

How to eliminate self-reference: a précis

Philippe Schlenker

Received: 6 February 2006 / Accepted: 16 May 2006
© Springer Science+Business Media B.V. 2006

Abstract We provide a systematic recipe for eliminating self-reference from a simple language in which semantic paradoxes (whether purely logical or empirical) can be expressed. We start from a non-quantificational language L which contains a truth predicate and sentence names, and we associate to each sentence F of L an infinite series of translations $h_0(F), h_1(F), \dots$, stated in a quantificational language L^* . Under certain conditions, we show that none of the translations is self-referential, but that any one of them perfectly mirrors the semantic behavior of the original. The result, which can be seen as a generalization of recent work by Yablo (1993, *Analysis*, 53, 251–252; 2004, *Self-reference*, CSLI) and Cook (2004, *Journal of Symbolic Logic*, 69(3), 767–774), shows that under certain conditions self-reference is not essential to any of the semantic phenomena that can be obtained in a simple language.

Keywords Paradox · Semantics · Self-reference · Yablo's paradox

1 Is self-reference semantically eliminable?

The Liar (*This sentence is not true*) is paradoxical by virtue of its self-reference. If Tr is the truth predicate, the Liar can be seen as a sentence $\neg\text{Tr}(s)$ named by a constant s (something we will represent as a pair $\langle s, \neg\text{Tr}(s) \rangle$, with the convention that the term s denotes the formula $\neg\text{Tr}(s)$). It used to be thought that self-reference¹ is *always* a crucial ingredient of semantic paradoxes. Yablo (1993, 2004) showed that this was not so; he constructed an infinite series of sentences none of which is self-referential but

¹ By *self-reference*, we mean direct or *indirect* self-reference (this broad notion is sometimes called 'circularity' in the literature). It has long been known that paradoxes can be produced with indirect self-reference—consider for instance a pair of sentences a and b , where a says: b is true, and where b says: a is false

which, taken together, yield a paradox. In its simplest form, Yablo's paradox consists of an infinite set of linearly ordered sentences, each of which claims that all the sentences following it are false. Using the notation we just introduced, the series can be represented as the set $\{(s(i), \forall k(k > i \rightarrow \neg \text{Tr}(s(k))) : i \geq 0)\}$, where for each integer i , $s(i)$ is intended to name the formula $\forall k(k > i \rightarrow \neg \text{Tr}(s(k)))$ [when the intended denotation is clear, we will often use $s(i)$ in the metalanguage to refer to the formula that the term $s(i)$ is supposed to denote, i.e. $\forall k(k > i \rightarrow \neg \text{Tr}(s(k)))$]. We may then reason as follows: If *all* sentences in the series are false, we obtain an obvious contradiction because what $s(0)$ asserts should be true. If *some* sentence, say $s(i)$, is true, it must be the case that for all $k > i$ (and hence in particular for all $k > i + 1$), $s(k)$ is false. But this should suffice to make $s(i + 1)$ true, which again yields a contradiction. Thus no bivalent valuation can be found for Yablo's series (Yablo, 2004 discusses other versions of his paradox, a point to which we return below).

There has been considerable debate to determine whether Yablo's result involves some 'concealed' self-reference (see Leitgeb, 2002 for a particularly clear discussion). In the present paper we will assume that it does not, and we will ask instead *how far Yablo's result can be generalized*. Thus we will ask whether self-reference ever plays a crucial semantic role, be it in the production of paradoxes or of other semantic phenomena. We will show that *to the extent that Yablo's sentences are not self-referential*, self-reference can be systematically eliminated from a simple language in which both logical and empirical paradoxes can be expressed (for brevity we will henceforth omit the condition *to the extent that Yablo's sentences are not self-referential*, which should be understood to prefix all of our claims). Cook (2004) already showed that Yablo's result can be generalized quite a bit. Specifically, he considered a bivalent logical system with infinite conjunction in which every sentence is of the form $\bigwedge_{i \in I} F(S_i)$, where $\{S_i : i \in I\}$ is a (possibly infinite) class of sentence names and where F is the falsity predicate. Both the Liar and Yablo's paradox can be formulated in this system, respectively as $\{(S_0, \bigwedge_{k \in \{0\}} F(S_k))\}$ and $\{(S_i, \bigwedge_{k > i} F(S_k)) : i \geq 1\}$ (where i ranges over the integers). The paradoxicality is apparent in the fact that no valuation can be found for these sentences if F is really interpreted as the falsity predicate. Interestingly, Cook defines an operation of 'unwinding' which transforms any set of formulas with an assignment of denotations to the sentence names into another such set which (i) does not involve any (direct or indirect) self-reference, but which (ii) shares important semantic properties with the 'original'. Cook's goal was to define the *simplest* framework in which Yablo's construction could be somewhat generalized. As a result, the syntax of his language is rather idiosyncratic, since *only* formulas of the form $\bigwedge_{i \in I} F(S_i)$ are deemed well-formed (in fact, the Truth-Teller cannot be straightforwardly defined in this system).

