \varphi$, obtaining longer and longer chains of reference. Also, ω-chains delivered by Proposition 3.3 are not well-founded, for each sentence directly refers to the one coming next

and indirectly to all the ones that come after itself. Unfortunately, given the way they are obtained, their members are self-referential as well, albeit only indirectly. Let's take another look at the proof of Proposition 3.3. Given a formula φ with exactly one free variable, each sentence on the list is of the form $\varphi(\dot{s}(t, \dot{n}(S\overline{n})))$, with $n \in \omega$. Then, for every $n \in \omega$,

$$\varphi(\dot{s}(t, \dot{n}(S\overline{\#\varphi(\dot{s}(t, \dot{n}(S\overline{n})))}))) \tag{14}$$

is also on the list. Since $n < \#\varphi(\dot{s}(t, \dot{n}(S\overline{n})))$, $\varphi(\dot{s}(t, \dot{n}(S\overline{n})))$ refers to (14), as we have just established. But since the term $\overline{\#\varphi(\dot{s}(t, \dot{n}(S\overline{n})))}$ occurs in (14), it's also the case that (14) (directly) refers to $\varphi(\dot{s}(t, \dot{n}(S\overline{n})))$. Thus, by the transitivity of reference each sentence on the list is self-referential.

§5. In the last section I introduced notions of reference by mention, by quantification, direct reference, reference, self-reference, and well-foundedness. Now it's time to see how to put them to use. In particular, they should enable us to provide reasonable formulations of the metamathematical problems singled out in §2. The notion of direct self-reference should allow us to determine which sentences say of themselves that they are Rosser-provable, Σ_n-true, and Π_n-true, that is, we should be able to identify regular Henkin-Rosser sentences, Σ_n-truth tellers, and Π_n-truth tellers (with $n \neq 1$). In order to do that we need to spell out what it means for a directly self-referential sentence to ascribe a certain property to itself. As Halbach and Visser [10] point out, this task is highly nontrivial.

The main issue stems from the way properties are to be individuated. To determine whether a sentence of arithmetic ascribes to itself a property P we need to know what it takes for a formula $\varphi(x)$ to express P. Another way of putting it is the following: What does it take for two formulae $\varphi(x)$ and $\psi(x)$ of $\mathscr{L}$ to express the same property? The answer to this question depends on how intensional we believe properties should be. Unlike sets, pluralities, or classes, properties are usually considered to be intensional entities of some kind. As a consequence, it wouldn't be enough that $\varphi(x)$ and $\psi(x)$ are equivalent in $\mathbb{N}$, in PA, or in some other arithmetical theory for them to express the same property. Otherwise, $\mathsf{Bew}(x)$ and $\mathsf{Bew}^{\mathsf{R}}(x)$ as introduced in §2 would express the same property (provided that PA is ω-consistent). Clearly, a stronger notion of equivalence is required. On the other hand, we shouldn't go too far and claim that any syntactic difference in formulae implies a difference in the properties they express, for this is certainly too strong. For instance, it seems that $\varphi(x) \wedge \psi(x)$ and $\psi(x) \wedge \varphi(x)$ do express the same property, despite being different formulae.

I see at least two reasonable ways of understanding the equivalence between formulae expressing the same property: an intensional and a hyperintensional one. According to the former, logically equivalent formulae (with the same number of free variables) express the same property. This means, for instance, that $\varphi(x)$ and $\varphi(x) \wedge x = x$ express the same property. Consequently, every directly self-referential sentence ascribes to itself the property of being self-identical, as well as every other logical property expressed by a valid formula; as long as we accept that if φ ascribes property P to χ, then $\varphi \wedge \psi$ also ascribes property P to χ.

If all of this seems undesirable, as Halbach and Visser suggest, one might alternatively consider allowing only minor syntactic variations in formulae, such as the ones involved in normalizations. In that case, expressing a property would be a hyperintensional relation, but closed under some sort of relevant notion of logical equivalence. Unfortunately, this also has counterintuitive consequences. For example, $\mathsf{Bew}(x)$ and $\mathsf{Bew}(x) \vee \mathsf{Bew}(x)$ would not express the same property. Perhaps there is a way of including this kind of transformations

while excluding the undesirable ones, but I am so far sceptical about it. In any case, what it means for a formula of the language of arithmetic to express a property is beyond the scope of this paper. In this section I just provide two notions of self-ascription, each of which is based on one of the ways of understanding the equivalence between formulae expressing the same property considered in this section (cf. Definitions 5.1 and 5.2).

