A fixed point theory over stratified truth

A fixed point theory over stratified truth

Andrea Cantini[*] [1]

[1] DILEF, Università degli Studi di Firenze

Received XXXX, accepted XXXX
Published online XXXX

Key words truth, stratification, μ-calculus
MSC (2010) 03A05, 03E70

We present a theory of stratified truth ST_μ with a μ-operator, where terms representing fixed points of stratified monotone operations are available. We prove that ST_μ is relatively intepretable into Quine's NF (or subsystems thereof). The motivation is to investigate a strong theory of truth, which is consistent by means of *stratification*, i.e. by adopting an implicit type theoretic discipline, and yet is compatible with self-reference (to a certain extent). The present version of ST_μ is an enhancement of the theory presented in [2].

1 The theory ST_μ of stratified truth

We investigate a strong theory ST_μ of truth, which is consistent by means of *stratification*, i.e. by adopting an implicit type theoretic discipline. To a certain extent, the theory is compatible with forms of self-reference and it is actually a sort of generalized μ-calculus (see [1] for a general reference and [12]).

Before describing the precise syntax of the theory and an appropriate notion of stratification, let us anticipate a few informal considerations. First of all, the whole universe structure has a built-in *reflection mechanism*: if A represents a given proposition, there is an object represented by $[A]$, which is transformed into *a statement of a higher level* by applying the truth predicate T to $[A]$. This means that, in order to preserve consistency, we have to keep track of this natural level stratification: if $[A]$ is given type level i, $T([A])$ is assigned type $i+1$.

We also stress that truth is regarded as a predicate T, which has a wider domain than usual truth predicates: T applies *not quite to sentences of an inductively defined formal language*, but to objects of the given universe, which play the role of propositions (henceforth termed as *propositional objects*). Hence the universe can be regarded as a sort of abstract syntax, which is closed under constructors representing *logical operations*, and which includes *propositional objects* $[A]$ and *predicative objects* $[x|A]$.

No type restriction is imposed in forming $[A]$ and $[x|A]$; but, as we shall see in the next subsection, the *use* of these expressions has to be suitably restricted according to the initial type-theoretic intuition.

If we exclude equality which is assumed as given, truth T is the only predicate of our language. But we also assume that there is a primitive operation pred, such that, roughly, $\mathrm{pred}(t,s)$ expresses the fact that the object represented by s falls under the concept represented by t. T and pred allow to express predication: if $[x|A]$ represents a predicate P defined by a given formula A, the result of the application of P to a, is rendered by $\mathrm{pred}([x|A],a)$. Hence the claim that a falls under P simply becomes the claim that $\mathrm{pred}([x|A],a)$ *is true*, i.e. $T(\mathrm{pred}([x|A],a))$.

One may wonder if self-referential constructions are, to a limited extent, allowed in the present framework. The answer is positive, but we stress that the present treatment is different from self-reference in the case of standard formal languages $\mathcal{L}$ (e.g. first order Peano arithmetic or ZFC), where one makes use of a substitution operation acting on Gödel numbers of $\mathcal{L}$-formulas and $\mathcal{L}$-terms. In the present framework, we assume that the universe is closed under a fixed point operator acting on terms, which depend *extensionally* on their parameters

This paper arose from the talks presented at the Conference *Ouroboros 2018. Formal Criteria of Self-Reference in Mathematics and Philosophy*, Hausdorff Center for Mathematics, Bonn 16.2-18.2, 2018, and at the Conference in honour of Gerhard Jäger's 65th birthday, Bern, December 12-13, 2018. The research has been suppported by PRIN 2017 (Florence Research unit), and the University of Florence (Fondi di ateneo 2018-2019).

[*] Corresponding author: e-mail: andrea.cantini@unifi.it

and are *homogeneously stratified*, i.e. arguments and values are assigned the same type. Roughly, the idea is that, if a propositional function F is extensional in a parameter x of given type i, then there is a fixed point c of type i, i.e. such that $F(c) = c$. Of course, one has to clarify what is the meaning of *extensional in* x and the notion of type assignment in a formally untyped framework. Once clarified, these requirements are sufficient to sterilize self-reference and Liar's arguments: according to the basic intuition, if $\mathrm{tr}(y)$ is the object representing the truth of y, $\mathrm{tr}(y)$ is *type raising*, i.e. is assigned type one greater than y itself. Hence no fixed point of $\mathrm{tr}(y)$ or of its negation will arise.