How do Yablo's and Cook's results bear on our understanding of truth? One of the important lessons of Kripke (1975) was that an adequate theory of truth should be 'risky', in the sense that a sentence may turn out to be Liar-like (or, for that matter, Truth-Teller-like) *depending on some empirical facts*. To take the simplest example, *It is raining and this sentence is false* (where *this* refers to the entire sentence) should turn out to be paradoxical just in case it is indeed raining; if it is not, the sentence should arguably be classified as false (this conclusion follows from the assumption, made in particular in the Strong Kleene trivalent logic, that a conjunction one of whose conjuncts is false is *ipso facto* false as well). Formally, this Empirical Liar can be analyzed as a pair $\langle e, R \wedge \neg \text{Tr}(e) \rangle$, where R is an atomic proposition (here: *It is raining*). Equally easily, we can define an Empirical Truth-Teller as the pair $\langle e', R \wedge \text{Tr}(e') \rangle$, approximating the English sentence *It is raining and this sentence is true*. The problem, of course,

is that Yablo's and Cook's constructions have nothing to say about these cases, since they only deal with purely logical paradoxes, not with empirical ones. Nor do they have anything to say about the *general* problem of determining whether any semantic phenomena (maybe not paradoxes) crucially depend on self-reference. In order to obtain a definite answer, a piecemeal approach is inadequate. It is not enough to show that Phenomena X , Y or Z (often paradoxes), which were thought to depend on self-reference, can be imitated without it; rather, we want to show that *every* semantic phenomenon that can be obtained in a specified language can be replicated without self-reference. We will show that this is indeed the case. Although we will restrict attention to a rather simple language (one with a truth predicate and sentence names but no quantifiers), we will show how to eliminate self-reference by *translating each sentence with a quantified formula that does not involve any self-reference*.

The rest of this paper is organized as follows. We outline the goals of the translation in Sect. 2, and show that the simplest procedure one could adopt happens to fail. A successful procedure is defined and illustrated in Sect. 3, and its main properties are discussed in Sect. 4, which is followed by some concluding remarks.

2 First steps towards a translation procedure

The language we will consider (call it L) contains a distinguished category of sentence names, which may only appear as arguments of the truth predicate Tr ; in turn, Tr is the only predicate of sentences. L contains some empirical vocabulary (*It is raining*, etc.) but no quantifiers. It will be expedient to divide the interpretation I' of L into three parts: (i) a classical interpretation I for the non-metalinguistic part of the vocabulary (i.e. everything except sentence names and Tr), (ii) a specification N' of the denotation of sentence names, and (iii) a specification of the extension and anti-extension of Tr . We can thus see I' as being an extension of the (classical) interpretation defined by the pair $\langle I, N' \rangle$. We assume throughout Kripke's theory of truth (Kripke, 1975), in the version obtained when formulas are evaluated according the Strong Kleene trivalent logic. For Tr to qualify as a truth predicate in an interpretation I' , I' should be a *fixed point*: the sentences that are true according to I' should be precisely those that fall in the extension of Tr , and similarly the sentences that are false according to I' should be precisely those that fall in the anti-extension of Tr . Given our stipulation that Tr may only take sentence-denoting names as arguments, the 'fixed point' condition is simply that for each sentence F , $I'(F) = \text{true}$ if and only if $F \in I'^+(\text{Tr})$ and $I'(F) = \text{false}$ if and only if $F \in I'^-(\text{Tr})$. Following Kripke's lead, we will *not* assume that any one fixed point is the 'right' one. Rather, we will require that a successful translation should imitate the behavior of the original with respect to *all* fixed points (in a sense that will be made precise shortly).

How do we intend to eliminate self-reference, then? Since we wish to translate self-reference away in *all* the sentences of the language, we should in particular eliminate it in the Liar. But there is little hope of doing so unless we resort to an infinite series of *quantificational* sentences (where we count as 'quantificational' sentences that involve infinite conjunction or disjunction). The reason is this: if a (finite or infinite) set of non-quantificational sentences is linearly ordered and if every sentence is required to refer only to sentences that come 'after' it, we can be sure that some bivalent valuation

exists for the set.² But of course no bivalent valuation can be found for the Liar, nor should one be available for its translation. Thus we probably have no choice but to translate the Liar with an infinite series of quantificational sentences, just as is the case in Cook’s ‘unwinding’ translation—or for that matter in Yablo’s paradox. Since we want our translation scheme to apply to *all* sentences, we will systematically associate to each sentence of L an infinite series of translations in a quantificational language L^* .

But if each sentence of L receives infinitely many translations in L^* , it is certainly reasonable to require that for any fixed point I^* of L^* , all the translations have the same value according to I^* —so that any one of them, or the equivalence class of them all, can be taken as ‘the’ translation of the original sentence (we henceforth call this requirement the Uniformity Condition).³ The simplest idea would be to generalize the construction that Yablo gave to obtain his paradox. The Liar $\langle s, \neg\text{Tr}(s) \rangle$ was, in effect, replaced with an infinite series $\{ \langle s(i), \forall k(k > i \rightarrow \neg\text{Tr}(s(k))) \rangle : i \geq 0 \}$. In any fixed point, the Liar has no truth value, since it is paradoxical; and it can be shown that the same applies to each of the sentences that partake in Yablo’s series. Generalizing somewhat, we could try to translate each formula F of the original language with an infinite series of formulas of the form $h_i(F) = \forall k(k > i \rightarrow [F]_k)$ (for $i \geq 0$), where $[F]_k$ is obtained from F by replacing each occurrence of the form $\text{Tr}(c)$ with $\text{Tr}(c(k))$. Of course we would still have to ensure that the initial denotation function N' for sentence names is replaced with a new function N^* , which guarantees that for each sentence name s that denotes a formula F according to N' , s is in L^* a *function symbol* and N^* guarantees that $s(i)$ denotes $h_i(F)$.