Setting this issue aside, there are several ways in which a sentence can ascribe the property P expressed by $\varphi(x)$ to itself. Before we turn to our definitions, it's important to distinguish the way we are interested in from other ways. On the one hand, a sentence can ascribe P to itself and, at the same time, to others, or it can ascribe P just to itself. For instance,

$$\forall x(\mathsf{Sent}(x) \rightarrow \neg\mathsf{Bew}(x)) \tag{15}$$

says of all sentences in $\mathscr{L}$ that they are unprovable in PA, including (15) itself, whereas PA's Gödel sentence $\neg\mathsf{Bew}(\ulcorner\gamma\urcorner)$ ascribes the same property just to itself. On the other hand, a sentence can ascribe a single property to itself (and perhaps other sentences), or it can ascribe a property to itself (and perhaps other sentences) and, at the same time, ascribe other properties to other sentences. For example, the strong diagonal lemma delivers true identities $t = \ulcorner t = t\urcorner$ and $s = \ulcorner s = s \wedge \mathsf{Bew}(\ulcorner 0 = 0\urcorner)\urcorner$. While $t = t$ just ascribes self-identity to itself, $s = s \wedge \mathsf{Bew}(\ulcorner 0 = 0\urcorner)$ ascribes self-identity to itself as well as provability to $0 = 0$. It seems the notion we are most interested in is that of a sentence ascribing a single property just to itself. However, given the lack of individual constants for numbers other than 0 and predicate symbols for properties other than identity in $\mathscr{L}$, it is frequently the case that one and the same sentence ascribes different properties to different sentences. Take, for example, $\mathsf{Bew}(\ulcorner\varphi\urcorner)$. It says of φ that is provable, but it also says something about each sentence whose code is smaller than φ. In the following two definitions of self-ascription we try to avoid this the best we can.

DEFINITION 5.1 (Self-ascription). *A sentence ψ ascribes the property expressed by $\varphi(x)$ to itself if and only ψ is directly self-referential and there's a sentence χ such that one of the following conditions holds*:

1. *χ is of the form $\varphi(t)$, it is logically equivalent to ψ, and $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.*
2. *χ is of the form $\forall\vec{v}_1(\neg\gamma_1(\vec{v}_1) \vee \cdots \vee \forall\vec{v}_n(\neg\gamma_n(\vec{v}_1, \ldots, \vec{v}_n) \vee \delta(\vec{v}_1, \ldots, \vec{v}_n))\ldots)$, it is logically equivalent to ψ and, for every $\vec{k}_1, \ldots, \vec{k}_n \in \omega$ such that $\mathbb{N} \vDash \gamma_i[\overline{\vec{k}}_1/\vec{v}_1]\ldots[\overline{\vec{k}}_n/\vec{v}_n]$, there's a term t such that $\delta[\overline{\vec{k}}_1/\vec{v}_1]\ldots[\overline{\vec{k}}_n/\vec{v}_n]$ is of the form $\varphi(t)$, and $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.*

In other, simpler but less accurate, words, ψ says of itself that it's a φ if ψ is of the form $\varphi(t)$ for some t denoting ψ modulo logical equivalence; or if it's roughly of the form $\forall v(\gamma(v) \rightarrow \delta(v))$ modulo logical equivalence, and for every $k \in \omega$ satisfying γ, there's a term t denoting ψ such that $\delta(\overline{k})$ is $\varphi(t)$.

If we didn't require ψ to be self-referential in Definition 5.1, every formula, even the non-self-referential ones, would ascribe some property to itself. For every formula ψ is, e.g., logically equivalent to $\psi \wedge \ulcorner\psi\urcorner = \ulcorner\psi\urcorner$. In turn, clause 2 of Definition 5.1 is intended to guarantee that ψ is logically equivalent to a sentence that only says of ψ that it satisfies $\varphi(x)$. However, the contrary is very often unavoidable. Consider the following identity:

$$t = \ulcorner\forall x(x = t \rightarrow x = x)\urcorner.$$

$\forall x(x = t \rightarrow x = x)$ ascribes the property of being self-identical to itself. But it is logically equivalent to $\forall x(x = x \rightarrow x = x)$ that ascribes the same property to every

sentence. Finally, note that the reason why we write δ in Definition 5.1 instead of φ is that δ could be different from φ but of the form $\varphi(s(x))$ for some open term $s(x)$, and $t = s(\overline{k})$ for all $k \in \omega$ satisfying the γ_i. For example, let $\neg\psi$ be logically equivalent to

$$\forall x(x = t \rightarrow \mathsf{Bew}(\dot{\neg}x)),$$

where t denotes ψ. $\neg\psi$ ascribes to itself the property expressed by $\mathsf{Bew}(x)$ here, but it's $\mathsf{Bew}(\dot{\neg}x)$, the formula that acts as δ in Definition 5.1.

