

Draft of February 2019—please do not cite without permission.

Paradox with just self-reference¹

T. Parent (Virginia Tech)

parentt@vt.edu

1. Self-describing sentences

Yablo (1993) argues that paradox can be generated without self-referential terms. But he also denies that “self-reference *suffices* for paradox” (p. 251). This is obvious if he means just that not every sentence with self-reference is paradoxical (e.g., ‘This sentence is a sentence.’) Yet, since he justifies the claim by citing Tarski and Gödel, Yablo likely meant something stronger. A natural, alternate reading is: Self-reference does not suffice for any paradoxical sentence (in a semantically open language).² This is a standard view; however, in what follows it is argued to be false.

¹ I thank Bradley Armour-Garb, Axel Barceló, John G. Bennett, Matteo Bianchetti, Ethan E. Brauer, Tim Button, Dave Chalmers, Anil Gupta, Cory Johnson, Gary Kemp, Dan Linford, William Lycan, David McCarty, Michaelis Michael, Jay Newhard, Yves Peraire, Bryan Pickel, Martin Pleitz, Graham Priest, Panu Raatikainen, David Ripley, Keith Simmons, Peter Woodruff, and Min Xu for helpful discussion on various versions of this material. I also thank an audience at the 2018 Central Division meeting of the American Philosophical Association.

² Following Tarski (1944), a language L is “semantically open” iff: L does not contain its own semantic terms, i.e., L does not contain semantic terms that are defined on expressions of L.

As far as I know, Gödel and Tarski never say that self-reference (in a semantically open language) does not breed paradox. In fact, Gödel’s second incompleteness theorem implies that a formal system defined on the language *cannot* show itself to be paradox-free, unless the system is unsound. (There is also textual evidence that Tarski (1933/1983, pp. 159-162) believed that unconstrained metalinguistic reference, incl. self-reference, is indeed sufficient for paradox, though his reasons were different from those given here.)

Suppose we define the following, self-referring name:

(1) ‘**d**’ names ‘**d**’

Let us also assume the following disquotational principle:

(DQ) \ulcorner ‘ τ ’ names μ \urcorner entails $\ulcorner \tau = \mu \urcorner$.

So to illustrate, if ‘Hesperus’ names Phosphorus, then Hesperus = Phosphorus. I take such a principle to be familiar and fairly uncontroversial.

Similarly, (DQ) indicates that from (1) it follows:

(2) **d** = ‘**d**’

This should not seem too odd. (2) is true, since its two terms are co-referential. It thus says that **d** is self-identical, though it uses two different names to say that (viz., **d** itself and its quotation).³

Nevertheless, there is something strange afoot. One clue is that it is hard to say whether **d** satisfies the following open formula:

(#) x is the first term of this very substitution instance of (#).

If we assign ‘**d**’ to ‘ x ’, it seems indeterminate whether the formula is satisfied. Apparently, we must actually replace the variable by a term⁴ before there is a fact of the matter. For the truth-value differs depending on whether **d** or its quotation replaces the variable:

³ All this assumes that **d** is a linguistic *type* rather than a particular token or occurrence. If a self-referential term were literally *token-reflexive* (as is sometimes said), then the two tokens in (2) would refer to different items, and (2) would be akin to the falsity ‘this very token = this very token’. But in general, classical logic must individuate expressions by type. Otherwise formal rule $\ulcorner \forall x x=x \urcorner$ is unsound whenever a metalinguistic expression replaces ‘ x ’. I elaborate on this further in Parent (ms.).

⁴ As usual, terms of a language can be defined in non-semantic terms by a list (or in the case of quote names, by a recursive clause like: If τ is a term, then so is $\ulcorner \tau \urcorner$).

(3) **d** is the first term of this very substitution instance of (#).

(4) '**d**' is the first term of this very substitution instance of (#).

Clearly, (3) is true. But in (4), the first term is not **d** but rather the quotation of **d**. However, since (4) *claims* that **d** is first, (4) is false. Yet since **d** = '**d**', this can suggest that a single object both satisfies and fails to satisfy (#).

