M. RANDALL HOLMES

TARSKI'S THEOREM AND *NFU*

Abstract. The Tarski paradox of the undefinability of truth is proved by a diagonalization argument similar to the argument of Russell's paradox. In *ZFC*, Russell's argument shows that the universal class (and large classes generally) do not exist. In other set theories, such as Jensen's variant *NFU* of Quine's "New Foundations", large classes such as the universe may exist; the diagonalization arguments lead to somewhat different restrictions on the existence of sets in the presence of different axioms. In this paper, we explore the possibility that semantics expressed in *NFU* may have somewhat different restrictions imposed on them by the diagonalization argument of Tarski. A language L is definable in *NFU*, in which the stratified sentences of the language of *NFU* can be encoded (but, it should be noted, as a proper subclass of L). Truth for sentences in L is definable in *NFU*, and the reason that a suitably adapted Tarski argument fails to lead to paradox is not that truth for L is undefinable in *NFU*, but that quotation becomes a type-raising operation, causing the predicate needed for the "Tarski sentence" to be unstratified.

1. INTRODUCTION

The well-known theorem of Tarski that truth of sentences in any reasonably expressive language L cannot be defined in the language L itself is proven by a diagonalization argument similar to the argument involved in Russell's paradox. The paradox of Russell shows us that some restriction on comprehension axioms in set theory is required, but it does not prescribe the restriction. The traditional approach involves "limitation of size", and is embodied in Zermelo set theory and its extensions. It is usual to think that Russell's paradox excludes "large" sets like the universe, but this is actually not the case. An alternate solution to Russell's paradox (and other paradoxes) was proposed by Quine (*1937*) in his system "New Foundations" (*NF*): comprehension restricted to stratified formulae. The consistency question for this theory remains open (as is generally known); what is less generally known is that the validity of the general approach of Quine to resolving the paradoxes has been demonstrated. Jensen (*1969*) showed that the theory *NFU* ("New Foundations" with the extensionality axiom weakened to allow urelements, but with the same comprehension axiom as "New Foundations" itself), is consistent relative to the usual set theory and remains consistent if the axioms of Infinity and Choice are adjoined. In these set theories, "large" sets like the universal set are provided by the comprehension axiom, and the resolution of the set-theoretical paradoxes proceeds along a different route. An all-purpose reference for set theories of this type, which can serve as a substi-

C. Anthony Anderson and M. Zeleny (eds.), Logic, Meaning and Computation, 469-478.

tute for most of the specific references cited in the paper, is Thomas Forster's excellent book *Set theory with a universal set* (*Forster 1992*), although the emphasis there is on *NF* rather than *NFU*.

Analogously, we have discovered that approaching the Tarski paradox of the definability of truth in a Quine-style set theory allows a different resolution of the problem; the paradox does not preclude the definability of truth any more than the Russell paradox precludes the existence of a universal set. We will present an infinitary language (with sentences of finite length but with infinitely many primitive predicates) which does define its own truth predicate, but avoids the paradox of Tarski because *quotation* turns out to be a "type-raising operation"! This appears to suggest an alternate approach to semantic paradoxes in general, analogous to the alternate approach to set-theoretical paradoxes embodied in *NFU*.

It should be noted that the infinitary nature of the language used is not accidental. Consideration of the paradox packaged succinctly in the phrase "the smallest natural number not describable in less than a billion words" reveals that any language which expresses its own semantic relations can be expected to have short names for each and every natural number, and so infinitely many atomic names (in a language without names, we can say, equivalently, "definite descriptions" in Russell's sense).

2. TARSKI'S THEOREM

We briefly (and informally) review the proof of the theorem of Tarski.[1] Suppose that we have encoded the formulae of a language L in such a way that they can be discussed in L (as numbers, for example). For each formula ϕ with one free variable x, let "ϕ" be the code for ϕ. We can think of formulas with one free variable as "definable predicates", holding of an object if substitution of a name for that object for the free occurrences of x in the formula yields a true sentence. Suppose, moreover, that truth of encoded sentences of L is a predicate definable in L.

We then consider the following, which should be expressible formally as a formula in one free variable x:

> The formula encoded by x, when each free occurrence of the variable "x" is replaced with the code x itself, yields the code of a sentence which is not true.

Less formally, for this to be true for an encoded formula "ϕ" in place of x means:

> The formula ϕ (as a predicate) does not hold of "ϕ".

[1] For details, see *Andrews 1986*.

