Logique & Analyse **177–178** (2002), 3–14

WHAT IS A SELF-REFERENTIAL SENTENCE? CRITICAL REMARKS ON THE ALLEGED (NON-)CIRCULARITY OF YABLO'S PARADOX

HANNES LEITGEB

1. *Introduction*

Since Stephen Yablo (1993) has introduced his 'Paradox without self-reference' there has been a lively discussion on whether Yablo's paradox is self-referential or not (see Sorenson (1998) and Beall (2001a), where an overview of further contributions can be found; see also Beall (1999), and Beall and Colyvan (2001b), footnote 3 on p. 402). The discussion has not been about the *meaning* of the general term 'self-referential' but rather about its *extension* with respect to truth-paradoxical sentences involving a quantification over an infinite sequence of sentences. My aim is to point out that the whole discussion is substantially flawed because (i) two different notions of self-referentiality and circularity have been used (and sometimes mixed up) in the discussion, and, worse, (ii) both notions are unclear and inadequate as explications of our pre-theoretical term 'self-referential'. This is not just relevant in itself, but it rather points to a much more general claim: we hypothesize that (iii) any non-sophisticated explication of self-referentiality is expected to fail.

At first we will outline the two notions referred to in (i) and (ii) — which has not been done by the debaters themselves — and of course we will try to be as fair as possible in our reconstruction. If there is a standard view on self-referentiality at all, the first notion is probably closer to it than the second one, and the former may also be more easily and less vaguely described than the latter. However, we are unable to offer clear-cut definitions for either of them (which is part of the problem). Besides the difficulties of proper formulation, we will show that both notions are strongly deficient concerning what they should be explications for. This is not meant to entail that the discussion about Yablo's paradox has been futile or amiss. In the contrary, the discussion teaches us that we do not have a conception of semantical circularity as clear as we would have thought to have, and as we should have. We also do not want to blame the debaters for a lack of rigorous explication concerning the notions of self-referentiality and circularity, since

that has definitely not been their topic. We just want to stress that the question of whether Yablo's paradox is circular or not, is not to be settled without much more theoretical elaboration. We finish the paper with the following open question: what might a formally correct and materially adequate definition of self-referentiality look like? Even if the machinery of some of the renowned approaches to truth and circularity, which has been developed after Kripke's (1975) paper, is to be applied in order to answer this question, it is by far not clear how this is done satisfyingly, or so we will argue. But the question is urgent since any answer to it might improve our understanding of the semantics of natural languages essentially[1], independently of whether the question is answered by an adequate explication of self-referentiality, or by a sound argument in favour of the rejection of the question as being ill-posed.

2. *A Notion of Self-Referentiality*

Let us first focus on some of the usual instances of self-referential sentences containing a truth predicate 'Tr'. Let, e.g., 'a' denote the sentence '$Tr(a)$', and let 'b' denote the sentence '$\neg Tr(b)$', i.e., $a =$ '$Tr(a)$' and $b =$ '$\neg Tr(b)$'. In this case, both a and b are usually said to refer to themselves (and to nothing else), such that a says of itself that it is true, while b says of itself that it is not true (a may thus be called a 'truth-teller', and b a '(strengthened) liar'). This usage of 'refers to' and 'says of' seems to presuppose that the usual reference relation ref, which holds between (singular or general) terms and their referents, is extended or complemented by a reference relation holding between *sentences* and objects (but where the referents of the sentences are not the truth values of the sentences).

In a first approximation, a singular sentence might thus be defined to refer to all the referents of all of its singular terms, and only to them. This is, presumably, the most common way of thinking about self-referentiality. If ref_1 is the so-defined reference relation for sentences, and if ref is the usual reference relation for terms, we can define: $ref_1(x, y) \leftrightarrow_{\text{df}}$ x *is a sentence* $\wedge$ $\exists z(z$ *is a singular term* $\wedge$ x *contains* $z \wedge z\ ref\ y)$. Since 'a' ref '$Tr(a)$',

[1] But, of course, a thorough explication of self-referentiality for sentences would also improve our understanding of the semantics of *formal* languages. This paper has e.g. been stimulated by a study (Leitgeb (2001)) of axiomatic theories of truth which extend arithmetic but which have no standard models, where one regretfully becomes aware of the lack of such an explication. The problems of self-referentiality that are raised in the following sections are even more difficult in the context of such non-standard theories of truth, since already the reference of terms is non-standard and thus unclear. However, we are going to omit any discussion of *that* type of problem.

