How to Eliminate Self-Reference: A Précis[1]

P. Schlenker (UCLA & Institut Jean-Nicod)

Second Draft (last modified on November 4, 2005)

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 $h_0(F)$, $h_1(F)$, ..., 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, 2004) and Cook (2004), shows that under certain conditions self-reference is not essential to *any* of the semantic phenomena that can be obtained in a simple language. [A longer and more technical version of the analysis is developed in 'The Elimination of Self-Reference'].

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 $\neg Tr(s)$ named by a constant *s* (something we will represent as a pair <s, ¬Tr(s)>, with the convention that the term *s* denotes the formula $\neg Tr(s)$). It used to be thought that self-reference 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 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**), ∀k (k>**i** → ¬Tr(s(k))>: i≥0}, where for each integer i, $s(\boldsymbol{i})$ is intended to name the formula $\forall k\ (k>\boldsymbol{i} \rightarrow \neg Tr(s(k))$ [when the intended denotation is clear, we will often use $\underline{s(\boldsymbol{i})}$ in the meta-language to refer to the formula that the term $s(\boldsymbol{i})$ is supposed to denote, i.e. $\forall k\ (k>\boldsymbol{i} \rightarrow \neg Tr(s(k))$]. We may then reason as follows: If *all* sentences in the series are false, we obtain an obvious contradiction because what $\underline{s(\mathbf{0})}$ asserts should be true. If *some* sentence, say $\underline{s(\mathbf{i})}$, is true, it must be the case that for all k>**i** (and hence in particular for all k>**i+1**), $\underline{s(k)}$ is false. But this should suffice to make $\underline{s(\mathbf{i+1})}$ true, which again yields a contradiction. Thus no bivalent valuation can be found for Yablo's series.

There has been considerable debate to determine whether Yablo's result involves some 'concealed' self-reference. In the present paper we will assume that it does not, and we will ask instead *how far Yablo's result can be generalized*. 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 $\wedge_{i\in I}F(S_i)$, 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,

[1]I thank the following for helpful discussions: Denis Bonnay, Serge Bozon, Paul Egré, Marcus Kracht, Tony Martin, Benjamin Spector, Albert Visser, as well as audiences at IHPST, UCLA, U. of Amsterdam, and ECAP'05. Special thanks to Denis Bonnay and Albert Visser for extremely helpful corrections and suggestions, and to Paul Egré, who provided 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).

respectively as $\{<S_0, \wedge_{k\in\{0\}} F(S_k)>\}$ and $\{<s_i, \wedge_{k>i}F(S_k)>: i\geq 1\}$ (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 $\wedge_{i\in I}F(S_i)$ 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 $<e, R \wedge \neg Tr(e)>$, where R is an atomic proposition (here: *It is raining*). Equally easily, we can define an Empirical Truth-Teller as the pair $<e', R \wedge Tr(e')>$, 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 given language can be replicated without it. We will show that this is indeed the case. Although we will we 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 Section 2, and show that the simplest procedure one could adopt happens to fail. A successful procedure is defined and illustrated in Section 3, and its main properties are discussed in Section 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 <I, N'>. 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)=true if and only if F∈I'$^{+}$(Tr) and I'(F)=false if and only if F∈I'$^{-}$(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 of 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 <s, ¬Tr(s)> was, in effect, replaced with an infinite series {<s(**i**), ∀k (k>**i** → ¬Tr(s(k))>: i≥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 $h_i(F) = \forall k\,(k>\boldsymbol{i} \rightarrow [F]_k)$ (for i≥0), where $[F]_k$ is obtained from *F* by replacing each occurrence of the form *Tr(c)* with *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 $h_i(F)$.

---

[2] Here is an argument, which was pointed out to me by Tony Martin. Consider a series of sentences of the form $\{<c_k, f_k(c_{k+1}, ..., c_{k+n_k})>: k\geq 0\}$, where for each k≥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 $c_0, ..., c_n$ is *acceptable* just in case for each i≤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 $c_0, ..., c_n$. 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 $\{<c_k, f_k(c_{k+1}, ..., c_{k+n_k})>: k\geq 0\}$; a slightly more general result would be desirable).