But in fact, 'this very substitution instance of (#)' has a different denotation in (3) versus (4). Because of that, the different truth-values between (3) and (4) do not really show that a single object satisfies and fails to satisfy a single formula. For in a regimented language, the shift in denotation would be indicated explicitly in the formalism (as with different readings of an ambiguous term, e.g., 'bank₁' versus 'bank₂'). Once this shift is made explicit, there will no longer be a *single* open formula which **d** satisfies and fails to satisfy.⁵

Still, we can capture the oddity in the area by a somewhat different means. But the root idea is the same. Namely, some formulae will yield a *self-describing expression* if a variable is replaced by **d**—but not if the variable is replaced by the quotation of **d**.

2. *The Lagadonian paradox*

The basic plan is to define a predicate 'L(x)' that **d** satisfies and '**d**' fails to satisfy—based on **d** satisfying and '**d**' failing a condition relevantly like (#). Though naturally, the new formula will also be unlike (#), in that it will not allow equivocation on 'this very substitution instance' or the like.

⁵ Paradoxes due to shifty deictic terms have been explored in detail by van Fraassen (1970) and Smullyan (1984).

Importantly, however, the new paradoxes are not just further cases of deictic paradox.

Consider then the following formula, defining the predicate ‘ x is Lagadonian.’⁶

(*) x is Lagadonian iff x is the first term in S , where S is the coordinated substitution instance of (*) in which y is the first term.

As a definition, each variable in (*) should be seen as universally quantified. But just to simplify matters, I will assume below that the quantifiers on ‘ x ’ and ‘ y ’ have been removed, and treat (*) as an open formula with ‘ x ’ and ‘ y ’ free.

In (*), let us define a “coordinated” substitution instance⁷ of a formula $\ulcorner \Phi(x, y) \urcorner$ (with exactly ‘ x ’ and ‘ y ’ free) as follows:

Def. S is a *coordinated substitution instance* [for short, a “CSI”] of a formula $\ulcorner \Phi(x, y) \urcorner$ iff there is a term τ such that $S = \ulcorner \Phi(\tau, \tau) \urcorner$.

Thus, in a CSI, the term replacing ‘ y ’ will be the quotation of the term replacing ‘ x ’. And note well that a CSI is defined exclusively in formal (hence, non-semantic) terms.

All tolled, the definition of ‘ x is Lagadonian’ works as follows. Take a CSI of (*), where *qua* CSI, the term replacing ‘ y ’ = the quotation of the term replacing ‘ x ’. The CSI states, for some specific x , a condition on which x is Lagadonian. In particular, it says that x is Lagadonian iff $x =$ the first term of that very CSI.

⁶ ‘Lagadonian’ is adapted from Lewis (1986), who borrows the term from Jonathan Swift. In Lewis, a “Lagadonian” language is one where an object is named by the object itself. Above, ‘Lagadonian’ is not so wide-ranging in its application conditions, though my choice of term is inspired by an obvious similarity with Lewis’ notion.

⁷ Just to be clear, ‘substitution instance’ is not a semantic term. It is defined here as follows: Given a formula $\ulcorner \Phi(x_1 \dots x_n) \urcorner$, a substitution instance of the formula is where some term τ has replaced at least one of the variables (and where any quantifiers binding the replaced variables have been deleted).

An example will help. Consider that one CSI of (*) is where ‘*x*’ and ‘*y*’ are replaced, respectively, with ‘‘**a**’’ and its quotation. Such a CSI then indicates a condition on which ‘**a**’ is Lagadonian:

- (5) ‘**a**’ is Lagadonian iff ‘**a**’ is the first term in *S*, where *S* is the coordinated substitution instance of (*) in which ‘‘**a**’’ is the first term.

The reader can verify that ‘**a**’ fails to be Lagadonian according to (5).

Of note, if we suppose that **b** = ‘**a**’, the following also gives us a condition on which ‘**a**’ is Lagadonian:

- (6) **b** is Lagadonian iff **b** is the first term in *S*, where *S* is the coordinated substitution instance of (*) in which ‘**b**’ is the first term.

There is no conflict between the two conditions on which ‘**a**’ is Lagadonian—provided that the term satisfies one condition iff it satisfies the other. That is indeed the case here, since ‘**a**’ is the first term of neither (5) nor (6). Hence, ‘**a**’ is not Lagadonian by either measure.