As to the justification, self-reference will follow by the set theoretic representation of logical constructors and as a consequence of (a variant of) the Knaster-Tarski theorem.

1.1 The syntax of ST_μ

Definition 1.1 The language $\mathcal{L}_T$. Besides the logical constants $\wedge$, $\neg$, $\forall$ countably many individual variables, parentheses, $\mathcal{L}_T$ includes:

(i) a unary predicate T for truth, a binary predicate $=$ for equality;

(ii) binary function symbols id, pred, and, pair; unary function symbols tr, neg, all, left, right, suc; an individual constant 0;

(iii) binding operators: $[-\,|\,-]$ (abstraction); μ (fixed point).

id, pred, tr internally represent *basic constructors* (i.e. constructors for atomic formulas), while neg, and, all internally represent *logical constructors*; pair, left, right represent an ordered paring operation and its corresponding projections.

If E is an expression, i.e. a term or a formula of the language, $FV(E)$ denotes the set of free variables of E.

Definition 1.2 [Terms and formulas of ST_μ] We give a simultaneous inductive definition of the notions of (1) *term*; (2) *term positive (negative) in a list $\vec{x}$ of variables*[1]; (3) *formula*; (4) *formula positive (negative) in a list $\vec{x}$.*

(i) Variables and the individual constant 0 are terms;

(ii) 0 is positive in $\vec{x}$, and every variable v is positive in $\vec{x}$, for every $\vec{x}$; if v does not occur in $\vec{x}$, v is also negative in $\vec{x}$; if $y \notin FV(E)$, E being an expression, then E is $\mathrm{Pos}(y)$ as well as $\mathrm{Neg}(y)$;

(iii) if t, s are terms, then

- $t = s$, Tt are formulas;
- $\mathrm{all}(t)$, $\mathrm{suc}(t)$, $\mathrm{neg}(t)$, $\mathrm{tr}(t)$, $\mathrm{left}(t)$ and $\mathrm{right}(t)$ are terms, as well as $\mathrm{id}(t,s)$, $\mathrm{and}(t,s)$, $\mathrm{pair}(t,s)$, $\mathrm{pred}(t,s)$;

(iv) if A, B are formulas, then $\neg A$, $A \wedge B$, $\forall x A$ are formulas, and $FV(\forall x A) = FV(A - \{x\})$; if A is a formula, $[x|A]$ is a term such that $FV([x|A]) = FV(A - \{x\})$;

(v) if t is positive (negative) in $\vec{x}$ and s is positive (negative) in $\vec{x}$, then $\mathrm{pair}(t,s)$, $\mathrm{and}(t,s)$, $\mathrm{id}(t,s)$, $\mathrm{left}(t)$, $\mathrm{right}(s)$, $\mathrm{suc}(t)$, $\mathrm{neg}(t)$, $\mathrm{all}(t)$ are all positive (negative) in $\vec{x}$;

(vi) if t is positive (negative) in $\vec{x}$, $\mathrm{pred}(t,s)$, $T(\mathrm{pred}(t,s))$, are positive (negative) in $\vec{x}$;

(vii) if t is positive (negative) in $\vec{x}$, then $T(\mathrm{neg}(\mathrm{pred}(t,s)))$ is negative (positive) in $\vec{x}$;[2]

(viii) if A is $\mathrm{Pos}(\vec{x})$ ($\mathrm{Neg}(\vec{x})$), then $\neg A$ is $\mathrm{Neg}(\vec{x})$ ($\mathrm{Pos}(\vec{x})$); if A, B are $\mathrm{Pos}(\vec{x})$ ($\mathrm{Neg}(\vec{x})$), then $A \wedge B$, $\forall v A$, $[y|A]$ are $\mathrm{Pos}(\vec{x})$ ($\mathrm{Neg}(\vec{x})$) (provided y not occurring in $\vec{x}$);

[1] The case of a single variable x is of course included.