'b' ref '$\neg Tr(b)$', and since '$Tr(a)$' and '$\neg Tr(b)$' are sentences which contain the singular terms 'a' and 'b' respectively, we have that '$Tr(a)$' ref_1 '$Tr(a)$', and '$\neg Tr(b)$' ref_1 '$\neg Tr(b)$'. If we finally define self-referentiality in the way that $selfref_1(x) \leftrightarrow_{\text{df}} x\ ref_1\ x$, it immediately follows that $selfref_1$('$Tr(a)$') and $selfref_1$('$\neg Tr(b)$') just as expected.

Moreover, if ref_1^* is the transitive closure of ref_1, i.e., if the extension of 'ref_1^*' is the smallest superset of the extension of 'ref_1' having the property that if $x\ ref_1^*\ y$, $y\ ref_1^*\ z$, then also $x\ ref_1^*\ z$, we can define the circularity of sentences in the way that $circular_1(x) \leftrightarrow_{\text{df}} x\ ref_1^*\ x$. E.g., let $c =$ '$\neg Tr(d)$' and $d =$ '$\neg Tr(c)$' (this is usually called a 'liar cycle'): thus, $c\ ref_1\ d$, $d\ ref_1\ c$, therefore $c\ ref_1^*\ c$ and $d\ ref_1^*\ d$, and it follows that $circular_1(c)$ and $circular_1(d)$ though $\neg selfref_1(c)$ and $\neg selfref_1(d)$. If furthermore 'e', besides 'b' itself, also denotes '$\neg Tr(b)$', we have that $self\ ref_1(e)$, because $e = b$, but $\neg circular_1$ ('$Tr(e)$') and thus also $\neg selfref_1$ ('$Tr(e)$'). If one regarded the latter as being unintended, our definitions might be easily adapted accordingly.

Now let us turn to Yablo's paradox, or rather a version of it, and thus to sentences which are not singular: let $s_0, s_1, s_2, \ldots$ be an infinite sequence of sentences, such that

$$\begin{aligned} s_0 &= \text{'}\forall x(P_1(x) \rightarrow \neg Tr(x))\text{'},\\ s_1 &= \text{'}\forall x(P_2(x) \rightarrow \neg Tr(x))\text{'},\\ s_2 &= \text{'}\forall x(P_3(x) \rightarrow \neg Tr(x))\text{'}, \ldots \end{aligned}$$

Let us assume that the extension of every n-th antecedent predicate (with $n = 1, 2, 3, \ldots$) in the sequence above is the set $s_n, s_{n+1}, s_{n+2}, \ldots$ of sentences. In this case, every sentence s_n is usually said to refer to each of the sentences $s_{n+1}, s_{n+2}, \ldots$ (and to nothing else), such that s_n says precisely of each of the latter sentences that it is untrue. More generally, if '$A[x]$' and '$B[x]$' are formulas which contain 'x' as the only free variable, and which contain no further singular terms, every general sentence of the form '$\forall x(A[x] \rightarrow B[x])$' might be defined to refer to all and only As, i.e.: '$\forall x(A[x] \rightarrow B[x])$' $ref_1\ y \leftrightarrow_{\text{df}} A[y]$. According to our definitions from above, no sentence s_n is self-referential or circular, which is precisely what has been claimed by Yablo and what has been re-emphasized by Sorenson (1998). This indicates that our definitions of $selfref_1$ and $circular_1$ are close to what Yablo, and probably the majority of philosophers, seem to understand by 'self-referential' and 'circular'. The driving idea behind these definitions is to define a reference relation for sentences on the basis of the reference relation for the terms occurring within the sentences. This constitutes the first notion of circularity that has been used in the discussion on Yablo's paradox; according to this notion the paradox turns out to be non-circular.