Call the formula informally described above ψ; replacing the variable x in ψ with the code "ψ" yields a sentence which asserts its own falsehood! Roughly speaking, the resulting sentence says

> The predicate ψ does not hold of "ψ",
> but for ψ to hold of "ψ" means exactly:
>
> The predicate ψ does not hold of "ψ",
> so the sentence denies itself.

An obvious requirement for the argument to work is that the notion of substitution of a specific object for a variable in (the coded version of) a formula be definable in L; this holds (for the usual kind of coding using numbers) in any theory as strong as arithmetic. The resolution of the contradiction apparently must be (and indeed must be in the usual context) that the predicate "is true" of encoded sentences of L cannot actually be defined in L.

3. PRELIMINARIES IN *NFU*

NFU is a first-order, one-sorted theory with equality, membership, and a unary predicate of sethood.[2] The axioms of *NFU* are as follows:

Extensionality: Sets with the same elements are the same.

Urelements: Objects which are not sets have no elements.

Stratified Comprehension: For each variable x and formula ϕ which is "stratified", $\{x|\phi\}$, the set of all x such that ϕ, exists.

A formula is said to be "stratified" if types can be assigned to each variable occurring in the formula in such a way as to obtain a formula of the simple theory of types.

Observe that the defining formula $x \notin x$ of the Russell class is not stratified, but the defining formula $x = x$ of the universe *is* stratified, so there is a universe V.

We usually introduce the Axiom of Infinity by introducing the projection relations for an ordered pair which has the same relative type as its projections; this is equiconsistent with the Axiom of Infinity in a more usual form, and implies it, although it is inessentially stronger. Functions and relations can then be defined in the usual way, and the Axiom of Choice can be stated in the usual equivalent forms: our favorite form asserts that disjoint partitions of sets have choice sets. Another form worthy of notice is the assertion that V can be well-ordered.

[2]This form of presentation of *NFU* was suggested by Quine himself in his remarks accompanying *Jensen 1969*.

Note that the usual Kuratowski definition of the pair ($\langle x, y\rangle = \{\{x\}, \{x, y\}\}$), which can be used in *NFU* without the assumption of Infinity, yields a pair with relative type two higher than the types of its projections. This is inconvenient but not impossible to work with; it has odd effects on the relative types of functions and their arguments, for example.

Cardinal numbers (including natural numbers) are defined as equivalence classes of sets under the obvious equivalence relation; ordinal numbers are defined as equivalence classes of well-orderings under similarity. The objects which occasion the paradoxes of Cantor and Burali-Forti in naive set theory (the cardinality of the universe and the order type of the ordinals) actually exist in *NFU* but do not have quite the expected properties.

The cardinality of the universe is clearly not less than the cardinality of the power set of the universe (the set of all *sets*), so Cantor's theorem in its usual form cannot hold. The situation can be clarified by considering the form of Cantor's theorem which can be proven in the theory of types: there, we cannot even ask whether the cardinalities of a set A and its power set $\mathcal{P}\{A\}$ are the same, because the types of these two sets are different. What can be proven, in *NFU* as in the theory of types, is that the cardinality of $\mathcal{P}_1\{A\}$, the set of one-element subsets of A, is strictly less than the cardinality of $\mathcal{P}\{A\}$. In *NFU*, we can draw the further conclusion that $|\mathcal{P}_1\{V\}| < |\mathcal{P}\{V\}| < |V|$ (there are "fewer" singletons of objects than there are objects). This should not be too surprising, since the function which takes each object to its singleton has an unstratified definition.

The role of the set of one-element subsets of A in the argument above inspires the definition of a parallel operation on cardinals: $\mathrm{T}\{|A|\}$ is defined as $|\mathcal{P}_1\{A\}|$. It is straightforward to show that this operation on cardinals does not depend on the choice of A; it is unstratified and does not define a (set) function. For each cardinal $|A|$, the cardinal $\exp(|A|) = 2^{|A|}$ is defined (following Marcel Crabbé rather than Specker to obtain a slightly stronger definition) as $\mathrm{T}^{-1}\{|\mathcal{P}\{A\}|\}$; the function exp has a stratified definition, but it is partial (because T^{-1} is partial). Observe that exp thus defined is also the natural exponentiation function for the theory of types; the operations T and T^{-1} can be thought of in the context of the theory of types as projecting cardinals to "the same" cardinals in higher or lower types (from the standpoint of the usual set theory; we have seen that $\mathrm{T}\{|V|\} \neq |V|$ in *NFU*, so this cannot be our position from the *NFU* standpoint).