[2] See the σ-interpretation in Definition 5.6 in NF for a motivation, and the axiom implying that T is well-defined on predication $\mathrm{pred}(f,t)$.

(ix) if t is a term with y free and positive in y, $\mu y.t$ is a term where y is bound; moreover, if y is distinct from the variables in the list $\vec{x}$ and t is $\mathrm{Pos}(\vec{x})$ ($\mathrm{Neg}(\vec{x})$), then $\mu y.t$ is $\mathrm{Pos}(\vec{x})$ ($\mathrm{Neg}(\vec{x})$).

Remark 1.3

1. The present notion of positivity is motivated by the fact that *under the intended set theoretic interpretation certain terms and formulas are monotone in their free parameters with respect to* $\subseteq$, see Definition 5.6 and Fact 5.2.

   Just to avoid misunderstandings, we stress that, if t, s are in $\mathrm{Pos}(\vec{x})$, $\mathrm{id}(t,s)$ is in $\mathrm{Pos}(\vec{x})$, but $t=s$, $\neg t=s$ are *neither* in $\mathrm{Pos}(\vec{x})$ nor $\mathrm{Neg}(\vec{x})$. On the contrary, at the atomic level, $\mathrm{Pos}(\vec{x})$ and $\mathrm{Neg}(\vec{x})$ are preserved only by pred-operation and by $Tpred$-application, and inherited under abstraction $[-|-]$.

   E.g. $T(\mathrm{pred}([x|T(\mathrm{pred}(a,x))],y))$ is positive in a, as well as $\mathrm{neg}(\mathrm{pred}([x|T(\mathrm{pred}(a,x))],y))$, while

   $$T(\mathrm{neg}(\mathrm{pred}([x|T(\mathrm{pred}(a,x))],y)))$$

   is negative in a.

2. The set theoretic language can be regarded as a sublanguage of $\mathcal{L}_T$, once we stipulate

   $$x \in a := T(\mathrm{pred}(a,x))$$
   $$a \subseteq b := \forall x(x \in a \to x \in b)$$

   Note that a is positive in $x \in a$, as expected.

Remark 1.4 The syntactical structures of terms in ST_μ can be entangled and nested, as illustrated by the simple examples below.

(i) If $t(y,x,z)$ is positive in y, x and negative in z, $\mu y.t(y,x,z)$ is a term positive in x and negative in z, and hence $\mu x \mu y.t(y,x,z)$ is a term negative in z.

(ii) $A(a,b,x) := T(\mathrm{pred}(a,x)) \wedge \neg T(\mathrm{pred}(b,x))$ is positive in a, negative in b, and hence $\mu y.[x|A(y,b,x)]$ is a term negative in b by the definition above, clause (ix).

By the preceding item, the term

$$[u|\neg T(\mathrm{pred}(\mu y.[x|A(y,b,x)],u))]$$

is positive in b and hence we have the term

$$\mu z.[u|\neg T(\mathrm{pred}(\mu y.[x|A(y,z,x)],u))].$$

(iii) Consider the formulas $A(x,y,a)$, $B(x,y,a)$ which are positive in x and negative in y. Then $\mu x[u|A(x,y,u)]$ is negative in y and hence $B(z,\mu x[u|A(x,y,u)],a)$ is positive in z and in y.

Hence the term $\mu z.[z|B(z,\mu x[u|A(x,y,u)],a)]$ is positive in y, and we finally have the term

$\mu y.[y|\mu z.[z|B(z,\mu x[a|A(x,y,a)],a)]]$.

(iv) The previous item makes clear in which sense the present version ST_μ of the theory is an enhancement over the theory discussed in [2]: ST_μlanguage is more expressive in that it contains terms which are not well-formed in the old theory, e.g. those containing abstracts $[x|A]$.