3. *Another Notion of Self-Referentiality*

Yablo's claim that his paradox without self-reference 'is not in *any* way circular' (Yablo (1993), p. 251; his italics) has evoked dissenting comments by Priest (1997) (in turn, Sorenson (1998) defends Yablo's claim against Priest's attack) and Beall (2001a); the latter — as we want to argue now — employing a different notion of circularity. Since this second notion is perhaps not as familiar as the first one, we will quote some typical passages by Priest and Beall in this section in order to expose its underlying intuitions. Let us present Priest's and Beall's view first with respect to the liar sentence again: Priest calls our attention to the fact that the name of the liar sentence occurs on both sides of an equation, i.e., $b =$ '$\neg Tr(b)$', which 'makes it a fixed point of a certain kind, and, in this context, codes the self-reference' (Priest (1997), p. 236). As Priest rightly adds, the presence of a fixed point does not immediately follow from the equation itself, since, strictly speaking, 'b' occurs on the right-hand side within an opaque context; it rather follows from the fact that b is a fixed point of a certain syntactical mapping, say, f. In the arithmetical context, where expressions are 'identified' with their Gödel codes, f maps the code of each formula A to the code of the formula which is the result of concatenating '$\neg Tr$' and parentheses with the numeral $\underline{A}$ of the code of A. E.g., neglecting the reference to codes, $f($'$\neg Tr(b)$'$) =$ '$\neg Tr(\underline{\neg Tr(b)})$'. Thus b is a fixed point of f, or, rather, b is a fixed point of f *up to arithmetical equivalence*, i.e., the formula encoded by the f-image of the code of b is *equivalent* to b in the standard model of arithmetic, i.e., it is arithmetically true that $\neg Tr(\underline{\neg Tr(b)}) \leftrightarrow \neg Tr(b)$ (given that 'Tr' has been added to the arithmetical language).

Stated crudely, a sentence might thus be defined as circular, such that $circular_2(x) \leftrightarrow_{\text{df}}$ *x is a sentence* $\wedge \exists f(f$ *is a syntactical mapping* $\wedge f(x) = x)$, where the fixed point property expressed by the last conjunct of the definiens might again be relaxed in some way. Alternatively, but following the same line of thought, circularity might be defined in the way that $circular_{2'}(x) \leftrightarrow_{\text{df}}$ *x is a sentence* $\wedge \exists y \exists z \exists f(y$ *is a term* $\wedge$ *x contains* $y \wedge$ $y\ ref\ z \wedge f$ *is a syntactical mapping* $\wedge f(z) = z)$. Since '$\neg Tr(b)$' is a sentence which is the fixed point of a syntactical mapping as described by Priest, we have that $circular_2($'$\neg Tr(b)$'$)$ as expected. Since '$\neg Tr(b)$' contains the term 'b', 'b' ref '$\neg Tr(b)$', and '$\neg Tr(b)$' is a fixed point of a syntactical mapping as sketched before, it also follows that $circular_{2'}($'$\neg Tr(b)$'$)$. Similar claims hold for the truth-teller sentence from above. We omit a discussion of the other examples of the last section for reasons which we are going to explain when we turn to our critical examination below. Note that circularity is now defined directly rather than by a detour via the reference (or self-reference) of sentences. If intended, the definitions of '$circular_2$'

and '$circular_{2'}$' might be strengthened further to encompass also transitive closure (just as in the case of 'ref_1^*').

