CRÍTICA, Revista Hispanoamericana de Filosofía. Vol. 38, No. 113 (agosto 2006): 27–37

FREGE'S HIDDEN ASSUMPTION

HARTLEY SLATER
Philosophy Department
University of Western Australia
slaterbh@cyllene.uwa.edu.au

SUMMARY: This paper is concerned with locating the specific assumption that led Frege into Russell's Paradox. His understanding of reflexive pronouns was weak, for one thing, but also, by assimilating concepts to functions he was misled into thinking, one could invariably replace a two-place relation with a one-place property.

KEY WORDS: Russell's Paradox, Paradox of Predication, lambda terms, reflexive pronouns

RESUMEN: Este trabajo se ocupa de localizar el supuesto específico que llevó a Frege a la Paradoja de Russell. Por una parte, su comprensión de los pronombres reflexivos era débil pero, por otra parte, al asimilar los conceptos a las funciones pensó equivocadamente que uno siempre podía reemplazar una relación de dos lugares con una propiedad de un solo lugar.

PALABRAS CLAVE: paradoja de Russell, paradoja de la predicación, términos lambda, pronombres reflexivos

I

It is well known that it was Russell's Paradox that alerted Frege to the trouble with his system, for substitution of the set-abstract for 'x' in

$$x \in \{y | y \notin y\} \equiv x \notin x,$$

produces a contradiction. It is less well known that there is not the same trouble with

$$< x, x > \in \{< y, z > | y \notin z\} \equiv x \notin x.$$

For substitution of the set-abstract for 'x' here does not produce a contradiction (cf. Slater 1984). Substituting the set abstract for 'x' in the first case yields something of the form '$a \in a \equiv a \notin a$', but the comparable substitution in the second case merely produces something of the form '$< b, b > \in b \equiv b \notin b$'. What this suggests is that '$x \notin x$' cannot be analysed as involving simply a predicate of x rather than a relation between x and x, the larger moral being that not all relations between a thing and itself can be a matter of that thing falling under a concept, i.e.

$\neg(R)\ (\exists P)(x)(Rxx \equiv Px)$.

This is defended further in what follows, but it is what might have led Frege to think otherwise which is the main interest in the present paper.

Carnap's notes on Frege's lecture course on the *Begriffsschrift* show precisely where Frege went astray. Of course, other passages in Frege might also be quoted, but these lecture notes (recently published in English) are particularly clear in this regard. For there is an argument in them about concepts and relations, and specifically about the possibility of generating certain (one-place) concepts from certain two-place relations. One can put the problem highlighted by Russell's Paradox directly in terms of relations and concepts, to bring it closer to these kinds of expression. Remembering the general format of lambda abstraction reduction is $\lambda yFy[a] = Fa$, the trouble with the Paradox of Predication,

$$(\exists P)(x = \lambda tPt.\neg\lambda tPt[x]) \equiv \lambda yQy[x],$$

is that one seems to be required to equate the 'λyQy' on the right with '$\lambda y(\exists P)(y = \lambda tPt.\neg\lambda tPt[y])$', i.e. with the expression formed by abstracting the 'x' from the left hand side of the equivalence, to produce a concept of x. But this does not work. For there is then a contradiction when 'λyQy' replaces 'x'. One gets, because '$(\exists P)(\lambda yQy = \lambda yPy)$' is guaranteed, something of the form

$$\neg c[c] \equiv c[c].$$

On the other hand, there is no paradox if one abstracts from each 'x' separately, i.e. with

$$(\exists P)(x = \lambda tPy.\neg\lambda tPy\ [x]) \equiv \lambda y\lambda z(\exists P)(z = \lambda tPt.\neg\lambda tPt[y])[x][x].$$

In this case, substituting '$\lambda y\lambda z(\exists P)(z = \lambda tPt.\neg\lambda tPt[y])$' for '$x$' does not produce a contradiction. Hence the left hand side, $(\exists P)(x = \lambda tPt.\neg\lambda tPt[x])$, cannot be analysed as involving simply a concept of x rather than a relation between x and x. A paradox only arises when taking the two argument relational expression on the left (with the two arguments identified the same) to be equivalent to a single-subject with constant predicate expression, as when there was just '$\lambda y(\exists P)(y = \lambda tPt.\neg\lambda tPt[y])[x]$' on the right.

