*Draft of March 2018—please do not cite without permission.*

# Paradox with just self-reference[1]

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 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, 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 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**' n**a**mes '**d**'

Let us also assume the following disquotational principle:

(DQ) ⌜'τ' names μ⌝ entails ⌜τ = μ⌝.

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).

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

(#) The last term of this very substitution instance of (#) is *x*.

If we assign '**d**' to '*x*', it seems indeterminate whether the formula is satisfied. Apparently, we must actually replace the variable by a term[3] before there is a fact of the matter. For the truth-value differs depending on whether **d** or its quotation replaces the variable:

[3] It will be important throughout that neither 'term' nor 'substitution instance' is defined using semantic notions. Terms of a language can be defined just by a list (or in the case of quote names, by a recursive clause like: If τ is a term, then so is ⌜'τ'⌝). Whereas, a substitution instance is defined here as follows: Given a formula Φ with $x_1 \ldots x_n$ free, a substitution instance of the formula is where at least one of the free variables has been replaced by a term. Some logicians define a 'substitution instance' just with respect to quantified sentences rather than open formulae, but it reduces clutter to leave quantifiers aside, and in the present discussion, no harm is done.

(3) The last term of this very substitution instance of (#) is **d**.

(4) The last term of this very substitution instance of (#) is '**d**'.

Clearly, (3) is true. But in (4), the last term is not **d** but rather the quotation of **d**. However, since (4) *claims* that **d** is last, (4) is false. Yet since **d** = '**d**', all 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., '$\text{bank}_1$' versus '$\text{bank}_2$'). Once this shift is made explicit, there will no longer be a *single* open formula which **d** satisfies and fails to satisfy.[4]

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—owing to **d** satisfying and '**d**' failing to satisfy 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.

[4] 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.'[5]

(*) If *S* is the coordinated substitution instance of (*) in which *x* is the last term, then *y* is Lagadonian iff the last term in *S* is *y*.

To act as a proper definition, all three variables in (*) should be universally quantified. But for simplicity's sake, I assume below that the quantifiers on '*x*' and '*y*' have been removed, and treat (*) as an open formula with '*x*' and '*y*' free.

Let us now stipulate that a sentence *S* is a "coordinated substitution instance" [a.k.a., a "CSI"] of an open formula ⌜Φ(*x*, *y*)⌝ (with exactly '*x*' and '*y*' free) iff there is a term τ such that *S* = ⌜Φ( 'τ', τ)⌝. Note that a CSI is defined exclusively by its formal (thus, non-semantic) features: A CSI of such a formula is where the term replacing '*x*' is the quotation of the term replacing '*y*'.

Roughly, the definition of 'Lagadonian' works as follows. Take a CSI of (*). The CSI defines, for some specific *y*, a condition on which *y* is Lagadonian. In particular, the CSI defines *y* as Lagadonian iff *y* = the last term of that very CSI.

An example will help. Consider that one CSI of (*) is where the quote-name ''**a**'' replaces '*y*' and the quotation of the quote-name replaces '*x*'. Such a CSI then defines a condition on which the term '**a**' is Lagadonian:

(5) If *S* is the coordinated substitution instance of (*) in which ''**a**'' is the last term, then '**a**' is Lagadonian iff the last term in *S* is '**a**'.

[5] '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.

Clearly, '**a**' fails to be Lagadonian according to (5). Contra the final clause, the last term of (5) itself is not '**a**' but rather the quotation of '**a**'.

Of note, if we suppose that **b** = '**a**', the following also defines a condition on which '**a**' is Lagadonian:

(6) If *S* is the coordinated substitution instance of (*) in which '**b**' is the last term, then **b** is Lagadonian iff the last term in *S* is **b**.

There need not be any conflict between two conditions on which a term is Lagadonian—as long as it satisfies one condition iff it satisfies the other. That is indeed the case here, since '**a**' is the last 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 last term used by a CSI of (*) is self-referring. (That is the only case where the last clause of the CSI correctly names its last term.) Thus, one condition on which **d** is Lagadonian is given by:

(7) If *S* is the coordinated substitution instance of (*) in which '**d**' is the last term, **d** is Lagadonian iff the last term in *S* is **d**.

Again, this CSI is distinctive since it consistently identifies its own last term. Accordingly, (7) determines 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 defines a condition on which '**d**' is Lagadonian:

(8) If *S* is the coordinated substitution instance of (*) in which ''**d**'' is the last term, then '**d**' is Lagadonian iff the last term in *S* is '**d**'.