In order to state the T-schema and the comprehension schema, we extend *the discipline of types* to arbitrary expressions E of the new language. The basic idea is that predication makes sense only in agreement with a suitable modification of Quine's stratification: informally, a predicate (represented by) t truly applies to s, where s is assigned type i, only if t is assigned type $i+1$. Similarly, truth has an *implicit hierarchical structure*: when we apply the predicate T to (a propositional object represented by) the term t, T must be assigned a level higher than the type assigned to t.

Definition 1.5 (Stratification of terms and formulas) If E is an expression, E is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence and to each T-occurrence of E, so that:

1. all free occurrences of the same variable in any subexpression of E have the same type;

2. in each expression of the form $\mathrm{pred}(t, s)$ the type of t is *one greater than the type of its argument* s; $\mathrm{pred}(t, s)$ is assigned the type of t;

3. each expression of the form $\mathrm{tr}(t)$ is assigned a type one greater than the type of t; in each expression of the form $T(t)$ the predicate T is assigned *a type one greater than the type of* t;

4. in each expression of the form $t = s$, $\mathrm{id}(t, s)$, $\mathrm{pair}(t, s)$, $\mathrm{and}(t, s)$ the type of t is the same type as s; $\mathrm{id}(t, s)$, $\mathrm{pair}(t, s)$, $\mathrm{and}(t, s)$ are assigned the same type of t (and hence of s);

5. each expression of the form $\mathrm{neg}(t)$, $\mathrm{all}(t)$, $\mathrm{left}(t)$, $\mathrm{right}(t)$, $\mathrm{suc}(t)$ is assigned the same type of t;

6. each term of the form $[x \mid C]$ is assigned a type one greater than the type assigned to x, and all the free occurrences of x in C receive the same type;

7. in each expression of the form $\forall x A$, if x is free in A, then the free occurrences of x in A and the occurrence of x in $\forall x$ receive the same type;

8. each term of the form $\mu y t(y, \vec{x})$ is assigned the same type as y and t, and all the free occurrences of $\vec{x}$ in t receive the same type.

9. A formula (term) is $n + 1$-stratified iff it is stratified by means of $0, \ldots, n$.

Remark 1.6 Within the same statement, different occurrence of T can be assigned different type labels and this makes sense of the idea of *typical ambiguity* in the semantical framework we are dealing with. Observe also that the definition of stratification imposes a *homogeneity condition* on $\mathrm{and}(t, s)$ and $\mathrm{pair}(t, s)$; a semantical justification is to be found in the Quinean interpretation developed in section 5 below.

Definition 1.7 We inductively introduce $A \mapsto [A]$ with $FV(A) = FV([A])$:

$$
\begin{aligned}
&[t = s] := \mathrm{id}(t, s)\\
&[T(t)] := \mathrm{tr}(t)\\
&[\neg A] := \mathrm{neg}([A])\\
&[A \wedge B] := \mathrm{and}([A], [B])\\
&[\forall x A] := \mathrm{all}([x \mid A])
\end{aligned}
$$

Definition 1.8 (P-Form) If an object x is in the range of the logical constructors, then it is called a *P-form*:

$$
\begin{aligned}
Pfor(x) \quad \Leftrightarrow \quad & \exists y(x = \mathrm{tr}(y)) \vee \exists z(x = \mathrm{neg}(z) \vee x = \mathrm{all}(z)) \vee\\
& \vee \exists u \exists v(x = \mathrm{id}(u, v) \vee x = \mathrm{and}(u, v))
\end{aligned}
$$

Roughly, a P-form is an object which is (possibly) apt to represent a proposition. Observe that $Pfor(x)$ is stratified (assign 1 to x, u, v, z, 0 to y).