When we apply this procedure, the Liar is translated as Yablo’s paradox, all of whose members are undefined in any fixed point; in this case the translation is immediately successful. We achieve equal success with the Truth-Teller, which has the property that it can variably take the values *true*, *false* or *undefined* in different fixed points - though in Kripke’s *least fixed point* it takes the value *undefined*. Following our procedure, $\langle t, \text{Tr}(t) \rangle$ is translated as $\{ \langle t(i), \forall k(k > i \rightarrow \text{Tr}(t(k))) \rangle : i \geq 0 \}$, and it can be checked again that all these formulas must have the same truth value, but that it can be set arbitrarily to *true*, *false* or *undefined* (though sentences that ‘talk about’ the Truth-Teller will have *their* truth values determined by this initial choice). It can also be checked that in Kripke’s least fixed point all these formulas have the value *undefined*. This is certainly a property we would like to impose quite generally on a successful translation procedure: the behavior of F in the least fixed point of L should be identical to that of $h_i(F)$ in the least fixed point of L^* . So far, it would seem that

² Here is an argument, which was pointed out to me by Tony Martin. Consider a series of sentences of the form $\{ \langle c_k, f_k(c_{k+1}, \dots, c_{k+n_k}) \rangle : k \geq 0 \}$, where for each $k \geq 0$, f_k is a Boolean function. We show that for any such series there exists a bivalent valuation. Let us say that an assignment of truth-values to c_0, \dots, c_n is *acceptable* just in case for each $i \leq n$, (1) or (2) holds:

- (1) for some k such that c_k is an argument of $f_i, k > n$
- (2) (1) fails, and the truth value assigned to c_i is as required by the value of f_i .

For each n , there is an acceptable assignment of bivalent values to c_0, \dots, c_n . We can simply start with an arbitrary value for c_n and any other sentence which has at least an argument c_m for $m > n$. We then compute the values of the other sentences as the f_k dictate. Thus the binary tree of all acceptable assignments has arbitrarily long branches. By Koenig’s Lemma, it has an infinite branch, which is the desired valuation. (Note that as stated the argument only applies to series of the form $\{ \langle c_k, f_k(c_{k+1}, \dots, c_{k+n_k}) \rangle : k \geq 0 \}$; a slightly more general result would be desirable).

³ This property is similar to what Cook (2004) calls ‘recurrence’.

our translation method, which is a close relative of Cook’s ‘unwinding’ procedure,⁴ delivers the desired results.

Unfortunately, in more complex cases the method fails. In fact, we cannot even guarantee in the general case that all the translations of a given sentence share the same truth value (i.e. we cannot guarantee that the Uniformity Condition is satisfied). To construct a counter-example, we must examine (the translation of) a set of sentences that contains an ‘infinite reference path’ (i.e. a series of sentences s_1, s_2, \dots , where s_1 refers to s_2 , which refers to s_3 , which refers to s_4 , etc—without end, and without repetition).⁵ Specifically, let us consider in the original language the set $\{ \langle s_{i'}, \text{Tr}(s_{i'+1}) \rangle : i' \geq 0 \}$, which can be seen as an infinite series of sentences (arranged from left to right) which each say: *The sentence immediately to my right is true*. The translation of each pair $\langle s_{i'}, \text{Tr}(s_{i'+1}) \rangle$ in the original (left-to-right) series comes out as another infinite series $\{ \langle s_{i'}(i), \forall k (k > i \rightarrow \text{Tr}(s_{i'+1}(k))) \rangle : i \geq 0 \}$, which we can arrange from top to bottom. All the translations taken together will thus form an infinite table, in which each sentence says: *each sentence which is below me in the column immediately to my*

⁴ Cook’s method can be summarized as follows (we use a notation which is as close as possible to the one developed in our text; Cook’s notations are different).

- First, he defines a lexicographic order between pairs of indices: $\langle a, b \rangle < \langle c, d \rangle$ iff $a < c$ or $(a = c$ and $b < d)$
- Second, he stipulates that each pair $\langle S_i, \wedge_{k \in K} F(S_{i,k}) \rangle$ is ‘translated’ with the pairs $\langle S_{i,n}, \wedge_{k \in K, \langle n,i \rangle < \langle m,i_k \rangle} F(S_{i,k},m) \rangle, n \geq 0$. Cook then proves that any assignment I of truth values to formulas which is ‘acceptable’ (i.e. which ensures that F is indeed interpreted as the falsity predicate) gives a uniform truth value to all the translations.

⁵ If we prohibit infinite reference paths, the Uniformity Condition will be satisfied in any fixed point. Consider a sentence s , together with all the sentences that it refers to, together with all the sentences that *these* refer to, etc. We obtain in this way a finite set T with members t^1, \dots, t^n (T is finite because otherwise there would be an infinite reference path, contrary to our assumption). For each t^k , there is a Boolean function f_k such that in each fixed point, the value \mathbf{t}_k of t_k is $f_k(\mathbf{t})$, where \mathbf{t} is the n -tuple of values of t^1, \dots, t^n . Now call the translations of level i of these sentences t^1_i, \dots, t^n_i respectively. By the definition of the translation procedure, each t^k_i makes a claim about the series of the truth values of $\langle t^1_{i+1}, \dots, t^n_{i+1} \rangle, \langle t^1_{i+2}, \dots, t^n_{i+2} \rangle, \dots$. Let us write these tuples of truth values as $\mathbf{t}_{i+1}, \mathbf{t}_{i+2}$, etc. In any fixed point, the truth value of t^k_i is determined by the number of members of the series that make f_k true / false / indeterminate (specifically, t^k_i is true just in case each member of the series makes f_k true; t^k_i is false just in case some member of the series makes f_k false; and t^k_i is indeterminate otherwise). But there are only 3^n distinct n -tuples of trivalent values. So one of the tuples—call it \mathbf{t}^* —must recur infinitely many times in the series $\mathbf{t}_{i+1}, \mathbf{t}_{i+2}, \dots$. Let us assume that \mathbf{t}^* occurs in particular in position $i + k^* + 1$ (thus $\mathbf{t}_{i+k^*+1} = \mathbf{t}^*$). $t^k_{i+k^*+1}$, which makes a claim about the series $\mathbf{t}_{i+k^*+2}, \mathbf{t}_{i+k^*+3}, \dots$, has a certain value \mathbf{t}^{*k} , which by assumption is the k th coordinate of \mathbf{t}^* . Now we claim that $t^k_{i+k^*}$ must have the very same value \mathbf{t}^{*k} as $t^k_{i+k^*+1}$. This is because the truth value of $t^k_{i+k^*}$ is determined by the number of members of the series $\mathbf{t}_{i+k^*+1}, \mathbf{t}_{i+k^*+2}, \dots$, which make f_k true / false / indeterminate. But this series is identical to $\mathbf{t}_{i+k^*+2}, \mathbf{t}_{i+k^*+3}, \dots$, except that we have added one additional occurrence of \mathbf{t}^* (since $\mathbf{t}_{i+k^*+1} = \mathbf{t}^*$). But there were already infinitely many occurrences of \mathbf{t}^* in $\mathbf{t}_{i+k^*+2}, \mathbf{t}_{i+k^*+3}, \dots$, so we certainly haven’t changed the number of members of the series that make f_k true / false / indeterminate. Since this reasoning can be repeated for each k , we see that $\mathbf{t}_{i+k^*} = \mathbf{t}_{i+k^*+1} (= \mathbf{t}^*)$. By iterating the reasoning, we observe that whenever \mathbf{t}^* occurs in the series, it occurs in all the preceding positions as well. But since \mathbf{t}^* appears infinitely many times, this shows that \mathbf{t}^* must occupy the entire series. Therefore the Uniformity Condition is satisfied for all the sentences in T —and hence in particular for s . (Thanks to an anonymous referee for suggesting that this point be addressed. The method we use to prove that the Uniformity Condition is satisfied when infinite reference paths are prohibited is also applied in the proof of the ‘Uniformity Property’ of Schlenker (2006)—which in turn incorporates a suggestion due to D. Bonnay.)