The Burali-Forti paradox does not afflict *NFU*, because it depends on the theorem of naive set theory or Zermelo-style set theory that the order type of the ordinals below α is equal to α itself. Observe that the order type of the ordinals below α is an object two types higher than α in the theory of types; one will then not be surprised to find that the corresponding theorem of *NFU* asserts that the order type of the ordinals below α in the natural order is $\mathrm{T}^2\{\alpha\}$, where $\mathrm{T}\{\alpha\}$ is defined as the order type obtained by replacing the objects ordered by any order of type α with their singletons. Now the

reasoning of the Burali-Forti paradox proves that $T^2\{\Omega\} < \Omega$, where Ω is the order type of the natural order on the ordinals; Ω proves to be greater than the order type of the segment below it and less than the largest ordinals. The sequence of iterated images of Ω under T can have no smallest element, but it is not a set. Such "sequences" (failing to be sets) of iterated images of objects under type-raising operations will play an important role below.

Note the role of type-raising operations such as the T operations on cardinal and ordinal numbers in avoiding paradox. The approach to avoiding semantic paradox here will allow the definition of notions such as truth and synonymy, at the price of treating *quotation* as a type-raising operation of this general kind and reference as an external relation with stratification restrictions similar to those on membership.

4. SEMANTICS IN *NFU*

The theory *NFU* + Infinity + Choice cannot describe its own semantics. This is fortunate, because the inconsistency of the theory would follow! However, it can describe the semantics of an infinitary language which captures the semantics of all *stratified* sentences of *NFU* in a certain guarded sense (if it did this unqualifiedly, it would still imply inconsistency of the theory!).

The obstruction to *NFU* introspecting on its own semantics is that the relation $\in$ is not a set; it should be clear that for any relation R which is a set, the set $\{x \mid \sim xRx\}$ is definable, and so the existence of the set $\in$ would imply the existence of the Russell class.

We use an idea of Grishin (*1972*) for a rather different purpose[3]. Observe that the relation of *inclusion* is a set; it has a stratified definition in which the two related objects are of the same type. Then observe that any formula $x \in y$ can be expressed as $\{x\} \subseteq y$, in which the non-relation $\in$ has been replaced by the relation $\subseteq$. One can then replace the expression $\{x\}$ by a variable X restricted to the set of singletons $\mathcal{P}_1\{V\}$.

In general, a formula in equality, membership, and the primitive projection relations can be translated into a formula in equality and the relations induced by inclusion and projection relations on n-fold singletons for each n in such a way that each variable of relative type i is replaced by a variable restricted to the set $\mathcal{P}_1^{N-i}\{V\}$ for a fixed natural number N large enough that $N - i$ is nonegative in each case needed. Note that any such formula can be expressed in the language with primitive unary predicates corresponding to each set in the universe and binary relations corresponding to each set relation in the universe. The specific unary and binary predicates we need are much more restricted, being simply the predicates of being an n-fold singleton for each n, equality, and the relations induced on n-fold singletons by inclusion and the projection relations; so it may seem to be overkill to use *all* sets and relations,

[3]See also *Forster 1992*, pages 64–66.

but there is no small set in *NFU* which contains all of the sets we need to encode these predicates (even though the class of sets we need is externally countable!).

If the sentence to be translated is not stratified, the translation process yields a formula in which some particular variable is replaced by variables representing different iterated singletons of objects represented by that variable; since the relation between the mth and nth iterated singletons of the same object for $m \neq n$ is not captured by any set relation in *NFU*, the translated sentence will not be usable in the construction below.

We now introduce an infinitary language L, having a primitive predicate "membership in x" for each object x in the universe and a primitive binary predicate R for each set of pairs such that xRy is to hold exactly when $(x, y) \in R$. The logical operations allowed in L are the usual propositional connectives and quantifiers (L is infinitary only in having infinitely many constants; sentences of L are finite in length). The terms of L are variables indexed by the natural numbers.

As we indicated above, each sentence of *NFU* can be translated into a sentence of the language L. There are *countable* sublanguages L_i of L capable of expressing all sentences of *NFU* using no relative types other than 0-$(i-1)$, but the "union" of the L_i's is not a set, and there is no set of exactly the sentences of L which encode sentences of *NFU*. If there were such a set, paradox would ensue.

We develop the semantics of L in some detail, and indicate how truth of coded sentences of L is definable in *NFU*. Use of the type-level pair of the previous section will be essential to allow induction on structures built up using pairing.