[3] This property is similar to what Cook 2004 calls 'recurrence'.

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, <t, Tr(t)> gets translated as {<t(**i**), ∀k (k>**i** → Tr(t(k))>: i≥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 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 that all the translations of a given sentence share the same truth value. Let us consider in the original language the set $\{<s_{i'}, Tr(s_{i'+1})>: i'\geq 0\}$, which can be seen as an infinite series of sentences that each say: *The sentence following me is true*. The translation of each pair $<s_{i'}, Tr(s_{i'+1})>$ comes out as an infinite series $\{<s_{i'}(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow Tr(s_{i'+1}(k))>: i \geq 0\}$. The problem is that *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):

	$\underline{s_0(.)}$	$\underline{s_1(.)}$	$\underline{s_2(.)}$	…
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 $\underline{s_{i'}(\mathbf{0})}$, $\underline{s_{i'}(\mathbf{1})}$, $\underline{s_{i'}(\mathbf{2})}$, etc. of a given sentence $\underline{s_{i'}}$ of the initial language. For instance, the first column indicates that $\underline{s_0(\mathbf{0})}$, which is the formula $\forall k\ (k>\mathbf{0} \rightarrow Tr(s_1(k))$, has the value *false*; that $\underline{s_0(\mathbf{1})}$, which is the formula $\forall k\ (k>\mathbf{1} \rightarrow Tr(s_1(k))$, has the value true; that $\underline{s_0(\mathbf{2})}$, which is the formula $\forall k\ (k>\mathbf{2} \rightarrow Tr(s_1(k))$, has the value true; and so on. Now each translation $\underline{s_i(\mathbf{i'})}$ claims that $\underline{s_{i+1}(\mathbf{i'+1})}$, $\underline{s_{i+1}(\mathbf{i'+2})}$, etc. are all true; in other words, it talks about those sentences whose values appear immediately to the right and below the position of $\underline{s_i(\mathbf{i'})}$ itself. The valuation we have displayed is defined by $I(s_{i'}(\mathbf{i}))$=false if i'≤i; and $I(s_{i'}(\mathbf{i}))$=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 $\underline{s_0(\mathbf{0})}$, which claims that $\underline{s_1(\mathbf{1})}$, $\underline{s_1(\mathbf{2})}$, ... are all true, should indeed be false because $\underline{s_1(\mathbf{1})}$ is false. By contrast, $\underline{s_0(\mathbf{1})}$ claims that $\underline{s_1(\mathbf{2})}$, $\underline{s_1(\mathbf{3})}$ ... are all true, and it should be true, since *they* are all true. The

[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: <a, b> < <c, d> iff a<c or (a=c and b<d)
-Second, he stipulates that each pair $<S_i, \wedge_{k\in K} F(S_{i_k})>$ is 'translated' with the pairs $<S_{i,n}, \wedge_{k\in K, <n,\, i> <<m,\, i_k>} F(S_{i_k,\, m})>$, n≥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.

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 $\underline{s_0}$ fail to have the same value[5].

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 $\{<s'(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k \wedge \neg Tr(s'(k'))>: i \geq 0\}$. Each sentence $\underline{s'(\mathbf{i})}$ says: as far as you go after rank i in the series $\neg Tr(s'(.))$, 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 semantically 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 semantic 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 would in fact be false, which contradicts the assumption; and if each sentence is false, each sentence should be true - again a contradiction[6].

This mechanism can be adapted to the general case. The idea is that for *any* truth function *F*, a series of sentences of the form $\{\forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k \wedge F(k'))$: $i \geq 0\}$ 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(.) has infinitely many true members). This guarantees that the Uniformity

[5] 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:
$<S_0, F(S_1)>$, $<S_1, F(S_2)>$, ... , i.e. all the pairs of the form $<S_i, F(S_{i+1})>$ for $i \geq 0$. The translations obtained from Cook's procedure are the pairs $<S_{i,k}, \wedge_{m>k}\ F(S_{i,m})>$, $i, k \geq 0$. We now consider the following assignment of truth values, where # represents the value 'neither true nor false':

$S_0, .$	$S_1, .$	$S_2, .$	$S_3, .$	…
#	#	#	#	
true	#	#	#	
true	false	#	#	
true	false	true	#	
true	false	true	false	
…	…	…	…	
true	false	true	false	
…	…	…	…	

It can be checked that this assignment is indeed 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*.