With respect to the Paradox of Predication, we therefore see that, while '*x* is a property which *y* does not possess' expresses a relation between *x* and *y*, that does not mean that, if '*y*' is replaced by '*x*', the result is a one-place property of *x*. And it is also very clear why this is so (see Slater 2004, 2005). For the predicate in the diagonal case, '*x* is a property which *x* does not possess' is itself something that varies with the subject specified. What is predicated of *a* would be that it is a property which *a* does not possess, but what is predicated of *b* would be that it is a property which *b* does not possess. In other words, the general predicate can be taken to be 'is a property which it does not possess', and this contains a pronoun, which is a contextual item with no direct representation in a context-free language. This pronominal predicate is functional, in other words, and the nearest one can get to any concept it expresses, in a language without pronouns, is:

$$\lambda y \lambda z(\exists P)(z = \lambda t Pt.\neg \lambda t Pt[y])[s],$$

with '*s*' not entirely free, but limited to repeating the subject of the sentence. Alternatively, by taking the second entry of '*x*', in the diagonal case, as the subject one could take the general predicate to be 'does not possess the property it is', and the nearest one could get to any concept this expresses, in a context independent language would be

$$\lambda z \lambda y(\exists \mathrm{P})(\mathrm{z} = \lambda t Pt.\neg \lambda t Pt[y])[s].$$

The attempt to construe these predicates as expressing a functional concept of their subject becomes needless, however, once the required subject term is actually attached, since the whole sentence is then revealed to be simply, though irreducibly relational, by each time being the analysis which was paradox free:

$$\lambda y \lambda z(\exists P)(z = \lambda t Pt.\neg \lambda t Pt[y])[x][x].$$

What Frege first missed with reflexive forms was the functionality of such pronouns. In the *Begriffsschrift* he says (Frege 1972, p. 127):

> The proposition that Cato killed Cato [can be considered in three ways, involving three different functions]. Here, if we think of 'Cato' as replaceable at the first occurrence, then 'killing Cato' is the function. If we think of 'Cato' as replaceable at the second occurrence, then 'being killed by Cato' is the function. Finally, if we think of 'Cato' as replaceable at both occurrences, then 'killing oneself' is the function.

But 'killing *s*', with '*s*' a pronoun, is not expressible in a context-free language, and, in addition, there are, in this case, other functions expressible in such a language which Frege does not mention: the two-place ones where the same name may (but need not) be put in both places: 'killing' and 'being killed by'. So why did Frege think that by abstracting 'Cato' from both occurrences in 'Cato killed Cato' one obtains a one-place function rather than a two-place one?

Here is the passage in Carnap's notes that provides the answer. Frege says (Frege 2004, p. 155):

> A function of two arguments, e.g. x–y, can be transformed into a function of one argument in two different ways, either by saturation (x–2), or by identifying the two argument places (x–x). Functions of two arguments that always have a truth-value as value are relations. Therefore we can transform the relation $x > y$ into a concept, e.g. $x > 0$ (the concept of a positive number). Or we can form the concept $x > x$.

If Frege could get the reflexive concept at the end, from the relation he started with, then a comparable derivation would produce a concept of x from the relation between x and x, on the left in the Paradox of Predication. So clearly this cannot be done.