Moreover, we will make an assumption strengthening our set theory as well. Observe that $\mathrm{T}\{n\} = n$ holds for $n = 0, 1, 2, \ldots$, where T is the type-raising operation on cardinal numbers of the previous section. It is impossible to prove the assertion "$\mathrm{T}\{n\} = n$ for each natural number n" in *NFU* + Infinity + Choice; the attempt to prove it by induction fails because the condition on which induction is to be carried out is unstratified and fails to define a set. Nonetheless, this assertion, called Rosser's Axiom of Counting,[4] is consistent with *NFU* + Infinity + Choice (and strengthens it essentially).

Atomic terms of L are variables. There will be countably many variables v_i for i ranging through the non-negative integers. The fact that each singleton corresponds to a predicate means that we could just as well have a name for each constant in the universe, but we will refrain from this extravagance.

Atomic formulas of L are statements of membership in sets and participation in binary relations. The sentence "$v_i \in A$" is encoded as $((i, \{A\}), 0)$, while the sentence "$(v_i, v_j) \in R$" is encoded as $(((i, j), \{R\}), 1)$. The use of the singleton operation is to preserve stratification; the relative type of a

[4]Proposed for *NF* in *Rosser 1953*.

coded sentence is to be the same as the intended relative type of its variables, while the sets represented by predicates, unary or binary, are one type higher.

Similarly, we define codes for formulas inductively:

$$\text{“}\sim\phi\text{”} = (\text{“}\phi\text{”}, 2);\quad \text{“}\phi \wedge \psi\text{”} = \big((\text{“}\phi\text{”}, \text{“}\psi\text{”}), 3\big);\quad \text{“}(\forall v_i)(\phi)\text{”} = \big((i, \text{“}\phi\text{”}), 4\big).$$

It is straightforward to define the predicate "x is a code for a formula" and the operation of substitution of a term for a variable by induction; *NFU* provides more than enough set theoretical machinery for this, but it is crucial that the type-level pair is used to allow structural induction on pairs.

Notice that a type raising operation T on coded sentences can be defined: replace each set x which appears representing a unary relation with $\mathcal{P}_1\{x\}$ and each relation R with the corresponding relation on singletons. Our assumption of the Axiom of Counting ensures good behavior of this type-raising operation; in particular, it implies that coded sentences which do not involve any constants will actually be sent to themselves, as one would expect! $\mathrm{T}\{\text{“}\phi\text{”}\}$ can be obtained from the singleton of "ϕ" by an inductively defined set function. If this operation is applied to a (doubly) encoded sentence of *NFU*, it actually has the effect of systematically *lowering* types by one.

One proceeds to define "satisfaction" as a predicate of pairs $(\{\text{“}\phi\text{”}\}, f)$, where f is a function taking each natural number i to an intended value for v_i, in the usual way; the singleton operation is used to preserve stratification. The satisfaction predicate, which is stratified and does define a set, can then be used to define the set of true closed sentences of L in the obvious way. Notice that the closed sentences of each L_i can actually be encoded as numbers; there are only countably many sentences in each of these languages.

5. THE TARSKI "PARADOX" IN THE LANGUAGE L

The "paradoxical" situation which now arises is this. We have defined the notions of "sentence of L" and truth of sentences of L via stratified constructions in *NFU*. But we know that L itself can express any stratified sentence of *NFU*. Thus, it appears that L may be able to capture its own truth predicate.

To see what actually happens, it is useful to look at the world of *NFU* as it is seen from the standpoint of L. Sentences of *NFU* are effectively translated into sentences of a type theory (not internally describable in L). The types are the iterated images of V under $\mathcal{P}_1$; the membership relations are inclusion (of singletons in general sets, between the top type and the type below it) and the relations induced by inclusion on iterated singletons (between successive lower types). Notice that this is a "downward" type theory in which there is a top type and no bottom type; in all other respects it is precisely analogous to the usual kind of type theory. A consequence of our assumption of the Axiom of Counting is that the type-raising operation T on "pure" translated

sentences of *NFU* (involving no constants other than those used to represent types) preserves truth value.[5]

Our construction of the satisfaction predicate for L involved the relative type of coded sentences of L and two types above that. In the translations of sentences involving satisfaction of the sentence ϕ into L, the double singleton of the code "ϕ" will appear in place of the code itself. Moreover, there is a further reflection into lower types involved if we consider the translation of a sentence *about* satisfaction into L; this will mention the relation of satisfaction using its double singleton and any formulas to be satisfied via their *quadruple* singletons.

Suppose we try to replicate the argument of Tarski. We would use the predicate:

The predicate represented by formula ϕ is not true of "ϕ".