In fact for any *x*, all the relevant CSIs will agree that *x* is not Lagadonian—with one exception. This is when the first term used by a CSI of (*) is self-referring. (That is the only case where the first term is correctly named by the second term.) Thus, one condition on which **d** is Lagadonian is given by:

- (7) **d** is Lagadonian iff **d** is the first term in *S*, where *S* is the coordinated substitution instance of (*) in which ‘**d**’ is the first term.

Again, this CSI is distinctive since it consistently identifies its own first term. Accordingly, (7) indicates that **d** is indeed Lagadonian. And in light of (2), it then follows by the indiscernability of identicals that ‘**d**’ is Lagadonian.

But here arises the paradox. We can also show that ‘**d**’ is not Lagadonian. Consider that the following CSI also states a condition on which ‘**d**’ is Lagadonian:

(8) ‘**d**’ is Lagadonian iff ‘**d**’ is the first term in S , where S is the coordinated substitution instance of (*) in which “‘**d**’” is the first term.

As in (7), the embedded antecedent of (8) is true. But unlike (7), the embedded consequent is *false*. After all, the first term in (8) is not ‘**d**’ but rather the quotation of ‘**d**’. So (8) implies that ‘**d**’ is not Lagadonian. And this along with the previous argument shows that ‘**d**’ both is and is not Lagadonian. Hence, ‘‘**d**’ is Lagadonian’ is a sentence of a semantically open object language that violates the law of noncontradiction.

3. Interlude on intensionality

Thus, despite being defined in nonsemantic terms, the predicate ‘ x is Lagadonian’ is pathological. As the case of ‘**d**’ shows, the truth-value it yields depends on what is substituted for the variable. Some have said this shows that the Lagadonian-predicate is intensional, and that this deflates the significance of the paradox. At the least, however, ‘ x is Lagadonian’ would be a novel intensional context; the intensionality does not owe to a propositional attitude verb, an idiom like ‘so called’, or substitution into quotes.

But second—and much more importantly—the fact remains that we can generate a contradiction *in a semantically open language* simply by exploiting self-reference.⁸ (It is also

⁸ It might be complained that if ‘ x is Lagadonian’ is intensional, then the substitution of co-referring terms is invalid in relation to the predicate, and thus, the derivation of the paradox is unsound, since it depends on such a substitution. But for one, the Newhard version of the paradox (see below) does not so depend. Second, the substitution of co-referring terms is supposed to be sound in a classical language. So to levy this complaint is, really, just to admit that ‘ x is Lagadonian’ makes the language nonclassical.

notable that, unlike in the standard semantic paradoxes, the paradoxical sentences here do not use the negation operator.) Many have proposed ways to avoid these sorts of problems, of course. But one needs to *modify* the language in order to do so. In the present case, one might restrict self-reference in some way, such as forbidding a term like **d**. Yet the point would stand that, absent any such restrictions, there are wff in a semantically open language that are both true and false. Whether one wishes to call it a case of “intensionality” does not change the fact that such a language, as it stands, is nonclassical.⁹

And in fact, recent work by Jay Newhard (unpublished) has rendered the concern otiose. For Newhard demonstrates that no switch in co-referring terms is necessary to derive a contradiction. He first considers the following instance of (*), which he stresses is *not* a coordinated substitution instance of (*):

(8+) ‘**d**’ is Lagadonian iff ‘**d**’ is the first term in S , where S is the coordinated substitution instance of (*) in which ‘**d**’ is the first term.

Nonetheless, since the definition at (*) is fully general, (8+) should still be seen as laying down a condition on which ‘**d**’ is Lagadonian. In particular, it can be seen as saying that ‘**d**’ is Lagadonian iff ‘**d**’ is the first term of (7), where (7) is the CSI of (*) that begins with ‘**d**’. Observe, moreover, that (7) trivially satisfies that condition. Hence, (8+) implies that ‘**d**’ is

⁹ Conceivably, even the Liar can be seen as intensional, since deriving the contradiction may require the substitution of co-referring terms. See Tarski (1933/1983, p. 158), where the substitution is explicit. See also Tarski (1944, pp. 347-8), where he uses only the indiscernability of identicals, yet still effectively performs a substitution, as is the case above. (But n.b., some versions of the Liar may not require such a move, e.g., where we assume that there is a fixed point for ‘ x is not true’; see Gupta & Belnap 1994, ch. 3.) Regardless, it is incidental whether one wishes to call ‘**d** is Lagadonian’ a case of “intensionality;” it nonetheless violates principle of non-contradiction within a semantically open language.