II

How could Frege have missed the fact that only a reflexive relation, and not a reflexive concept is derivable? Clearly it was Frege's background in Mathematics that got him into trouble. Specifically, if there is no derivation of the supposed concepts from the given relations (both in the case of '$x > x$', and in the case of the Paradox of Predication), then Frege must have been working entirely on the basis of his understanding of mathematical functions —which is also nicely illustrated in the passage above. For there is no doubt that, given any function of two variables, $f(x, y)$, one can invariably obtain a function of one variable, by identifying the two arguments: $f(x, x) = g(x)$. One important case where this undoubtedly happens is in Cantor's diagonal procedure, for instance. But the seeming parallel case with relations and predicates, which generates Russell's Paradox, works very differently, as we have seen. So, while it is well known that Frege thought of concepts as functions, the analogy between the mathematical case, and the 'truth-value' case must limp just at this point. The point to note is that the 'truth-value' case involves an equivalence, not an identity, and we now know, from all logic

texts subsequent to Frege, that identity is not equivalence. Frege himself had a curious system, which allowed him to conflate, to some extent, identities and equivalences; but this has not been followed, and for good reason. For the expression for an identity, like '$a = b$', is between two names, whereas the expression for an equivalence, like '$p \equiv q$' is between two sentences. Thus 'if and only if' has quite a different grammar from 'is identical to'. Maybe by 'sentence' Frege meant 'nominalised sentence', since those certainly are referential expressions, and we can, as a result, say 'John's being a bachelor is the same as him being an unmarried male'. But since Frege did say 'sentence' we have every right to correct him. One cannot say, for instance, 'John is a bachelor is identical with John is an unmarried male'.

Sentences are the sort of expression that enters into equivalences, so they are not referring terms which can enter into identities (cf. Prior 1971, p. 35), and specifically, therefore, sentences are not referential terms with the same reference as 'The true' or 'The false', as Frege thought. If anything at all like '$Pa = T$', or '$Rab = F$' holds, it is with '=' as material equivalence, 'T' a tautology, and 'F' a contradiction. And then one has that '$Pa \equiv T$' and '$Rab \equiv F$' are equivalent to 'Pa' and '$\neg Rab$' respectively, making 'Pa' and 'Rab' quite unlike mathematical functions, and 'T' and 'F' nothing like their values.

On the specific question of a reflexive relation being a function of one argument, certainly one might be able to define a function $f(x)$ such that, say,

$$Rxx \equiv (f(x) = 1),\ \neg Rxx \equiv (f(x) = 0).$$

But not only does that not make the relation the function, also the right hand sides of these equivalences cannot be figured as involving the same predicate of x. For '$f(x) = 0$' is not contradictory, but merely contrary to '$f(x) = 1$'. If '$f(x) = 0$' was replaced by '$f(x) \neq 1$' there would be the same predicate of x; but no specific function would then be defined.

The propositional equivalences above though, namely '$Pa \equiv T$' and '$Rab \equiv F$', maybe still suggest that predicative expressions are functions of some sort. So we must delve deeper. The question in Carnap's case is whether from the relation $\lambda y \lambda x(x > y)$, one can obtain the concept $\lambda x(x > x)$, as well as the concept $\lambda x(x > 0)$. The second reduction is straightforward, since applying the two-term relation to 0 one gets the concept of being greater than 0:

$$\lambda y \lambda x(x > y)[0] = \lambda x(x > 0).$$

But the first reduction hits a problem. Proceeding as before one might try

$$\lambda y \lambda x(x > y)[x] = \lambda x(x > x),$$

but in '$\lambda y \lambda x(x > y)$' the '$x$' is a bound variable, and so the whole is equivalent to '$\lambda y \lambda z(z > y)$', and using that form one merely gets

$$\lambda y \lambda z(z > y)[x] = \lambda z(z > x),$$

i.e. the concept of being greater than x. Frege talks about getting his second, reflexive, concept by identifying the two variables, but he cannot be thinking that one can produce his second concept from $\lambda y \lambda z(z > y)[x][x]$, since while that produces the statement '$x > x$', it still does not identify the concept he mentions, because that statement is still analysed as a relation between two arguments, not as involving a single subject with a predicate expressing the concept $\lambda x(x > x)$. Certainly if we could form $\lambda x(\lambda y \lambda z(z > y)[x][x])$, we could get the desired $\lambda x(x > x)$, but the abstraction of x in that larger form is just as questionable, given that a relation between a thing and itself is not necessarily replaceable by a concept applicable to that thing.