1.2 Axioms of ST_μ

ST_μ consists of the classical logical calculus (say, Hilbert-style) with equality and, in addition:

1. Compositional T-axioms:

$$T(\mathrm{id}(x,y)) \leftrightarrow x = y;$$
$$T(\mathrm{neg}(\mathrm{id}(x,y))) \leftrightarrow \neg x = y;$$
$$T(\mathrm{tr}(x)) \leftrightarrow T(x);$$
$$T(\mathrm{neg}(\mathrm{tr}(x))) \leftrightarrow \neg T(x);$$
$$T(\mathrm{neg}(\mathrm{neg}(x))) \leftrightarrow T(x);$$
$$T(\mathrm{and}(x,y)) \leftrightarrow T(x) \wedge T(y);$$
$$T(\mathrm{neg}(\mathrm{and}(x,y))) \leftrightarrow T(\mathrm{neg}(x)) \vee T(\mathrm{neg}(y));$$
$$T(\mathrm{all}(f)) \leftrightarrow \forall x T(\mathrm{pred}(f,x));$$
$$T(\mathrm{neg}(\mathrm{all}(f))) \leftrightarrow \exists x T(\mathrm{neg}(\mathrm{pred}(f,x)))$$

2. T-consistency:

$$\neg(T(a) \wedge T(\mathrm{neg}(a)))$$

3. T is well-defined on predication:

$$T(\mathrm{pred}(f,x)) \vee T(\mathrm{neg}(\mathrm{pred}(f,x)))$$

4. Stratified β-conversion: if A is stratified,

$$\begin{array}{rcl} T(\mathrm{pred}([x|A],u)) & \leftrightarrow & T([A[x := u]]) \\ T(\mathrm{neg}(\mathrm{pred}([x|A],u))) & \leftrightarrow & T([\neg A[x := u]]) \end{array}$$

   Roughly, this schema states that, insofar as stratified conditions and truth contexts are involved, predicate abstraction and predicate application behave as inverse to each other.

5. Self-reference: if t is positive in the list y, $\vec{x}$ and stratified,

$$\forall \vec{x}(t(\mu y t(y,\vec{x}),\vec{x}) = \mu y t(y,\vec{x}))$$

6. μ-Extensionality: if two terms positive (negative) in $\vec{x}$, y and stratified, are pointwise equal, then the respective fixed points coincide:

$$\forall \vec{x} \forall y(t(y,\vec{x}) = s(y,\vec{x})) \rightarrow \forall \vec{x}(\mu y t(y,\vec{x}) = \mu y s(y,\vec{x}))$$

7. Minimality: if t is positive in the list y, $\vec{x}$ and stratified,

$$\forall \vec{x}(t(a,\vec{x}) \subseteq a \rightarrow \mu y t(y,\vec{x}) \subseteq a)^3$$

8. P-form:

$$\begin{array}{rcl} T(x) & \rightarrow & Pfor(x) \\ \\ \neg Pfor(x) & \rightarrow & T(\mathrm{neg}(x)) \end{array}$$

   The P-form axioms grant that true objects lie in the range of logical constructors; furthermore, any object inaccessible to logical constructors is classified as (representing) False.

[3] We follow the notation of remark 1.3.

9. Ontological axioms: basic constructors and logical constructors are injective, but not surjective, and their images are disjoint; standard projections. In details, if f, g are distinct basic or logical constructors, f unary and g binary (including pair), then:

$$f(x) = f(y) \to x = y$$
$$g(x,y) = g(u,v) \to x = u \wedge y = v$$
$$\forall y(0 \neq \mathrm{suc}(y))$$
$$\forall x \forall y \forall z(f(x) \neq g(y,z))$$
$$\exists x \neg Pfor(x)$$
$$\mathrm{left}(\mathrm{pair}(x,y)) = x \wedge \mathrm{right}(\mathrm{pair}(x,y)) = y$$

Remark 1.9 The equivalence between $T(\mathrm{neg}(\mathrm{tr}(x))$ and $\neg T(x)$ is *strongly non-kripkean* and makes the truth predicate closer to its classical counterpart. A similar comment holds for the clause involving predication. Note that suc satifies the standard axioms for successor.

2 Stratified truth in ST_μ

Not surprisingly, T is provably *internally undefined* on (the simplest variant of) the Liar; but, interestingly, T *internally believes* this fact.