Lagadonian. Yet it remains that (8) from the previous section shows, independently of all this, that ‘**d**’ is not Lagadonian. Contradiction. And here, no substitution of coreferential terms has been made.

4. *Further objections*

Some have objected, instead, that the paradox simply shows that the definition of ‘Lagadonian’ is not legitimate. But that is exactly the point: The definition of ‘Lagadonian’ leads to a sentence of a (semantically open) object-language which is both true and false.

Correlatively, such definition must indeed be excluded somehow. But—this is important—classical logicians have thus far not offered any stipulations against such a definition. As long as the language is semantically open, self-reference has been permitted without restriction. One example is with the Henkin (1949) construction in proving the completeness of predicate logic with identity, where each constant denotes itself in the relevant model. As a more humdrum example, suppose ‘ Wx ’ is defined in predicate logic to mean “ x is wff”: The definition might start by listing the singular terms and predicates—and the list of predicates may well include ‘ Wx ’ itself. This sort of thing has been done without any qualms whatsoever.

Yet this lack of restrictions is what permits the definition of ‘Lagadonian’, and what enables us to derive that **d** is and is not Lagadonian.¹⁰ So again, the point is precisely that the definition of ‘Lagadonian’ is illegitimate and thus, *novel restrictions on self-reference are required* for the language to remain classical.

¹⁰ Some have wondered whether it is legitimate for (*) to define x as “Lagadonian” by features of substitution-instances of (*). But that feature of the definition is enabled by unrestricted self-reference. So to place a ban on such a mechanism would already be to admit the key lesson, viz., that self-reference cannot be given free reign.

This is not to say that a classical language must be free of all self-reference. Nor does the paradox indicate, in particular, that such a language must refrain from identifying its own wffs. But it indeed reveals that a classical language must be cautious with self-reference, even though caution has largely been absent.

There is a variant on this objection, however, raised by a colleague (who prefers anonymity). The idea is that ‘Lagadonian’ is not problematic because of unrestricted self-reference, but because of a much more simple-minded error. The point is put as follows:

Imagine if I were to define a predicate ‘happy-sistered’ as follows. Where x is a sister of y , y is happy-sistered iff x is happy. Now, so long as there is a y that has two sisters, one happy and one not, we can show from my definition that such a y is both happy-sistered and not happy-sistered. Contradiction.

This is not a striking new paradox: it's just a bad definition. My definition says that whether someone is or is not happy-sistered is determined by any one of their sisters. But someone can have multiple sisters that differ in happiness, thus settling their happy-sisteredness in incompatible ways. The definition of ‘Lagadonian’ exhibits this same problem, but with names for an expression rather than sisters of a person.

However, there is a key disanalogy with the case of ‘happy-sistered’. The contradiction with ‘happy-sistered’ requires extra premises about the contents of the domain. For the contradiction is not derivable without assuming that the domain contains sisters of x that are neither uniformly happy, nor uniformly unhappy. In contrast, no such premises are necessary for Lagadonian paradox. The definitions at (1) and at (*) are sufficient to derive the contradiction.¹¹

¹¹ Granted, our derivation assumes that the domain contains substitution instances of (*). But that is secured by the unrestricted self-referential feature of (*) *ipso facto*.

Correlatively, it is not as if any language which includes the term ‘happy-sistered’ is nonclassical. It is only when such a language has an “uncooperative domain” that a contradiction results.¹² In contrast, any language that contains (1) and (*) is indeed nonclassical.