As in (7), the antecedent is satisfied by the very sentence of which it is a part. But unlike (7), the final clause is *false*. The last 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. (It is also 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. [6]

And in fact, recent work by Jay Newhard (unpublished) renders the concern otiose. For Newhard demonstrates that no substitution move is necessary to the paradox. He first considers the following instance of (*), which he stresses is *not* a coordinated substitution instance of (*):

(8+) If *S* is the coordinated substitution instance of (*) in which '**d**' is the last term, then '**d**' is Lagadonian iff the last term in *S* is '**d**'.

Nonetheless, since the definition at (*) is fully general, (8+) should still be seen as laying down a condition on which '**d**' is Lagadonian. Thus, (8+) legislates that '**d**' is Lagadonian iff '**d**' is the last term *of (7),* given that (7) *is* the CSI of (*) that ends with '**d**'. Observe, moreover, the final clause of (8+) is satisfied by (7). Hence, (8+) implies that '**d**' is 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. 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,

[6] 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). 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. 2). Though perhaps a fixed-point version of the Lagadonian paradox exists as well. But again, it is incidental whether one wishes to call '**d** is Lagadonian' a case of "intensionality;" it violates principle of non-contradiction regardless within a semantically open language.

but not in the case of '**d**'.) That is basically what the paradox exploits, when (7) consistently identifies its own last 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 Johnathan G. Bennett (in correspondence) on a related issue.) A self-referring *definite description* will still be required—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 has been indicated so far.

In particular, consider the following definition of the predicate '*x* is Laputan':

(†) If *S* is the CSI of (†) in which *x* is the last term, then *y* is Laputan iff both '**a**' is the last term in *S* and **a** = *y*.

As with (*), I shall treat (†) as an open formula with '*x*' and '*y*' free (merely for convenience). And a "CSI" of (†) should be understood as per above, namely, as an instance of (†) where '*x*' is replaced with the quotation of the term replacing '*y*'.

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) If *S* is the CSI of (†) in which '**a**' is the last term, then **a** is Laputan iff both '**a**' is the last term in *S* and **a** = **a**.

(10) If *S* is the CSI of (†) in which '**b**' is the last term, then **b** is Laputan iff both '**a**' is the last term in *S* and **a** = **b**.

Going by (9), **a** is Laputan since the last term in (9) is '**a**'. Hence, since **a** = **b**, this implies that **b** is Laputan. Yet (10) does not have '**a**' as its last term; thus, (10) rules that **b** is *not* Laputan. So, **b** is and is not Laputan.

Again, no self-referring name like **d** was employed here, although (9) and (10) indeed feature definite descriptions which denote their own (respective) containing sentences. And here too, the reliance on the indiscernibility of identicals is unnecessary, as per Newhard's maneuver from section 3. Consider here the following non-CSI:

(10+) If *S* is the CSI of (†) in which '**a**' is the last term, then **b** is Laputan iff both '**a**' is the last term in *S* and **a** = **b**.

As before, (10+) should be treated as definitional, even though it is not a CSI. More, the conjunctive clause of (10+) is true, given that the sentence being referred to is (9). Hence, (10+) indicates that **b** is Laputan. Yet it remains that (10) shows, quite independently of all this, that **b** is not Laputan. Contradiction.[7]

*5. Closing remarks*

In conversation, Tim Button objects that if unconstrained self-reference suffices for paradox, then plausibly, Peano Arithmetic can be shown unsound via the method of Gödel numbering. For apparently, 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 points out (in conversation), one would first need to show that something analogous to a "CSI" is arithmetically definable, and that is hardly obvious. Regardless, the import for classical *logic*

[7] Granted, (10+) does not use any self-referential expression. But the paradox here also depends on (10), which indeed features a self-referential descriptor; so the Newhard variant is still fairly labeled a paradox of self-reference.

seems undeniable: The paradoxes show that self-referential expressions must be restricted somehow, on pain of contradiction.

## References

Gupta, A. & Belnap, N. (1994). *The Revision Theory of Truth.* Cambridge, MA: MIT Press.

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.

Smullyan, R. (1984). Chameleonic languages. *Synthese* 60: 201–224.

Tarski, A. (1933/1983). Pojęcie prawdy w językach nauk dedukcyjnych. Towarszystwo 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: 425–438.

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