right is true. The problem is that in the general case *there is no way to guarantee that all of these sentences have the same truth value.* In fact, it is easy to display a coherent valuation in which they don't (this valuation could be extended to a fixed point for the entire language by using methods that are laid out in the Appendix):

	$s_0(\cdot)$	$s_1(\cdot)$	$s_2(\cdot)$...
0	false	false	false	...
1	true	false	false	...
2	true	true	false	...
3	true	true	true	...
...

The table should be read as follows: each column lists the values of the translations $s_{i'}(0), s_{i'}(1), s_{i'}(2), \dots$ of a given sentence $s_{i'}$ of the initial language. For instance, the first column indicates that $s_0(0)$, which is the formula $\forall k(k > 0 \rightarrow \text{Tr}(s_1(k)))$, has the value *false*; that $s_0(1)$, which is the formula $\forall k(k > 1 \rightarrow \text{Tr}(s_1(k)))$, has the value *true*; that $s_0(2)$, which is the formula $\forall k(k > 2 \rightarrow \text{Tr}(s_1(k)))$, has the value *true*; and so on. Now each translation $s_{i'}$ claims that $s_{i'+1}(i'+1), s_{i'+1}(i'+2), \dots$ are all *true*; in other words, it talks about those sentences whose values appear immediately to the right and below the position of $s_{i'}$ itself. The valuation we have displayed is defined by $I(s_i(i)) = \text{false}$ if $i' \leq i$; and $I(s_i(i)) = \text{true}$ otherwise. It can be checked that it is indeed coherent, in the sense that it is compatible with a valuation for which Tr really is interpreted as the truth predicate. To see this, observe that $s_0(0)$, which claims that $s_1(1), s_1(2), \dots$ are all *true*, should indeed be *false* because $s_1(1)$ is *false*. By contrast, $s_0(1)$ claims that $s_1(2), s_1(3), \dots$ are all *true*, and it should be *true*, since *they* are all *true*. The reasoning can be extended to the rest of the table. We thus have a valuation (and hence a fixed point) in which the Uniformity Condition is violated, since for instance the various translations of s_0 fail to have the same value.⁶

⁶ If it were applied to a trivalent logic evaluated with the Strong Kleene Scheme, Cook's 'unwinding' procedure would suffer from the same problem. To see this, consider the following pairs from the initial language:

$\langle S_0, F(S_1) \rangle, \langle S_1, F(S_2) \rangle, \dots$, i.e. all the pairs of the form $\langle S_i, F(S_{i+1}) \rangle$ for $i \geq 0$. The translations obtained from Cook's procedure are the pairs $\langle S_{i,k}, \wedge_{m>k} F(S_{i,m}) \rangle, i, k \geq 0$. We now consider the following assignment of truth values, where # represents the value 'neither true nor false':

S_0, \cdot	S_1, \cdot	S_2, \cdot	S_3, \cdot	...
#	#	#	#	...
true	#	#	#	...
true	false	#	#	...
true	false	true	#	...
true	false	true	false	...
...
true	false	true	false	...
...

It can be checked that this assignment is coherent. For example, $S_{0,0}$, which is the formula $\wedge_{m>0} F(S_{1,m})$ should indeed have the value # because $S_{1,1}$ has the value # while for all $m > 1, S_{1,m}$ has the value *false*. By contrast, $S_{0,1}$, which is the formula $\wedge_{m>1} F(S_{1,m})$ should have the value *true*, as desired, because for all $m > 1, S_{1,1}$ has the value *false*.