DEFINITION 5.2 (Hyperintensional self-ascription). *A sentence ψ ascribes the hyperintensional property expressed by $\varphi(x)$ to itself if and only if there's a sentence χ such that one of the following conditions holds*:

1. *χ is of the form $\varphi(t)$, belongs to the set containing the formula that results from applying (the normalizing) clauses 1–3 of Definition 4.5 to $\tau(\psi)$, and is closed under renaming of variables, commutativity and associativity of disjunction, conjunction, and the universal quantifier, and $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.*
2. *χ is of the form $\forall\vec{v}_1(\neg\gamma_1(\vec{v}_1) \vee \cdots \vee \forall\vec{v}_n(\neg\gamma_n(\vec{v}_1, \ldots, \vec{v}_n) \vee \delta(\vec{v}_1, \ldots, \vec{v}_n))\ldots)$, belongs to the set containing the formula that results from applying clauses 1–3 of Definition 4.5 to $[\tau(\psi)]$, and is closed under renaming of variables, commutativity and associativity of disjunction, conjunction, and the universal quantifier; and, for every $\vec{k}_1, \ldots, \vec{k}_n \in \omega$ such that $\mathbb{N} \vDash \gamma_i[\vec{\overline{k}}_1/\vec{v}_1]\ldots[\vec{\overline{k}}_n/\vec{v}_n]$, there's a term t such that $\delta[\vec{\overline{k}}_1/\vec{v}_1]\ldots[\vec{\overline{k}}_n/\vec{v}_n]$ is of the form $\varphi(t)$ and $\mathbb{N} \vDash t = \ulcorner\psi\urcorner$.*

Clearly, hyperintensional self-ascription entails self-reference, for only transformations that don't add or remove atoms are allowed in formulae expressing a certain property. Moreover, hyperintensional self-ascription entails self-ascription *simpliciter*, as expected.

Let's look at some examples. Clearly, all sentences obtained by (weakly) diagonalizing a predicate $\varphi(x)$, that is,

$$\forall x(x = \ulcorner\forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))\urcorner \rightarrow \forall y(\mathsf{Diag}(x, y) \rightarrow \varphi(y))),$$

hyperintensionally ascribe to themselves the property expressed by φ and, *a fortiori*, they also ascribe this property to themselves *simpliciter*. The same can be said of sentences that result from an application of the strong diagonal lemma to $\varphi(x)$, for they satisfy an identity of the form $t = \ulcorner\varphi(t)\urcorner$. As a consequence, both the weak and the strong Gödel sentences can be said to ascribe the property expressed by $\neg\mathsf{Bew}(x)$ to themselves.

Analogously, sentences that result form an application of either the weak or the strong diagonal lemma to $\mathsf{Bew}^{\mathsf{R}}(x)$ ascribe the property expressed by $\mathsf{Bew}^{\mathsf{R}}(x)$ to themselves, so they are all Henkin-Rosser sentences. In contrast, neither $0 = 0$ nor $0 \neq 0$ turn out to be Henkin or Henkin-Rosser sentences. Of course, nothing I said here precludes the existence of other Henkin-Rosser sentences. They could result from the application of alternative diagonalization procedures or be more 'accidental', as Halbach and Visser would put it. As I argued before, this is not an issue but, on the contrary, a desirable feature.

Likewise, genuine Σ_n- and Π_n-truth tellers (with $n \neq 1$) can be obtained by weakly or strongly diagonalizing the predicates $\mathsf{T}_{\Sigma_n}(x)$ and T_{Π_n}, respectively, whereas $0 = 0$, $0 \neq 0$, and other trivial fixed points do not qualify as truth tellers, for they don't ascribe any of the properties expressed either by $\mathsf{T}_{\Sigma_n}(x)$ or by T_{Π_n} to themselves.

The notions of reference introduced in §4 are non-trivial and intuitively appealing. Moreover, they have proved to be useful for the formulation of the metamathematical problems indicated by Halbach and Visser. Although it is not straightforward how to extend

the new notions to other languages, for instance, not containing individual constants or a standard interpretation—such as the language of set theory or the extension of $\mathscr{L}$ with a truth predicate, I hope they will shed light on investigations of reference for other formal and perhaps even natural languages.

§6. Acknowledgments. I am deeply indebted to Volker Halbach, with whom I had countless fruitful discussions on reference and self-reference over the last five years. I would also like to particularly thank Thomas Schindler, for great suggestions and encouragement, especially when it came to proofs. I should mention as well Eduardo Barrio, Catrin Campbell-Moore, Roy T. Cook, Martin Fischer, Hannes Leitgeb, Øystein Linnebo, Johannes Stern, Albert Visser, the Buenos Aires Logic Group, the MCMP logic community, and the Oxford logic group. Finally, I am grateful to two anonymous referees for the sensible improvements they proposed.