Turning again to the infinite sequence of sentences involved in Yablo's paradox, Priest first notes 'that each sentence refers to (quantifies over) only sentences later in the sequence. No sentence, therefore, refers to itself, even in an indirect, loop-like fashion. There seems to be no circularity' (Priest (1997), p. 237). This statement obviously concerns the notions of self-referentiality and circularity that we have sketched in §2. But then Priest reconstructs Yablo's paradox in an arithmetical setting: by Gödel's (generalized) diagonalization lemma, there is a formula '$s(x)$', such that $s(x) \leftrightarrow \forall k(k > x \rightarrow \neg Sat(k, \underline{s(x)}))$, where Sat is the usual satisfaction predicate. The latter equation 'shows that we have a fixed point ... here, of exactly the same self-referential kind as in the liar paradox' (Priest (1997), p. 238). $s(n)$ is (arithmetically equivalent to) what may be regarded as the n-th sentence of Yablo's infinite sequence of sentences. In this way, s may as well be regarded as a function, such that 'The function s is defined by specifying each of its values, but each of these is defined with respect to s ... It is now the function s that is a fixed point. s is the function which, applied to any number, gives the claim that all claims obtained by applying s *itself* to subsequent numbers are not true. Again the circularity is patent.' (Priest (1997), p. 239; his italics). Priest concludes: 'As we see, then, Yablo's paradox does involve circularity of a self-referential kind.' (Priest (1997), p. 242). This matches our provisional definition of '$circular_{2'}$' above, since, e.g.: '$\forall k(k > 0 \rightarrow \neg Sat(k, \underline{s(x)}))$' contains the term '$\underline{s(x)}$', '$\underline{s(x)}$' is the numeral of the code of the formula '$s(x)$', thus (modulo coding) '$\underline{s(x)}$' ref '$s(x)$', but '$s(x)$' is a fixed point of an appropriate mapping as sketched by Priest, and therefore it follows that $circular_{2'}$('$\forall k(k > 0 \rightarrow \neg Sat(k, \underline{s(x)}))$'). It is not clear to us whether a similar claim might be put forward concerning circularity in the sense of '$circular_2$', but that does not matter now. Priest, and also Beall, argue that all other formulations of Yablo's paradox may be shown circular in a similar way, independently of whether they have been formulated by means of a fixed point formula '$s(x)$' or not. The only difference between the different formulations is that in one case the circularity might be hidden more effectively than in another. 'Given all this, it follows that the reference of 'Yablo's paradox' is ... a circular sequence — a sequence containing fixed points, self-reference, etc.' (Beall (2001a)). Thus we hope that our definitions of $circular_2$ and $circular_{2'}$ are close to what Priest and Beall seem to understand by 'circular'. The driving idea behind the definitions is to define circularity by the truth of certain fixed point equations, where 'equation'

is understood broadly. This constitutes the second notion of circularity involved in the discussion on Yablo's paradox; according to this notion the paradox turns out to be circular indeed.

4. *Why Both Notions are Deficient*

Before we outline the weaknesses of the two notions of circularity introduced, let us first comment on a critical point concerning the latter notion. The notion of circularity which applies to Yablo's paradox according to Priest, has been questioned by Sorenson for the following reason: Sorenson points out that 'If I set up a sequence with a self-referential description, it does not follow that the *content* of what I specified is self-referential' (Sorenson (1998), p. 148; his italics). According to Sorenson, if Priest calls Yablo's paradox circular, this is so because Priest mixes up the indeed circular description of the paradox with the circularity of the paradox itself. Beall defends Priest's view against this remark of Sorenson's by claiming that 'The point, rather, is that any description D, used to fix the reference of 'Yablo's paradox' is such that D's satisfaction conditions require that the ... satisfier is circular (contains self-reference, a fixed point, etc.)' (Beall (2001a)). Beall thus seems to say that the members of Yablo's sequence are $circular_{2'}$ and not just *described* in a circular way. Although this issue is probably not finally settled, let us presume in the following that it is, and, say, in favour of Priest and Beall, since this is not the point which we are mainly interested in.