From a different angle, the objection is essentially reiterating that ‘Lagadonian’ is intensional. Whether ‘**d**’ has the property of being “Lagadonian” depends on whether ‘**d**’ bears a certain relation to itself or to its quotation, in the context of a sentence. Specifically, it is a matter of which term is being used to name ‘**d**’ in the sentence. (Compare: Whether Giorgione has the property of being “so called because of his size” depends on which name for Giorgione we are using in the context of the sentence; see Quine 1960, p. 153.) And yes, intensional predicates must be excluded from classical logic (just like sentences with ‘so called’ must be recast to avoid that idiom). So the upshot, I agree, is that the ‘Lagadonian’ predicate must be excluded from a classical language. But again, no constraints on self-reference have thus far been acknowledged.

5. *The Laputan paradox*

The Lagadonian paradox was arrived at, by reflecting on the peculiarity that **d** appears self-referential in a way that ‘**d**’ does not. (In the case of **d**, the term mentioned = the term used, but not in the case of ‘**d**’.) That is basically what the paradox exploits, when (7) consistently identifies its own first term. Unexpectedly, however, this kind of paradox does not depend on a self-referring name like **d**. (This was partly inspired by a point made by John G. Bennett (in correspondence) on a related issue.) A self-referring *definite description* will still be required—

¹² Even a language that defines the predicate ‘is a round square’ is classical, so long as nothing in the domain satisfies the predicate. (Whence, ‘There are no round squares’ is true only.) It is only with a round square in the domain that contradiction results, and then the blame really lies with the domain.

or more precisely, a definite description for the very sentence containing that description. Yet if a self-referring *name* is unnecessary to the paradox, then a classical language will require more restrictions than what was indicated above.

So in particular, consider the following definition of the predicate ‘*x* is Laputan’:

(†) *x* is Laputan iff ‘**a**’ is the first term in *S*, where *S* is the CSI of (†) in which *y* is the first term.

Qua definition, (†) should be seen as having all variables universally quantified, yet as with (*), I shall treat (†) as an open formula with ‘*x*’ and ‘*y*’ free (merely for convenience). Also, a “CSI” of the formula should be understood as per Def. above, namely, as a substitution instance of (†) where ‘*y*’ is replaced with the quotation of the term replacing ‘*x*’.

Suppose now that **a** = **b**, for an arbitrary **a** (regardless of whether **a** is a linguistic object or not). Then, the following CSIs will give conflicting verdicts on whether the object is Laputan:

(9) **a** is Laputan iff ‘**a**’ is the first term in *S*, where *S* is the CSI of (†) in which ‘**a**’ is the first term.

(10) **b** is Laputan iff ‘**a**’ is the first term in *S*, where *S* is the CSI of (†) in which ‘**b**’ is the first term.

Going by (9), **a** is Laputan since the first term in (9) is ‘**a**’. And since **a** = **b**, this implies that **b** is Laputan. Yet (10) does not have ‘**a**’ as its first term; hence, (10) rules that **b** is *not* Laputan. Hence, **b** is and is not Laputan. And again, no self-referring name like **d** is employed here, although (9) and (10) indeed feature definite descriptions which denote (respectively) the very sentences in which they occur.

Further, the reliance on the indiscernibility of identicals is unnecessary, as per Newhard’s maneuver from section 3. Consider here the following non-CSI:

(9+) **b** is Laputan iff ‘**a**’ is the first term in S , where S is the CSI of (†) in which ‘**a**’ is the first term.

As before, (9+) should be treated as definitional, even though it is not itself a CSI. More, ‘**a**’ is trivially the first term of the CSI of (†) where ‘**a**’ is first. Hence, (9+) indicates that **b** is Laputan. Yet it remains true that (10) shows, quite independently of all this, that **b** is not Laputan.

Contradiction.

By the way, (9+) itself is free of all self-reference. But the paradox here derives partly from (10), which indeed features a self-referential descriptor. (Also, (9+) makes a descriptive reference to (9), and (9) obviously indulges in much self-reference.) It is thus still fair to call the paradox from (9+) and (10) a paradox of self-reference.

6. Closing remarks

We began by noting how $\mathbf{d} = \mathbf{d}$ follows from (1), and then observing that the following is odd in relation to \mathbf{d} :

(#) x is the first term of this very substitution instance of (#).

Again, if ‘ x ’ is replaced by \mathbf{d} itself, the result seems true, but not if ‘ x ’ is replaced by the quotation of \mathbf{d} :

(3) \mathbf{d} is the first term of this very substitution instance of (#).