3 The translation

The problem we just outlined has a solution, however. In fact, it is derived from a different version of Yablo’s paradox. Let us consider an infinite series of sentences each of which says: *Infinitely many sentences that follow me are false*. In our notation, this can be represented as the set $\{\langle s'(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge \neg \text{Tr}(s'(k')))) : i \geq 0 \rangle\}$. Each sentence $s'(i)$ says: as far as you go after rank i in the series $\neg \text{Tr}(s'(\cdot))$, you will find a false sentence. This is equivalent to saying that there are infinitely many false sentences after rank i . As it turns out, however, the modifier *after rank i* is truth-conditionally idle because the claim is utterly insensitive to whatever happens in any finite initial segment of the sequence. As a result, all the sentences make exactly the same claim, namely that there are infinitely many false sentences in the series. Since they have the same truth-conditional content, they must also receive the same value in any interpretation (note that this also applies to interpretations that are not fixed points: nothing in the argument hinges on the fact that Tr is interpreted as the truth predicate). Once we have this result, it is easy to show that the series is paradoxical. If each sentence is true, what each sentence says should in fact be false, which contradicts the assumption; and if each sentence is false, each sentence should be true—again a contradiction.⁷

This mechanism can be adapted to the general case. The idea is that for *any* truth function $F(\cdot)$, a series of sentences of the form $\{\forall k(k > i \rightarrow \exists k'(k' > k \wedge F(k')) : i \geq 0\}$ is guaranteed to have a uniform value in any interpretation, for the simple reason that *all the sentences in the series have exactly the same semantic content* (they all assert that the series of truth values $F(\cdot)$ has infinitely many true members). This guarantees that the Uniformity Condition will be automatically satisfied. Using this observation, we are at last in a position to offer a correct translation procedure. One version can be defined as follows:⁸

- (a) *Translation:* For each positive integer i , $h_i(F) = \forall k(k > i \rightarrow \exists k'(k' > k \wedge [F]_{k'}))$, where k and k' are ‘fresh’ variables, and where $[F]_{k'}$ is obtained from F by replacing each occurrence of the form $\text{Tr}(c)$ with $\text{Tr}(c(k'))$.
- (b) *Denotation:* s denotes F according to N' iff $s(i)$ denotes $h_i(F)$ according to N^* . (1)

Before we discuss the general properties of this translation scheme, let us see how it applies to some important examples. As before, we write $\langle s, F \rangle$ for a pair of a formula F denoted by a sentence-denoting term s , and we write the set of translations-cum-denotation relation as $h(\langle s, F \rangle) = \{\langle s(i), h_i(F) \rangle : i \geq 0\}$.

(1) First, let us make sure that the translation is adequate for sentences that do not contain the truth predicate, say *It is raining*, symbolized as R , and named by a constant r (we henceforth call a sentence *Tr-free* if it does not contain the truth predicate). Since R contains no occurrence of the truth predicate, the translation procedure yields a sentence with vacuous quantification, as follows:

⁷ Yablo (2004) discusses a version of the paradox in which every sentence in the series says: *All but a finite number of the sentences following me are false*. This can be represented by a kind of ‘dual’ of the version we have in the text: $\{\langle s'(i), \exists k(k > i \wedge \forall k'(k' > k \rightarrow \neg \text{Tr}(s'(k')))) : i \geq 0 \rangle\}$. Yablo’s example would illustrate just as well the points we make in the text.

⁸ Alternative translation procedures are defined and characterized in Schlenker (2006). We could just as well have chosen a translation modeled after the version of Yablo’s paradox discussed in the preceding footnote (this would yield the alternative definition $h_i(F) = \exists k(k > i \wedge \forall k'(k' > k \rightarrow [F]_{k'})$).

$$h(\langle r, R \rangle) = \{ \langle r(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge R)) \rangle : i \geq 0 \} \tag{2}$$

It is immediate that in any interpretation all the translations are equivalent to R , as is desired.

(2) Second, we should check that a sentence that talks about the truth of a Tr -free sentence is properly translated. Let us consider a sentence (named by a constant r') which says that r is true, yielding a pair $\langle r', \text{Tr}(r) \rangle$. Its translation is given by:

$$h(\langle r', \text{Tr}(r) \rangle) = \{ \langle r(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge \text{Tr}(r(k')))) \rangle : i \geq 0 \} \tag{3}$$

We have already established that all the sentences $r(i)$ (for $i \geq 0$) are equivalent to R . It follows that in any fixed point all the sentences $r'(i)$ are also equivalent to R , and hence to $\text{Tr}(r)$, as is desired.

(3) Third, let us consider the Liar. We have no new work to do, since we already discussed its translation when we introduced the modified version of Yablo's paradox. As is desired, the Liar $\langle s, \neg \text{Tr}(s) \rangle$ gets translated as a Yablo-like series which is itself paradoxical, namely $\{ \langle s(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge \neg \text{Tr}(s(k')))) \rangle : i \geq 0 \}$. Since this series has a uniform value, we immediately obtain the result that in any fixed point each sentence in the series should be neither true nor false.

(4) Fourth, we should consider the Truth-Teller $\langle t, \text{Tr}(t) \rangle$. It is translated as $\{ \langle t(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge \text{Tr}(t(k')))) \rangle$. As before, the form of the translations guarantees that in any interpretation they must all share the same value. It is then easy to see that there are fixed points in which these sentences are true, others in which they are false, and yet others in which they are undefined.

(5) Fifth, let us reconsider our empirical versions of the Liar and of the Truth-Teller, which we gave respectively as $\langle e, R \wedge \neg \text{Tr}(e) \rangle$ and $\langle e', R \wedge \text{Tr}(e') \rangle$. They are translated as $\{ \langle e(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge R \wedge \neg \text{Tr}(e(k')))) \rangle$ and as $\{ \langle e'(i), \forall k(k > i \rightarrow \exists k'(k' > k \wedge R \wedge \text{Tr}(e(k')))) \rangle$. Reasoning by cases, we see that if R is false we simply obtain two series of false sentences; and if R is true, we obtain an infinite Liar and infinite Truth-Teller, as we wished.

4 Properties

It is time to see that our construction yields the desired results in the general case. This is done in two steps: we first reiterate that all the translations of a given sentence display a uniform semantic behavior (Property 1), and we then show that the behavior in question adequately mirrors that of the original (Property 2).