But let us now turn to the evaluation of the definitions of self-referentiality and circularity above. First of all, the 'definitions' are, at best, incomplete and/or vague, a fact for which, prima facie, the one who has stated them is to be blamed, i.e., *we* have to be blamed: the first notion of circularity has only been specified for singular sentences on the one hand, and universal conditional sentences on the other. It is clear that what we really aim at is a definition of '$selfref_1(x)$' and '$circular_1(x)$' where 'x' may denote *any* kind of sentence. Concerning our second notion of circularity, we actually should specify much more precisely what is meant by a fixed point *modulo coding, arithmetical equivalence, etc.*, and what kinds of syntactical mappings f are referred to in our to-be-a-definition of '$circular_2$' and '$circular_{2'}$' (we will return to this latter point below). Due to the lack of precise formulation it also impossible to compare the two notions of circularity seriously. In a nutshell, the two notions of circularity are still in a pre-theoretical state, which is *in itself* no problem, since this only brings about a request for further analysis. The actually urgent and remaining question is whether one of the notions of self-referentiality and circularity sketched above may be transformed into a serious scientific concept *at all*.

This may be called in doubt. Consider '$selfref_1(x)$' and '$circular_1(x)$': what is conspicuous about them is that they do not satisfy the following Equivalence Condition (EC): if A is self-referential/circular, and if B is logically equivalent to A, then also B is self-referential/circular. 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. In particular, if the extension of a predicate should be, say, free from any self-referential sentence but at the same time logically closed, a failure of EC might cause substantial problems. But EC indeed fails for the definitions of §2: e.g., if $b' =$ '$(P(a) \vee \neg P(a)) \vee \neg Tr(b')$', then b' is obviously $selfref_1$, although the logically equivalent '$P(a) \vee \neg P(a)$' is not. Accordingly, '$\forall x((A[x] \vee \neg A[x]) \rightarrow (A[x] \rightarrow B[x]))$' is $selfref_1$ since it satisfies (as any object does) '$A[x] \vee \neg A[x]$', although the logically equivalent '$\forall x(A[x] \rightarrow B[x])$' is not necessarily so, depending on the extension of '$A[x]$'. The sequence members of Yablo's paradox, as we have presented it in §2, are not $selfref_1$, however, there are logically equivalent formulations which are indeed $selfref_1$. This is particularly annoying if one thinks of the different 'versions' of Yablo's paradox (or paradoxes?) which may be found in the literature: e.g., our presentation is different from Yablo's original one, both are different from Priest's version above, and so on. Some of these version might turn out to be circular, some not. The immediate response to this problem is to revise our definition of self-referentiality (and of circularity, accordingly) such that, say, $selfref'_1(x) \leftrightarrow_{\text{df}} \exists y(y$ *is a sentence* $\wedge$ y *is logically equivalent to* $x \wedge y\ ref_1\ y)$. Indeed, we should even liberalize the notion of logical equivalence in the definiens furthermore to 'equivalent in the standard model of arithmetic (under arbitrary interpretations of 'Tr')', or even to 'equivalent according to (i.e., derivable from) Peano arithmetic or some proper fragment of the latter', since otherwise no philosopher may any longer argue in the following way: 'By Gödel's diagonalization lemma, we know that there is a sentence A such that A is equivalent to '$\neg Tr(\underline{A})$' in arithmetic. Thus there is a self-referential sentence, that is, A.' The problem is that such an A's being arithmetically equivalent to '$\neg Tr(\underline{A})$' does not *necessarily* entail A's being $selfref_1$ or $selfref'_1$. However, already the seemingly small alteration of the definition of '$selfref_1(x)$' in terms of *logical* equivalence has the effect that $selfref'_1$('$P(a) \vee \neg P(a)$') and $selfref'_1$('$\forall x(A[x] \rightarrow B[x])$') by the same reasoning as before, and, more generally, *every* sentence might turn out to be self-referential.