(4) ‘ \mathbf{d} ’ is the first term of this very substitution instance of (#).

This was not yet sufficient for a paradox, however, given the equivocation on ‘this very substitution instance of (#)’.

However, the suspicion was that the equivocal phrase could be eliminated by a metalinguistic descriptor, whose denotation could be determined algorithmically rather than

equivocally. More exactly, some descriptor of the form “ $\lambda y\Phi y$ ” might express a function that outputs (as its denotation) different sentences containing the descriptor, depending on what is inputted as y .

In (*), such a function was expressed by the descriptive phrase ‘the co-ordinated substitution-instance of (*) in which y is the first term.’ The phrase takes as input a term y , and outputs (as its denotation) the substitution-instance of (*) which has y as its first term. In particular, the descriptive phrase outputs (7) when $y = \mathbf{d}$, and outputs (8) when $y = \mathbf{“d”}$. The definition at (*) then determines whether x is Lagadonian by whether the first term of the relevant instance is identical to \mathbf{d} . The paradox results because when $x = \mathbf{d}$, the first term of the corresponding output is identical to \mathbf{d} only in the case of (7), not in the case of (8). And so, we get opposite verdicts on whether \mathbf{d} is Lagadonian.

I have redescribed matters in these function-theoretic terms, just to help convey how the definition of ‘Lagadonian’ is indeed a legitimate definition, formally speaking at least. But at the same time, it leads to a violation of the law of non-contradiction. The same can be said of the definition for ‘Laputan’, where a similar descriptive phrase expresses a function that outputs different instances of (\dagger), depending on whether $y = \mathbf{a}$ or $y = \mathbf{b}$. In both cases, we need some novel way of excluding such definitions, even though from a strictly formal point of view, the definitions seem entirely permissible.

In conversation, Tim Button has worried that if unconstrained self-reference suffices for paradox, then it seems Peano Arithmetic could be shown unsound via the method of Gödel numbering. For Gödel numbering enables something functionally like self-reference, and unrestricted self-reference now appears sufficient for paradox.

I admit that this objection is unsettling, and I hope to discuss it further in the near future. But nothing here yet demonstrates anything regarding arithmetic. As Ethan Brauer has emphasized (in correspondence), one would first need to show that something analogous to a “CSI” is arithmetically definable, and that is hardly obvious. And even then, the lesson may be just that Gödel numbering in the metalanguage must be restricted in some way, and not that the arithmetical object language is defective. At any rate, for now, my concern is only with classical *logic*: The paradoxes show that self-referential expressions must be restricted somehow in a classical system of symbolic logic, on pain of contradiction.

References

- Gupta, A. & Belnap, N. (1994). *The Revision Theory of Truth*. Cambridge, MA: MIT Press.
- Henkin, L. (1949). The completeness of the first-order functional calculus. *Journal of Symbolic Logic* 14: 159–66.
- Lewis, D. (1986). *On the Plurality of Worlds*, Malden, MA: Blackwell.
- Newhard, J. (unpublished). Comments on Parent’s ‘Paradox with just self-reference.’ Presented at the Central Division meeting of the American Philosophical Association, Feb. 2018.
- Parent, T. (ms.) The premise paradox. Available at <http://tparent.net/PremiseParadox.pdf>.
- Quine, W. (1960). *Word and Object*. Cambridge, MA: Harvard UP.
- Smullyan, R. (1984). Chameleonic languages. *Synthese* 60: 201–24.
- Tarski, A. (1933/1983). Pojęcie prawdy w językach nauk dedukcyjnych. Towarzystwo Naukowe, Warsaw. Reprinted as The concept of truth in formalized languages. In his *Logic, Semantics, and Metamathematics: Papers from 1923-1938*, 2nd edition, J. H. Woodger (ed. and trans.) and J. Corcoran (ed.), 152–269. Oxford: Oxford UP.

_____. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research* 4(3): 341–76.

van Fraassen, B. (1970). Inference and self-reference. *Synthese* 21: 425–38.

Yablo, S. (1993). Paradox without self-reference. *Analysis* 53(4): 251–2.