[6] 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: $\{<s'(\mathbf{i}), \exists k\ (k>\mathbf{i} \wedge \forall k'\ (k'>k \rightarrow \neg Tr(s'(k'))>: i \geq 0\}$. Yablo's example would illustrate just as well the points we make in the text.

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[7]:

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

Before we discuss the general properties of this translation scheme, let us see how it applies to some important examples. As before, we write <s, F> 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 $h(<s, F>)=\{<s(\mathbf{i}), h_i(F)>: i\geq 0\}$.

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:

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

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 <r', Tr(r)>. Its translation is given by:

(3) $h(<r', Tr(r)>)=\{<r'(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k\ \wedge Tr(r(k'))))>: i\geq 0\}$

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

3. Third, let us consider the Liar. We have no new work to do, since we already discuss its translation when we introduced the modified version of Yablo's paradox. As is desired, the Liar <s, ¬Tr(s)> gets translated as a Yablo-like series which is itself paradoxical, namely $\{<s(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k\ \wedge\neg Tr(s(k'))))>: i\geq 0\}$. 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 <t, Tr(t)>. It is translated as $\{<T(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k\ \wedge Tr(T(k'))))>$. 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 Truth-Teller, which we gave respectively as <e, R ∧ ¬Tr(e)> and <e', R ∧ Tr(e')>. They are translated as $\{<E(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k \wedge R \wedge\neg Tr(E(k'))))>$ and as $\{<E'(\mathbf{i}), \forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k \wedge R \wedge Tr(E'(k'))))>$. 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.

---

[7] Alternative translation procedures are defined and characterized in 'The Elimination of Self-Reference'. 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 $h_i(F)= \exists k\ (k>\mathbf{i} \wedge \forall k'\ (k'>k \rightarrow [F]_{k'}))$.

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 $\forall k\ (k>\mathbf{i} \rightarrow \exists k'\ (k'>k\ \wedge\ [F]_{k'}))$, 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[8].

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 <I, N'>) ordered by $\leq_{L'}$ and the set of equivalence classes of fixed points of L*

---

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

(compatible with <I, N*>) 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

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.

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'^{+}(Tr) and I'^{-}(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^{*+}(Tr) and I^{*-}(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 $I^{*+}(Tr)\cap h(L)=\{h_k(s): k\geq 0 \text{ and } s\in I'^{+}(Tr)\}$, $I^{*-}(Tr)\cap h(L)=\{h_k(s): k\geq 0 \text{ and } s\in I'^{-}(Tr)\}$.

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 $I^{*+}(Tr)\cap h(L)$ and $I^{*-}(Tr)\cap h(L)$ are fixed, so is the value of each of the members 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 $<E_i, A_i>$ (for ordinal i) of extensions and anti-extensions for the truth predicate. Combined with I and N*, each pair $<E_i, A_i>$ defines an interpretation I^*_i for L*, though I^*_i need not be a fixed point. But we show thtat for *some* ordinal i the desired interpretation is obtained.
f is the 'jump' operator, defined as: $f(<E_i, A_i>) = <\{s\in L^*: I^*_i(s)=1\}, \{s\in L^*: I^*_i(s)=0\}>$. As observed in Kripke 1975, f is monotonic (increasing).