These considerations sound strikingly familiar: when Hempel (1945) criticized Nicod's criterion of confirmation as being unacceptable, he did so for essentially two reasons: (i) 'First, the applicability of this criterion is restricted to hypotheses of universal conditional form' (p. 10); our definition of '$selfref_1(x)$' suffers from similar restrictions. (ii) 'Nicod's criterion makes confirmation depend not only on the content of the hypothesis, but also on its formulation' (p. 11); the same holds for our definition of '$selfref_1(x)$'. Hempel thus suggested to adopt Nicod's criterion just as a sufficient (but not necessary) condition of confirmation. However, if the intuitively plausible equivalence condition is added, one is stuck with the so-called paradoxes of confirmation. Similarly, we seem to be stuck with what might be called the paradoxes of self-referentiality, but this time not with the well-known self-referential and paradoxical sentences, but rather with a paradox affecting the notion of self-referentiality itself.

We could try to handle this problem by redefining self-referentiality in a way such that an *arithmetical* equivalence condition, analogous to EC, is satisfied, but where a self-referential sentence is not demanded to be equivalent to *some* sentence being self-referential in the sense of '$selfref_1(x)$', but where *all* of its equivalents are demanded to be so: $selfref_1''(x) \leftrightarrow_{\text{df}} \forall y(y \textit{ is a sentence} \wedge y \textit{ is arithmetically equivalent to } x \rightarrow y\ ref_1\ y)$. But that does not solve the problem either, as may be seen from an example employing 'deferred ostension' (compare Quine (1969), p. 40): consider again the sentence C = '$\forall x(A[x] \rightarrow B[x])$', and assume that it is $selfref_1$, i.e., the extension of '$A[x]$' contains C. Assume furthermore that '$A^*[x]$' is another formula, such that the extension of '$A^*[x]$' is disjoint from the extension of '$A[x]$' and does not contain linguistic items. Finally, let the extension of '$R[x, y]$' be a relation holding precisely between all the members of the extensions of '$A^*[x]$' and of '$A[x]$'. We use the first relata of R to refer to the second ones: '$\forall x(A^*[x] \rightarrow \forall y(R[x, y] \rightarrow B[y]))$' is intuitively equivalent to C above, but not $selfref_1$. If R is a mapping, we can write the equivalent sentence more succinctly as: $\forall x(A^*[x] \rightarrow B[R(x)])$. If R is arithmetical, C is not $selfref_1''$. Now we do not face the problem that every sentence turns out to be self-referential, but rather that (perhaps) no sentence turns out self-referential at all.

The concepts of circularity developed in §3 suffer again from being too liberal: here the problem is that virtually *every* sentence is the fixed point of *some* mapping, in particular if only equivalence in one or another sense is demanded. E.g.: let g map the code of each formula A to the code of the formula which is the result of concatenating the numeral $\underline{A}$ with the equality sign, $\underline{A}$ again, the conjunction sign, and finally A itself. Neglecting the reference to codes, it follows that, e.g., $g(\text{'}P(a)\text{'}) = \text{'}\underline{P(a)} = \underline{P(a)} \wedge P(a)\text{'}$. Thus '$P(a)$' is a fixed point of g, or, rather, it is a fixed point of g *up to*

arithmetical equivalence, i.e., the formula encoded by the g-image of the code of '$P(a)$' is *equivalent* to '$P(a)$' in the standard model of arithmetic, i.e., it is arithmetically true that $(\underline{P(a)} = \underline{P(a)} \wedge P(a)) \leftrightarrow P(a)$ (given that 'P' is part of the arithmetical language or has been added to the latter). Therefore, '$P(a)$' is $circular_2$. Moreover, consider any arithmetical formula '$t(x)$' whatsoever which has precisely one free variable. It follows that $t(x) \leftrightarrow \forall k(k > x \rightarrow (\underline{t(x)} = \underline{t(x)} \wedge t(x)))$ (we have added the vacuous quantification clause just in order to enhance the similarity to Priest's fixed point clause for '$s(x)$' above) and thus $circular_{2'}$('$\forall k(k > 0 \rightarrow (\underline{t(x)} = \underline{t(x)} \wedge t(0)))$'). As a final example, consider deferred ostension again: assume that $u =$ '$Tr(v)$', and that $v =$ '$2 + 2 = 4$'. Let h be any (say, arithmetical) mapping such that $h(u) = v$. Then '$Tr(v)$' is certainly equivalent to '$Tr(h(u))$', i.e., arithmetically equivalent under appropriate coding, therefore $Tr(v) \leftrightarrow Tr(h(\text{'}Tr(v)\text{'}))$, which implies that u is a fixed point of a certain kind and thus also $circular_2$. But intuitively u just says that '$2 + 2 = 4$' is true. Summing up, a proper characterization of what a circular sentence is demanded to be a fixed point of, or what a component of a circular sentence is demanded to be a fixed point of, and what this means precisely, is vital for the second approach to circularity. However, at least *we* do not see any obvious way to state such a characterization. This is also the reason why we did not check more examples in §3 for being circular, since *everything* turns out to be circular in the sense of §3, including Yablo's paradox.