But this predicate is either unstratified or senseless (depending on how it is read). Certainly the reference to the formula ϕ outside of quotes in the predicate is, for L, a reference to the double singleton of ϕ. Then it is necessary to realize that the truth predicate appearing in a sentence interpreted inside of L now refers to a property not of double singletons of formulas but of double singletons of double singletons of formulas; the question is then whether the term "ϕ" is to be understood as referring to the double singleton of ϕ, in which case it is not understood as a code for a formula (so we have the "senseless" interpretation) or as a quadruple singleton, in which case the formula is unstratified and so does not define a set in *NFU* or predicate in L.

The Tarski paradox is blocked by the fact that quotation of a formula is a type-raising operation, preventing diagonalization. If we had terms representing constants, quoting them would raise types in the same way. Since the diagonalization is blocked by stratification, it is not a problem that the set of true sentences of L (as represented by their double singletons) proves to be definable in L by following our development above in *NFU*.

6. MODELS USING EXTERMAL AUTOMORPHISMS

The same situation can be modelled in the usual set theory using models of initial segments of the cumulative hierarchy with external automorphisms. In fact, this is how *NFU* itself is best modelled.

We work in a nonstandard model of the usual set theory *ZFC* (or of "enough" axioms of *ZFC*) with an external automorphism j and nonstandard infinite ordinal α such that $j(\alpha) > \alpha$. We consider V_α, stage α in the cumulative hierarchy.

[5]See *S. Orey 1964* for a discussion of the need for the Axiom of Counting here.

It is shown elsewhere[6] that V_α is readily interpreted as a model of *NFU*. The trick is to observe that V_α contains the much earlier stage $V_{j(\alpha)}$ of the cumulative hierarchy which looks exactly like it. Construe each element of $V_{j(\alpha)+1}$ as a set with its elements replaced by the inverse images under j of its actual elements and each element of $V_\alpha - V_{j(\alpha)+1}$ as an urelement (notice that there are a lot of urelements!). This is achieved by defining a new membership relation "$x \in_{new} y$" as "$x \in j(y)$ and $j(y) \in V_{\alpha+1}$". It is straightforward to prove that V_α with membership relation $\in_{new}$ is a model of *NFU*.

The semantics we have been doing in this paper can be clarified in the same context. It is possible to define the semantics for the "full language" (in which each set corresponds to a unary predicate and each relation to a binary predicate) on V_α in $V_{\alpha+2}$. If we replace elements of sets with their images under j as in the model construction for *NFU* and inflate power sets with urelements in the same way (setting up the construction to be carried out in *NFU* as above), we observe that the semantics for the "full language" on $V_{j^2\alpha}$ can be expressed in V_α in a way which translates successfully into terms of $\in_{new}$. The construction is in fact exactly that of the previous sections.

If L is the "real" full language on V_α, the language, which we see internally to V_α in the construction above, is $j^2(L)$. We cannot derive the Tarski paradox here because the needed predicate would be:

Predicate ϕ does not hold (in $j^2(L)$) of $j^2(\text{“}\phi\text{”})$,

which cannot be expressed in our working model of set theory because j is external. This is precisely analogous to the way in which the argument for Tarski's paradox fails above.

A system resembling the system of this paper in its semantic features, although not apparently motivated by work in Quine-style set theory, is discussed by Hiller and Zimbarg (*1984*). Their system uses additional assumptions which make it far stronger than the system *NFU* + Infinity + Choice used here, even when this is further extended with the Axiom of Counting.

REFERENCES

Andrews, P.

1986 *An introduction to mathematical logic and type theory: To truth through proof,* Academic Press, New York, pp. 264–265.

Forster, T.

1992 *Set theory with a universal set,* Oxford Logic Guides, no. 20, Clarendon Press, London.

Grishin, V. N.

1972 The equivalence of Quine's *NF* system to one of its fragments (in Russian), *Nauchnotekhnicheskaya Informatsiya,* series 2, vol. 1, pp. 22–24.

[6]See *Forster 1992,* pages 67–68.

Hiller, A. P., and J. P. Zimbarg
1984 Self-reference with negative types, *The Journal of Symbolic Logic*, vol. 49, pp. 754–773.

Jensen, R. B.
1969 On the consistency of a slight(?) modification of Quine's *NF*, *Synthese*, vol. 19, pp. 250–263.

Orey, S.
1964 New Foundations and the axiom of counting, *Duke Mathematical Journal*, vol. 31, pp. 655–660.

Quine, W. V.
1937 New foundations for mathematical logic, *American Mathematical Monthly*, vol. 44, pp. 70–80.

Rosser, J. B.
1953 *Logic for mathematicians*, McGraw-Hill, New York; and Chelsea, New York, 1978.