BIBLIOGRAPHY

[1] Boolos, G., Burgess, J. P., & Jeffrey, R. C. (2007). *Computability and Logic* (fifth edition). Cambridge: Cambridge University Press.

[2] Carnap, R. (1937). *Logical Syntax of Language*. London: Routledge.

[3] Cook, R. T. (2006). There are non-circular paradoxes (but Yablo's isn't one of them!). *The Monist*, **89**, 118–149.

[4] Cresswell, M. J. (1975). Hyperintensional logic. *Studia Logica*, **34**, 25–38.

[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] Hájek, P. & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin: Springer.

[8] Halbach, V. (2009). Reducing compositional to disquotational truth. *Review of Symbolic Logic*, **2**, 786–798.

[9] Halbach, V. (2016). The root of evil. A self-referential play in one act. In van Eijck, J., Iemhoff, R., and Joosten, J. J., editors. *Liber Amicorum Alberti. A Tribute to Albert Visser*. London: College Publications, pp. 155–163.

[10] Halbach, V. & Visser, A. (2014a). Self-reference in arithmetic I. *Review of Symbolic Logic*, **7**, 671–691.

[11] Halbach, V. & Visser, A. (2014b). Self-reference in arithmetic II. *Review of Symbolic Logic*, **7**, 692–712.

[12] Heck, Jr., R. (2007). Self-reference and the languages of arithmetic. *Philosophia Mathematica*, **III**, 1–29.

[13] Henkin, L. (1952). A problem concerning provability. *The Journal of Symbolic Logic*, **17**, 160.

[14] Henkin, L. (1954). Review of G. Kreisel: On a problem of Henkin's. *The Journal of Symbolic Logic*, **19**, 219–220.

[15] Horwich, P. (1998). *Truth* (second edition). New York: Blackwell.

[16] Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford: Clarendon Press.

[17] Kleene, S. (1938). On notation for ordinal numbers. *The Journal of Symbolic Logic*, **3**, 150–155.

[18] Kreisel, G. (1953). On a problem of Henkin's. *Indagationes Mathematicae*, **15**, 405–406.

[19] 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.

[20] Leitgeb, H. (2005). What truth depends on. *Journal of Philosphical Logic*, **34**, 155–192.

[21] Löb, M. H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, **20**, 115–118.

[22] Milne, P. (2007). On Gödel sentences and what they say. *Philosophia Mathematica*, **III**(15), 193–226.

[23] Montague, R. (1962). Theories incomparable with respect to relative interpretability. *The Journal of Symbolic Logic*, **27**, 195–211.

[24] Priest, G. (1997). Yablo's paradox. *Analysis*, **57**, 236–242.

[25] Putnam, H. (1958). Formalization of the concept of about. *Philosophy of Science*, **25**, 125–130.

[26] Ryle, G. (1933). Imaginary objects. *Proceedings of the Aristotelian Society*, **12**(Suppl.), 18–43.

[27] Smoryński, C. (1981). Fifty years of self-reference in arithmetic. *Notre Dame Journal of Formal Logic*, **22**(4), 357–374.

[28] Smoryński, C. (1991). The development of self-reference: Löb's theorem. In Drucker, T., editor. *Perspectives on the History of Mathematical Logic*. Boston: Birkhäuser, pp. 110–133.

[29] Sorensen, R. A. (1998). Yablo's paradox and kindred infinite liars. *Mind*, **107**, 137–155.

[30] Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica Commentarii Societatis Philosophicae Polonorum*, **1**, 261–405, reprinted as *The Concept of Truth in Formalized Languages* in ***Logic, Semantics and Metamathematics***, pp. 152–278.

[31] Tarski, A. (1944). The semantic conception of truth: And the foundations of semantics. *Philosophy and Phenomenological Research*, **4**, 341–376.

[32] Urbaniak, R. (2009). Leitgeb, "About," Yablo. *Logique et Analyse*, **207**, 239–254.

[33] Visser, A. (1989). Semantics and the liar paradox. In Gabbay, D. M. and Günthner, F., editors. *Handbook of Philosophical Logic*, Vol. 4. Dordrecht: Reidel, pp. 617–706.

[34] Yablo, S. (1985). Truth and reflexion. *Journal of Philosphical Logic*, **14**, 297–349.

[35] Yablo, S. (1993). Paradox without self-reference. *Analysis*, **53**, 251–252.

MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY
LMU MUNICH
MUNICH, GERMANY
E-mail: Lavinia.Picollo@lrz.uni-muenchen.de