Property 1 *In any interpretation of the target language L^* , for any sentence F of the original language L , I^* assigns the same truth value to all the translations of F .*

As was mentioned above, this result follows from the general form of the translation procedure: the translations of F all have the form $\forall k(k > i \rightarrow \exists k'(k' > k \wedge [F]_{k'}))$ and hence they all make the same claim about the Boolean series $[F]_{k'}(k' \geq 0)$, namely that it has infinitely many true members.⁹

⁹ More precisely: under the Strong Kleene Scheme, $\forall k(k > i \rightarrow \exists k'(k' > k \wedge [F]_{k'}))$ is:

- (a) true iff for each natural number $k > i$, for some natural number $k' > k$, the formula $[F]_{k'}$ is true of k' ;
- (b) false iff for some natural number $k > i$, for each natural number $k' > k$, the formula $[F]_{k'}$ is false of k' ;
- (c) undefined otherwise.

For the translation to be adequate, we would like a sentence and its translations to display the same semantic behavior with respect to all fixed points. But of course we are talking about fixed points of *different* languages, so we must establish simultaneously (i) a correspondence between the fixed points of the initial language and the fixed points of the target language, and (ii) a correspondence between the semantic behavior of the sentences of the initial language and their translations. Point (i) needs some elaboration; the target language contains many sentences that do not translate any sentences of the initial language. For our purposes it is natural to *treat as equivalent* fixed points of the target language that agree on all the formulas that translate some sentences of the initial language. We thus define an equivalence relation between the fixed points of the target language, and we show that there is indeed a natural isomorphism j between the fixed points of the initial language and the equivalence classes of fixed points of the target language. Once j is defined, Point (ii) is straightforward: we show that for any fixed point I' of the initial language and for any fixed point I^* of the target language, if I' and I^* are homologues according to j , then I' assigns the same value to a formula F as I^* does to its translations $h_i(F)$.

To make things precise, we define an equivalence relation \approx between the fixed points of L^* as: $I_1^* \approx I_2^*$ iff I_1^* and I_2^* agree on the translations of sentences of L (we call the set of translations $h(L)$). We write $[I_1^*]$ for the equivalence class of I_1^* . For any set of sentences S , we can further define a partial order on interpretations by stipulating that $i \leq_s j$ just in case every sentence of S that has a classical truth value in i has the same value in j . Property 2 will now guarantee that a sentence and its translations do indeed display the same semantic behavior.

Property 2 *There is an isomorphism j between the set of fixed points of L (compatible with $\langle I, N' \rangle$) ordered by \leq_L and the set of equivalence classes of fixed points of L^* (compatible with $\langle I, N^* \rangle$) ordered by $\leq_{h(L)}$ and j guarantees that for every sentence F of L , for every fixed point I of L , $I'(F) = j(I')(h_i(F))$. (The proof is sketched in the Appendix).*

5 Concluding remarks

Two remarks should be made by way of conclusion.

(1) Even if one grants (as we do) that the sentences that make up Yablo's 'Universal Liar' are not self-referential, it might seem that our procedure works only because all the translations $h_0(F), h_1(F), \dots$ of a given sentence F have in a strong sense the same semantic content. Specifically, we observed that in any interpretation that extends $\langle I, N^* \rangle$ (not just fixed points), the translations of F share the same truth value because they all make the same claim about the Boolean series $[F_{k'}]$ ($k' \geq 0$), namely that it has *infinitely many* true members. In this respect, it might seem that self-reference has been replaced with strong synonymy ('strong' because it holds not just in fixed points, but in all interpretations). Is this an essential feature of our translation? Yes

Footnote 9 continued

The condition in (a) boils down to: for infinitely many natural numbers k' , the formula $[F]_{k'}$ is true of k' .

The condition in (b) boils down to: for only finitely many natural numbers k' is the formula $[F]_{k'}$ true or undefined of k' .

Neither paraphrase of the conditions in (a) and (b) makes any reference to i , and hence no matter what the value of i is, the truth conditions of the formulas are the same.

and no. If one restricts attention to variants of the procedure that are obtained when *infinitely many* is replaced with other quantifiers, it can be shown that only those quantifiers that guarantee strong synonymy are adequate (Schlenker, 2006). On the other hand, it is easy to modify the form of the original translation by adding to each $h_i(F)$ a conjunct $(\text{Tr}(i = i) \leftrightarrow (i = i))$, which is true in fixed points but not necessarily in other interpretations (the equivalence has no reason to hold when Tr is not interpreted as the truth predicate). This little modification suffices to destroy strong synonymy, but it does not prevent the translation from working adequately, since we are only interested in what it does in fixed points.

(2) There are many respects in which our result is partial. One obvious limitation is that our initial language contains no quantifiers. We would like to be in a position to eliminate self-reference in quantificational languages as well; this would seem to be a straightforward extension, but it goes beyond the present note. A more essential limitation is that we have assumed that our initial language only contains one sentence-denoting predicate, the truth predicate. Could the construction be reproduced if other sentence-denoting predicates were included in the original language? In general, the answer probably has to be in the negative. It would be easy to introduce in the original language a semantic predicate whose intended meaning is: *is self-referential*. But it is clear that there will be no way to translate adequately this notion using our procedure. A sentence with name s which says that s is *self-referential* should come out as true in at least some fixed points of the initial language. But in no fixed point of the target language should any of the translations come out as self-referential.

Acknowledgements I wish to thank the following for helpful discussions: Denis Bonnay, Serge Bozon, Paul Egré, Marcus Kracht, Tony Martin, Benjamin Spector, Albert Visser, Ede Zimmermann, and audiences at IHPST, UCLA, U. of Amsterdam, U. of Frankfurt, ECAP'05, and *Sinn und Bedeutung* '05. Special thanks to Denis Bonnay and Albert Visser for extremely helpful corrections and suggestions, and to Paul Egré for comments on a first draft of the present paper. The author gratefully acknowledges the financial support of the American Council of Learned Societies (Ryskamp Fellowship) and of UCLA. A longer and more technical version of this analysis is offered in Schlenker (2006).