It might be said that it is anachronistic to use closed lambda terms to try explicate what Frege was saying. The Paradox of Predication, for instance, is not obtainable in Frege's system, since he did not allow the application of a concept to another concept. So he most probably would have resisted the use of closed lambda terms in any explanation of what he meant by transforming the function $x > y$ into the function $x > x$. But if one cannot talk about Frege in a language he would not use, then one cannot criticise him on a scientific basis. One could not tell him, for instance, that sentences are not referring terms, as was done above, since 'for Frege' they are referring terms, and so one's remarks, it might be said, are not about what Frege was talking about, namely 'Frege sentences' which are referring terms, by definition. Popper, however, amongst several others, had a lot to say about this sort of thing in connection with closed societies, and pseudo-science.

Frege, in his article "Sense and Reference", wanted the 'F' and the 'a' to be both referring phrases in an elementary sentence such as 'Fa', taking the reference of the whole —a truth value— to be formed from the references of the parts. But only the singular term

is referential: both the predicate, and the sentence as a whole, are merely expressive (cf. Kneale and Kneale 1962, pp. 585–586; Prior 1971, p. 35; Wright 1983, p. 21; Slater 2000, *passim*). They are expressive of a concept and a proposition, respectively. A further argument Frege had for his 'truth-value' conclusion rested on what is commonly called his 'slingshot' (see Neale 1995, especially p. 765, and pp. 791–795). But the irony with this argument is that it is plainly invalid if complete individual terms are used for referential phrases (Neale 1995, pp. 795f), and Frege's extensional logic (unlike Russell's, for instance) did employ such complete referential terms. As it stands, that point provides merely an *ad hominem* argument against Frege, but it has been made very plausible that a better representation of referential phrases is obtained using certain other complete terms, namely Hilbert's epsilon terms (see Slater 2001, for instance), and so the inadequacy of Frege's 'slingshot' argument can be argued for much more generally.

The most crucial reason why sentences are not referring phrases, however, arises from the more basic fact mentioned above, that predicates are not referring phrases either. 'For Frege' they were, but Frege's thought at that point, of course, was what got him into his paradox about the concept *horse*. Following Cocchiarella (Cocchiarella 1986), we can remove that paradox. For we can distinguish, as before, the concept of being a horse (λxHx) from the predicate 'is a horse' ('$\lambda xHx[\ ]$'), and so see that it is the latter, and not the former, which is unsaturated. That is because 'being a horse' is a nominalised predicate, which hence is a referring phrase, while 'is a horse' is not nominalised, and contains a gap which needs to be filled before a complete thought can be expressed. If the predicate was a referring phrase, and referred to the concept, then that concept would certainly be unsaturated, but also the phrase 'the concept *horse*' would not refer to it, since that is saturated. Hence there would be Frege's paradox. But the predicate 'is a horse', while unsaturated, is not a referring phrase, and what does refer in the area is the nominalisation of a predicate, such as 'being a horse'. Indeed Frege himself, in his informal language, used nominalised predicates to refer to concepts, as when talking about 'being killed by Cato' and the like, in the first quotation above. But his 'official' position was that non-nominalised predicates had this purpose, so his theory was not in tune with his practise. Frege lacked a symbol for nominalised predicates in his formal language, which is what fundamentally led him to the conclusion that there is an inadequacy in natural language at this point, when it comes to expressing the semantical facts.

For there is no inadequacy in natural language when it comes to expressing the associated *natural language* semantic facts, and Cocchiarella has provided a symbolisation separating out predicates from their nominalisations, so natural language in this area can now be represented in a properly formal manner. Adopting Cocchiarella's symbolism we thereby move over to a clearer formal language without Frege's paradox, and with a clear distinction between predicates and their nominalisations, for a start.