The definition of the series $<E_i, A_i>$ is by induction: (1) $<E_0, A_0> := <\{h_k(s): k\geq 0 \text{ and } s\in I'^{+}(Tr)\}, \{h_k(s): k\geq 0 \text{ and } s\in I'^{-}(Tr)\}>$. (2) If i is a successor ordinal k+1, $<E_i, A_i> := f(<E_k, A_k>)$. (3) If i is a limit ordinal, $<E_i, A_i> := \cup_{k<i} <E_k, A_k>$.
We prove by induction that the property π(i) holds of all ordinals i:
π(i) : *for all i', i" for which* $i"\leq i'\leq i$, (a) $<E_{i"}, A_{i"}> \subseteq <E_{i'}, A_{i'}>$ *and (b)* $<E_{i"}\cap h(L), A_{i"}\cap h(L)> = <E_{i'}\cap h(L), A_{i'}\cap h(L)>$

(1) $\pi(0)$ is trivially true.
(2) 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 is a successor ordinal k'+1, $k' \leq k$ and $\langle E_k, A_k\rangle \subseteq \langle E_{k+1}, A_{k+1}\rangle$. If k is a limit ordinal, $\langle E_k, A_k\rangle = \cup_{k'<k}\langle E_{k'}, A_{k'}\rangle \subseteq \cup_{k'<k}\langle E_{k'+1}, A_{k'+1}\rangle \subseteq \langle E_{k+1}, A_{k+1}\rangle$. 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 for each $k' \leq k$, $\langle E_{k'} \cap h(L), A_{k'} \cap h(L)\rangle = \langle E_{k'} \cap h(L), A_{k'} \cap h(L)\rangle$, whence $\langle E_{k'+1} \cap h(L), A_{k'+1} \cap h(L)\rangle = \langle E_{k+1} \cap h(L), A_{k+1} \cap h(L)\rangle$. If k is a successor ordinal k'+1, $k' \leq k$ and $\langle E_{k'+1} \cap h(L), A_{k'+1} \cap h(L)\rangle = \langle E_{k+1} \cap h(L), A_{k+1} \cap h(L)\rangle$, i.e. $\langle E_k \cap h(L), A_k \cap h(L)\rangle = \langle E_{k+1} \cap h(L), A_{k+1} \cap h(L)\rangle$. If k is a limit ordinal, $\langle E_k \cap h(L), A_k \cap h(L)\rangle = \cup_{k'<k}(\langle E_{k'}, A_{k'}\rangle \cap \langle h(L), h(L)\rangle) = \langle E_0, A_0\rangle$ (again thanks to the Induction Hypothesis), and thus $\langle E_{k+1} \cap h(L), A_{k+1} \cap h(L)\rangle = \langle E_k \cap h(L), A_k \cap h(L)\rangle$.
(3) Suppose that i is a limit ordinal k. 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 $I'^{+}(Tr) = \{s$: for some $k \geq 0$, $h_k(s) \in I^{*+}\}$ and $I'^{-}(Tr) = \{s$: for some $k \geq 0$, $h_k(s) \in I^{*-}\}$. It is then immediate that $I^{*+}(Tr) \cap h(L) = \{h_k(s)$: $k \geq 0$ and $s \in I'^{+}(Tr)\}$, $I^{*-}(Tr) \cap h(L) = \{h_k(s)$: $k \geq 0$ and $s \in I'^{-}(Tr)\}$. 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', $I^*(h_i(F)) = I^*(/h_i(F)/_{\mathbf{0}/k'})$, where $/h_i(F)/_{\mathbf{0}/k'}$ is obtained from $h_i(F)$ by replacing every formula *Tr(c(k'))* with *Tr(c(**0**))*. Therefore for all $i \geq 0$, $I^*(h_i(F)) = I^*(\forall k\,(k > \mathbf{i} \rightarrow \exists k'\,(k' > k \wedge [F]_{k'}))) = I^*(\forall k\,(k > \mathbf{i} \rightarrow \exists k'\,(k' > k \wedge /[F]_k/_{\mathbf{0}/k'}))$
$= I^*([F]_{\mathbf{0}})$ [because quantification is vacuous]
$= I'(F)$ [because by construction for any c $I^*(Tr(c(\mathbf{0})) = I'(Tr(c))$]

2b. We can now reason as follows:

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

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
Schlenker, P. 2005. The Elimination of Self-Reference. Manuscript, UCLA & Institut Jean-Nicod.
Yablo, S. 1993. Paradox without self-reference. *Analysis* 53: 251–52.
Yablo, S. 2004. Circularity and Paradox, in *Self-Reference,* CSLI