5. *What Might a Non-Deficient Notion of Self-Referentiality Look Like?*

We leave the discussion with the following open question: *what might a formally correct and materially adequate definition of self-referentiality look like?* Perhaps the definitions stated in §2 and §3 may be repaired in some way. E.g., instead of changing our definition of '$selfref_1(x)$' by adding a logical equivalence condition, we might rather use something else, e.g., being intensionally isomorphic in the sense of Carnap or the like. In any case, if we wanted to follow the lines of §2 and §3, a much more sophisticated account of self-referentiality would have to be developed in order to classify a linguistic construction as ingenious as Yablo's paradox as being self-referential or not.

Perhaps there is an account of self-referentiality which has been introduced independently of the discussion about Yablo's paradox and which has been overlooked by the participants of the discussion? As far as we can see, this is not the case. None of the more recent standard textbooks on the topic

(see, e.g., Barwise and Etchemendy (1987), McGee (1991), Gupta and Belnap (1993), Cantini (1996)) outlines any explication of self-referentiality or circularity, although the notions are of course applied on the intuitive level. While self-referentiality by itself has certainly not been the main target of these books — that has rather been the development of theories of truth for languages *allowing for* self-referentiality — it is still elucidating to see how some further clarification in this respect has been evaded. E.g., Gupta and Belnap (1993), p. 273, state in a footmark: 'We mean 'self-reference' in a general sense: A sentence A is self-referential if it is about itself or is about a sentence B that is about A or ... We shall not try to make the notion of "aboutness" precise.' This seems to conform to our account of circularity in section §2 and suffers from the same shortcomings. However, we do not want to give the impression that the exciting formal and philosophical machinery that has evolved after Kripke's (1975) seminal paper might not be put to use in order to answer our question from above; it is just not so clear how this is to be done. Let us consider, e.g., Kripke's well-known and important notion of *groundedness*: a sentence is grounded if it is assigned a classical truth value in the least fixed point of a 'jump' operator which is defined on the class of three-valued models of a language with truth predicate. Although ungroundedness and self-referentiality are certainly 'intertwined' in more than one respect, one has to resist the temptation to regard ungroundedness simply as the formal elaboration of the informal notions of self-referentiality or circularity: both the liar paradox and Yablo's paradox may easily be shown ungrounded, but it is doubtful whether this tells us anything about their status with respect to self-referentiality. After all, both paradoxes indeed seem to be ungrounded intuitively, though not both of them seem to be circular intuitively. Moreover, if, as Kripke suggests, the so-called Strong Kleene scheme is used as the three-valued background semantics, a sentence like $a_1 =$ '$\neg Tr(a_1) \vee a_1 = a_1$' turns out to be grounded (given equality is considered classical), while sentences like $a_2 =$ '$\neg Tr(a_2) \vee \neg a_2 = a_2$' or $a_3 =$ '$Tr(a_3) \vee \neg Tr(a_3)$' do not, although they should perhaps behave similarly as far as self-referentiality is concerned (or should they?). On the other hand, if the so-called Weak Kleene scheme is used as the three-valued semantics, many quantified sentences involving the truth predicate may be proven ungrounded though, intuitively, not being self-referential at all (consider, e.g., the sentence '$\forall x(P[x] \rightarrow Tr(x))$', where the extension of P may even be set *arbitrarily*). This leads us to another question: in how far shall our envisioned explication of self-referentiality reflect our commitments to a certain semantics or logic? Let us draw an analogy: there are axiomatic set theories in which the axiom of foundation is dropped and where instead some anti-foundation axiom is introduced by which the existence of certain non-well-founded sets is proved (see Aczel (1988)). E.g., it may be shown that there are sets X and Y, such that $X = \{X\}$, and