But Cocchiarella's language naturally contains nominalisations of zero-place predicates, i.e. closed sentences, and misconceptions about such nominalisations also got Frege mixed up about truth (see Slater 2004, 2005). For it is not a sentence 'p', but its associated 'that'-clause ('λp', see Cocchiarella 1986, p. 217) which is referential, and the subject of judgements of truth and falsity. Thus it might be judged true or false that 5 is a prime number, for instance. But while 'that 5 is a prime number' therefore refers, it refers to a proposition, not a truth value, and so judgements of truth and falsity do not equate something with a truth-value, but instead predicate a truth-value of a proposition. That is to say, such judgements are not in the form of referential identities like '$\lambda Hd = T$' and '$\lambda Hd = F$', with 'T' and 'F' 'The true' and 'The false'. They are predicative remarks like '$T\lambda Hd$' and '$F\lambda Hd$', with 'T' and 'F' 'is true', and 'is false', where the lambda expressions obey the propositional truth schema: $T\lambda p \equiv p$. Certainly 'is a prime number' is then a function taking as values propositions which are true or false, but that means it is a *propositional* function, not a truth-value function like those in Frege's "Function and Concept" (see, e.g., Frege 1952, p. 28). As a result, the focus has to be on what sentences express.

So look at a reflexive case again. If A, B, and C each shave D then they do the same constant thing —shave D— but if they each shave themselves, or, in a ring, shave their neighbour on their left, say, then they do not do the same constant thing, since what they do merely has a common functional expression: shave $f(s)$, where the variable s is the subject. That means that a reflexive predicate is never, in itself, equivalent to a constant one-place predicate —although, contingently, of course, such a pair may be equivalent. They will be, for instance, if the number of objects involved is finite, since they then can be listed, and do not need to be described. Thus if all and only A, B, and C are self-shavers, then 'x is a self-shaver' is materially equivalent to 'x is one of A, B, and C'. But it is not *logically* equivalent to this disjunction, i.e. it does not say the same thing. For the variable in the predicate 'is a self-shaver', namely the

pronoun 'self', prevents the whole expressing a fixed property of its subject.

Of course one can put '*A* shaved *D*' differently. In line with the point made right at the start, one can say '*A* and *D* are a shaving (i.e. shaver-shaved) pair' in place of '*A* shaved *D*', and the predicate in that formulation is not functional even in the reflexive case: '*A* and *A* are a shaving pair'. But the latter does not predicate a fixed property of just *A*, since instead it predicates a fixed property of the ordered pair consisting of *A* and himself. Focussing on what they express, therefore, we see that a relational expression like '*Rxy*' invariably generates thoughts about the two objects *x* and *y*, and the reflexive, or diagonal expression '*Rxx*', as a result, generates thoughts about *x* and itself. Certainly the latter thoughts can be taken to be thoughts about the single subject *x*, as Frege saw with respect to 'Cato killed Cato'. But what Frege missed there was that the two ways in which this can be done each can be expressed with a pronoun, since each of those two thoughts about the subject involve a further function of it. Thus what is thought *about Cato* could be that he killed Cato, i.e. that he killed himself (more generally $Rxx \equiv \lambda yRyx[x]$ ($\equiv \lambda yRys[x]$)), and it could be that he was killed by Cato, i.e. that he was killed by himself (more generally $Rxx \equiv \lambda yRxy[x]$ ($\equiv \lambda yRsy[x]$)). Frege had no way to differentiate 'killing himself' from 'being killed by himself', but in both cases the pronoun is simply a context dependent replacement for the immediate subject, allowing the two expressions to be differentiated in a partly context-sensitive language as '$\lambda yRys$' and '$\lambda yRsy$'.

The tradition in modern logic has followed Frege in this respect, making a difference from the case with identities and equivalences. But the objective in all cases has been to give a representation of natural language structures and arguments, so these points about pronouns are in the same category as those about predicates and their nominalisations, and sentences and their nominalisations. It needs more than such a devise as a lambda term, however, to adequately formalise reflexive pronouns. For pronouns are context-sensitive elements, and so a context-sensitive formal language is required to symbolise them. As before, there is no way to represent reflexive pronouns as such in a completely context-free language. Certainly a pronoun *with its antecedent* can be represented in a context free way —thus '*a* is not a member of itself' is the same as '*a* is not a member of *a*'. But without that antecedent, the relevant context is unspecified, and the pronoun in the predicate 'is not a member

of itself' is revealed to be a limited variable, i.e. the referent of the pronoun is seen to be functional upon the subject supplied.