Appendix. Outline of a proof of Property 2

An interpretation point I' for the initial language L is fully determined by a specification of (i) the ground interpretation I , (ii) the denotation function N' for sentence-names, and (iii) the extension and anti-extension of Tr , i.e. $I'^+(\text{Tr})$ and $I'^-(\text{Tr})$. An interpretation for the target language L^* is determined by (i') the ground interpretation I , (ii'a) the denotation function N^* for functional names of sentences, (ii'b) an interpretation of the arithmetic vocabulary [which we assume to be in the standard model], and (iii') $I'^+(\text{Tr})$ and $I'^-(\text{Tr})$.

Definitions

- (i) I' is an *acceptable fixed point for L* just in case I' is a fixed point for L based on (i) the base interpretation I and (ii) the denotation function N' for sentence-names. I^* is an *acceptable fixed point for L^** just in case I^* is a fixed point for L^* based on (i') the base interpretation I , (ii'a) the denotation function N^* for functional sentence names, and (ii'b) the standard interpretation of the arithmetic vocabulary.

(ii) We write that $J(I', [I^*])$ just in case I' and I^* are acceptable fixed points for L and L^* respectively, and $I^{*+}(\text{Tr}) \cap h(L) = \{h_k(s) : k \geq 0 \text{ and } s \in I'^+(\text{Tr})\}, I^{*-}(\text{Tr}) \cap h(L) = \{h_k(s) : k \geq 0 \text{ and } s \in I'^-(\text{Tr})\}$.

(1) Let I' be an acceptable fixed point for L . We show that there is exactly one equivalence class of acceptable fixed points $[I^*]$ for L^* satisfying $J(I', [I^*])$.

- At most one: given N^* and I , the truth value of any member of $h(L)$ is fixed by the restriction of the interpretation of Tr to $h(L)$. As a result, once $I^{*+}(\text{Tr}) \cap h(L)$ and $I^{*-}(\text{Tr}) \cap h(L)$ are fixed, so is the value of each member of $h(L)$.
- At least one: we show how to construct an acceptable fixed point I^* for L^* which satisfies $J(I', [I^*])$. The construction is by stages: we define an increasing series $\langle E_i, A_i \rangle$ (for ordinal i) of extensions and anti-extensions for the truth predicate. Combined with I and N^* , each pair $\langle E_i, A_i \rangle$ defines an interpretation I_i^* for L^* , though I_i^* need not be a fixed point. But we show that for *some* ordinal i the desired interpretation is obtained.

f is the ‘jump’ operator, defined as: $f(\langle E_i, A_i \rangle) = \{\{s \in L^* : I_i^*(s) = 1\}, \{s \in L^* : I_i^*(s) = 0\}\}$. As observed in Kripke 1975, f is monotonic (increasing).

The definition of the series $\langle E_i, A_i \rangle$ is by induction: (1) $\langle E_0, A_0 \rangle := \{\{h_k(s) : k \geq 0 \text{ and } s \in I'^+(\text{Tr})\}, \{h_k(s) : k \geq 0 \text{ and } s \in I'^-(\text{Tr})\}\}$. (2) If i is a successor ordinal $k + 1$, $\langle E_i, A_i \rangle := f(\langle E_k, A_k \rangle)$. (3) If i is a limit ordinal, $\langle E_i, A_i \rangle := \cup_{k < i} \langle E_k, A_k \rangle$.

We note immediately that:

(*) (a) $\langle E_0, A_0 \rangle \subseteq \langle E_1, A_1 \rangle$ and (b) $\langle E_0, A_0 \rangle = \langle E_0 \cap h(L), A_0 \cap h(L) \rangle = \langle E_1 \cap h(L), A_1 \cap h(L) \rangle$

Proof Let s be any sentence of L . By Property 1, for all $k \geq 0, I_0^*(h_k(s)) = I_0^*(h_0(s))$. By the construction of $\langle E_0, A_0 \rangle$ (which treats in the same way the translations of different levels), $I_0^*(h_0(s)) = I_0^*(/h_0(s)/_{0/k'})$, where $/h_0(s)/_{0/k'}$ is obtained from $h_0(s)$ by replacing every formula $\text{Tr}(c(k'))$ with $\text{Tr}(c(0))$. Furthermore, $I_0^*(/h_0(s)/_{0/k'}) = I_0^*(\forall k(k > 0 \rightarrow \exists k'(k' > k \wedge [s]_0))) = I_0^*([s]_0)$ (because quantification is vacuous). But $I_0^*([s]_0) = I'(s)$ (again by the construction of $\langle E_0, A_0 \rangle$, which is derived from $I'^+(\text{Tr})$ and $I'^-(\text{Tr})$). In sum, $I_0^*(h_i(s)) = I'(s)$, and since I' is a fixed point, $I_0^*(h_i(s)) = 1$ (resp. $= 0$) iff $s \in I'^+(\text{Tr})$ (resp. $I'^-(\text{Tr})$), iff $h_i(s) \in E_0$ (resp. A_0). It follows that (a) $\langle E_0, A_0 \rangle \subseteq f(\langle E_0, A_0 \rangle)$, and hence $\langle E_0, A_0 \rangle \subseteq \langle E_1, A_1 \rangle$, and (b) $\langle E_0, A_0 \rangle = \langle E_0 \cap h(L), A_0 \cap h(L) \rangle = \langle E_1 \cap h(L), A_1 \cap h(L) \rangle$.

We now prove by induction that the property $\pi(i)$ holds of all ordinals i :

$\pi(i)$: for all i', i'' for which $i'' \leq i' \leq i$, (a) $\langle E_{i'}, A_{i'} \rangle \subseteq \langle E_{i'}, A_{i''} \rangle$ and (b) $\langle E_{i''} \cap h(L), A_{i''} \cap h(L) \rangle = \langle E_0, A_0 \rangle$.