$Y = \{Y_1\}, Y_1 = \{Y_2\}, Y_2 = \{Y_3\}, \ldots$; intuitively, X is circular with respect to the membership relation whilst Y is not. However, according to Aczel's anti-foundation axiom, X is *identical* to Y, and thus either both are circular, or both are not, or the notion of circularity is to be abandoned. On the other hand, this is not necessarily the case if only some different set theory is chosen which still allows for non-well-founded sets but which replaces the axiom of foundation differently, such that X and Y do not turn out to be identical. In analogy, every formal elaboration of our intuitive conception of circularity might be bound to depend on the choice of a corresponding semantical background theory. Does every such theory have its 'own' formal concept of self-referentiality?

Since every answer to our open question from above seems to presuppose an explication of what it means for a sentence to be 'about' something, or of what it means for a sentence to be a fixed point 'essentially', or of what it means for a sentence to be ungrounded in a properly 'circular' way, and since these issues strongly depend on the semantical assumptions to be started from, we either suspect that much philosophical work lies ahead of us before the question is finally settled, or that otherwise the question is ill-posed, i.e., that the talk of self-referentiality is to be banished from scientific contexts.[2]

Dept. of Philosophy
University of Salzburg
Austria
E-mail: Hannes.Leitgeb@sbg.ac.at

REFERENCES

Aczel, P. 1988. *Non-Well-Founded Sets.* Stanford: CSLI.

Barwise, J., and Etchemendy, J. 1987. *The Liar. An Essay on Truth and Circularity.* Oxford: Oxford University Press.

Beall, J.C. 1999. 'Completing Sorenson's Menu: A Non-Modal Yabloesque Curry'. *Mind* 108, pp. 737–739.

Beall, J.C. 2001a. 'Is Yablo's paradox non-circular?'. *Analysis* 61, pp. 176–187.

Beall, J.C. 2001b. 'Heaps of Gluts and Hyde-ing the Sorites'. *Mind* 110, pp. 401–408.

Cantini, A. 1996. *Logical Frameworks for Truth and Abstraction.* Amsterdam: Elsevier.

[2] I would like to thank A. Hieke, L. Horsten, V. Halbach for their comments, corrections, and suggestions.

Gupta, A., and Belnap, N. 1993. *The Revision Theory of Truth.* Cambridge: The MIT Press.
Hempel, C.G. 1945. 'Studies in the logic of confirmation (I.)'. *Mind* 54, pp. 1–26.
Kripke, S. 1975. 'Outline of a Theory of Truth'. *Journal of Philosophy* 72, pp. 690–716.
Leitgeb, H. 2001. 'Theories of Truth which have no Standard Models'. *Studia Logica* 68, pp. 69–87.
McGee, V. 1991. *Truth, Vagueness, and Paradox.* Indianapolis: Hackett Publishing Company.
Priest, G. 1997. 'Yablo's paradox'. *Analysis* 57, pp. 236–42.
Quine, W.V. 1969. 'Ontological Relativity'. In his *Ontological Relativity and Other Essays.* New York: Columbia University Press, pp. 26–68.
Sorenson, R. 1998. "Yablo's paradox and kindred infinite liars". *Mind* 107, pp. 137–55.
Yablo, S. 1993. 'Paradox without self-reference'. *Analysis* 53, pp. 251–52.