III

In conclusion, this paper has been concerned with Frege's understanding of reflexive expressions, and specifically about what led him to overlook the impossibility of it being generally the case that $(R)(\exists P)(x)(Rxx \equiv Px)$. His handling of pronouns was at fault, but also, when assimilating concepts to functions he was misled by a presumed affinity between identities and equivalences, and this had much larger consequences. Thus it was his move from 'x–y' to '$x > y$', in Carnap's passage above —and, of course, similar moves in other passages— that led him astray. Certainly 'x–y' is a mathematical function of x and y, but '$x > y$' is not. The former has two *arguments*, the latter two *subjects*, i.e. things the whole is saying something about. As a result the latter is a propositional, or logical function, of the form '$\lambda z \lambda t(z > t)[y][x]$', and identifying the appropriate variables in it merely turns this into another propositional function with two subjects, '$\lambda z \lambda t(z > t)[x][x]$', not a function of one variable '$\lambda z(z > z)[x]$' as with 'x–x'.

Returning to the original case of set-abstraction, we see that the predicate 'is not a member of itself' does not collect its subjects into a set, because there is a variable, 'itself', in this predicate, and so those things that are not members of themselves do not thereby have a common property —not even the common property of being members of the same set. Nevertheless, each one, paired with itself, is a member of a set of pairs. The Set Abstraction Axiom, in the case of elementary sentences, viz

$$(R^n)(\exists S)(x_1)(x_2) \ldots (x_n)(R^n x_1 x_2 \ldots x_n \equiv < x_1, x_2, \ldots x_n > \in S),$$

thus holds logically only if none of the variables are repeated in the relation. Of course, that still allows an equivalence to hold contingently in some cases, and even logically when some variables are repeated —in the latter case they simply must be repeated, as well, in the ordered set on the right. Clearly similar revisions are necessary not only with n-ary relations, but also with second and higher-order ones.

REFERENCES

Cocchiarella, N., 1986, *Logical Investigations of Predication Theory and the Problem of Universals*, Bibliopolis, Naples.

Frege, G., 2004, *Frege's Lectures on Logic*, trans. and ed. by Erich H. Reck and Steve Awodey, Open Court, Chicago.

———, 1972, *Conceptual Notation and Related Articles*, trans. and ed. by Terrell Ward Bynum, Clarendon, Oxford.

———, 1952, *Translations from the Philosophical Writings of Gottlob Frege*, trans. and ed. by Peter Geach and Max Black, Blackwell, Oxford.

Kneale, W. and M. Kneale, 1962, *The Development of Logic*, Clarendon Press, Oxford.

Neale, S., 1995, "The Philosophical Significance of Gödel's Slingshot", *Mind*, vol. 104, no. 416, pp. 761–825.

Prior, A.N., 1971, *Objects of Thought*, Clarendon Press, Oxford.

Slater, B.H., 2005, "Choice and Logic", *Journal of Philosophical Logic*, vol. 34, pp. 207–216.

———, 2004, "A Poor Concept Script", *Australasian Journal of Logic*: http://www.philosophy. unimelb.edu.au/ajl/2004/2004_4.pdf

———, 2001, "Epsilon Calculi", *The Internet Encyclopedia of Philosophy*: http://www.utm.edu/research/iep/e/ep-calc.htm

———, 2000, "Concept and Object in Frege", *Minerva —An Internet Journal of Philosophy*: http://www.ul.ie/∼philos/vol4/index.html

———, 1984, "Sensible Self-Containment", *Philosophical Quarterly*, vol. 34, pp. 163–164.

Wright, C., 1983, *Frege's Conception of Numbers as Objects*, Aberdeen University Press, Aberdeen.

Received: November 3, 2005; revised: July 10, 2006; accepted: September 5, 2006.