- (i) $\pi(0)$ is trivially true.
- (ii) Suppose that i is a successor ordinal $k + 1$. Then $\langle E_i, A_i \rangle = f(\langle E_k, A_k \rangle)$. By the Induction Hypothesis, for each $k' \leq k, \langle E_{k'}, A_{k'} \rangle \subseteq \langle E_k, A_k \rangle$. By the monotonicity of f , it follows that for each $k' \leq k, f(\langle E_{k'}, A_{k'} \rangle) \subseteq f(\langle E_k, A_k \rangle)$, i.e. that $\langle E_{k'+1}, A_{k'+1} \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$. If $k = 0$, observation (a) in (*) states that $\langle E_k, A_k \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$. If k is a successor ordinal $k' + 1, k' \leq k$ and thus $\langle E_{k'+1}, A_{k'+1} \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$, hence $\langle E_k, A_k \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$. If k is a limit ordinal, $\langle E_k, A_k \rangle = \cup_{k' < k} \langle E_{k'+1}, A_{k'+1} \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$. In all cases $\langle E_k, A_k \rangle \subseteq \langle E_{k+1}, A_{k+1} \rangle$, which together with the Induction Hypothesis yields part (a) of $\pi(i)$.

To prove part (b), we observe that given N^* the value of any member of $h(L)$ is fixed by the restriction of the interpretation of Tr to $h(L)$. From the Induction Hypothesis it follows that $\langle E_k \cap h(L), A_k \cap h(L) \rangle = \langle E_0, A_0 \rangle$, and therefore $\langle E_{k+1} \cap h(L), A_{k+1} \cap h(L) \rangle = \langle E_1 \cap h(L), A_1 \cap h(L) \rangle$. But by part (b) of (*), $\langle E_1 \cap h(L), A_1 \cap h(L) \rangle = \langle E_0, A_0 \rangle$. With the induction Hypothesis, this proves part (b) of $\pi(i)$.

- (iii) Suppose that i is a limit ordinal. Then $\langle E_i, A_i \rangle = \cup_{k < i} \langle E_k, A_k \rangle$, which given the induction Hypothesis immediately yields $\pi(i)$.

The series $\langle E_i, A_i \rangle$ is increasing on the ordinals and thus it must have a fixed point $\langle E_{i^*}, A_{i^*} \rangle$, which determines the desired interpretation: $J(I', [I^*_{i^*}])$.

(2) Let I^* be an acceptable fixed point for L^* . We show that there is exactly one acceptable fixed point I' for L satisfying $J(I', [I^*])$.

Given Property 1, for all $k, k' \geq 0, I^*(h_k(s)) = I^*(h_{k'}(s))$. Given I and N' , we can thus define an interpretation I' by $I'^+(\text{Tr}) = \{s : \text{for some } k \geq 0, h_k(s) \in I^{*+}(\text{Tr})\}$ and $I'^-(\text{Tr}) = \{s : \text{for some } k \geq 0, h_k(s) \in I^{*-}(\text{Tr})\}$. It is then immediate that $I'^+(\text{Tr}) \cap h(L) = \{h_k(s) : k \geq 0 \text{ and } s \in I'^+(\text{Tr})\}$, $I'^-(\text{Tr}) \cap h(L) = \{h_k(s) : k \geq 0 \text{ and } s \in I'^-(\text{Tr})\}$. All that remains to be shown is that I' is a fixed point.

- (2a) From Property 1, it follows that for any formula F of $L, I^*(/h_i(F)/_{0/k'})$, where $/h_i(F)/_{0/k'}$ is obtained from $h_i(F)$ by replacing every formula $\text{Tr}(c(k'))$ with $\text{Tr}(c(0))$. Therefore for all $i \geq 0, I^*(h_i(F)) = I^*(\forall k(k > i \rightarrow \exists k'(k' > k \wedge [F]_{k'})) = I^*(\forall k(k > i \rightarrow \exists k'(k' > k \wedge /[F]_{k'}_{0/k'})) = I^*([F]_0)$ [because quantification is vacuous] $= I'(F)$ [because by construction, for any $c, I^*(\text{Tr}(c(0))) = I'(\text{Tr}(c))]$

(2b) We can now reason as follows:

$$\begin{aligned}
 F \in I'^+(\text{Tr})(\text{resp. } I'^-(\text{Tr})) &\text{ iff for each } i \geq 0, h_i(F) \in I^{*+}(\text{Tr})(\text{resp. } I'^-(\text{Tr})) \\
 &\text{ iff for each } i \geq 0, I^*(h_i(F)) = 1 \text{ (resp. } = 0) \\
 &\quad \text{[because } I^* \text{ is a fixed point]} \\
 &\text{ iff } I'(F) = 1 \text{ (resp. } = 0) \text{ [from (2a)].}
 \end{aligned}$$

Taken together (1) and (2) show that J is a 1-1, onto function from the acceptable fixed points of L to the equivalence classes of acceptable fixed points of L^* . We henceforth write $[I^*] = J(I')$ for $J(I', [I^*])$. It is immediate from the meaning of J that $I'_1 \leq I'_2$ iff $j(I'_1) \leq_{h(L)} j(I'_2)$.

References

Cook, R. (2004). Patterns of paradox. *Journal of Symbolic Logic*, 69(3), 767–774.
 Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72, 690–716.
 Leitgeb, H. (2002). What is a self-referential sentence? Critical remarks on the alleged (non-)circularity of Yablo’s paradox. *Logique et Analyse*, (177–178), 3–14.
 Schlenker, P. (2006). *The elimination of self-reference*. *Journal of Philosophical Logic* (to appear).
 Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53, 251–252.
 Yablo, S. (2004). Circularity and paradox. In: *Self-reference*, CSLI.