Proposition 2.1

(i) For some closed term L,

$$\mathsf{ST}_\mu \vdash \neg T(L) \wedge \neg T(\mathrm{neg}(L)).$$

Morover

$$\mathsf{ST}_\mu \vdash T(\neg T(L) \wedge \neg T(\mathrm{neg}(L))).$$

(ii)

$$\mathsf{ST}_\mu \vdash T(\mathrm{neg}(\mathrm{pred}(f,x))) \leftrightarrow \neg T(\mathrm{pred}(f,x))$$

P r o o f. As to (i), by self-reference choose $L = \mathrm{neg}(L) = \mu y.\mathrm{neg}(y)$. Then apply logic, T-consistency and the axioms relating T with tr, neg and and.

(ii): by T-consistency and the axiom that T is well-defined on predication. □

Proposition 2.2 (Uniform stratified T-schema) *If A is stratified, ST_μ proves:*

$$\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x}))$$

P r o o f. We check by simultaneous induction on A

$$(T([A]) \leftrightarrow A) \wedge (T([\neg A]) \leftrightarrow \neg A)$$

If A is of the form $t = s$, $T(t)$, apply the corresponding axioms of ST_μ.

If A is of the form $B \wedge C$, even if $B \wedge C$ is stratified, $[B \wedge C]$ may be not.[4] However, by $\forall$-instantiation of the compositional axiom about T and $\wedge$, we obtain

$$T([B \wedge C]) \leftrightarrow T([B]) \wedge T([C])$$

[4] For instance, $T(x) \wedge T([Tx])$ is stratified (e.g. assign 0 to x, 1 to the first occurrence of T and 2 to the second); but $[T(x) \wedge T([Tx])]$ is not, as it fails to meet the homogeneity condition required by and.

where the left hand side is not stratified in general, while the right hand side is stratified and IH[5] can be safely applied to it.

If A is of the form $\neg B$, apply the compositional ST_μ-axioms involving negated $\wedge$, double negation, and IH.

Let us consider the case of a negated universal quantifier. Then we use the axioms relating T, $\neg\forall$, together with β-conversion and IH in the final step:

$$\begin{aligned} T([\neg\forall x A]) &\leftrightarrow T(\mathrm{neg}(\mathrm{all}([x|A]))) \\ &\leftrightarrow \exists u T(\mathrm{neg}(\mathrm{pred}([x|A], u))) \\ &\leftrightarrow \exists u T([\neg A[x := u]]) \\ &\leftrightarrow \exists u(\neg A[x := u]) \equiv \neg\forall x A \end{aligned}$$

The case of positive $\forall$ is similar. □

The stratified T-schema implies that T strongly deviates from the behaviour of self-referential truth predicates à la Kripke-Feferman, which cannot in general be applied to the truth axioms themselves, nor to *arbitrary* logical axioms. On the contrary, T *provably believes that it is two-valued and consistent*; further, it recognizes that each closure condition is also internally true.

Corollary 2.3 (i) ST_μ *proves:*

$$T([T(a) \vee \neg T(a)])$$
$$T([\neg(T(a) \wedge T(\mathrm{neg}(a)))])$$

(ii) *Moreover, if* Axiom *is an instance of a compositional T-axiom or T-welldefinedness,* ST_μ *proves* $T([\mathit{Axiom}])$.

Proof. Observe that the consistency statement as well as *tertium non datur* for T and the compositional axioms are stratified; hence the claim is a consequence of the stratified truth schema. □

Remark 2.4 One may wonder whether the fixed point property can be extended, e.g. up to include the constructors tr, pred and combinations thereof. It is immediate to see that the answer is negative. Indeed, assume that there exists e such that

$$e = \mathrm{neg}(\mathrm{tr}(e))$$

Then $T(e) \leftrightarrow T(\mathrm{neg}(\mathrm{tr}(e))) \leftrightarrow \neg T(e)$: contradiction! The reason is that, roughly, as we shall see in the model construction, neither $x \mapsto \mathrm{tr}(x)$ nor $x \mapsto \mathrm{pred}(y, x)$ are monotone (in the sense of set theoretic inclusion) with respect to x.

We conclude by showing that ST_μ proves that its truth predicate is indeed the fixed point of a natural positive operator. Let $\mathcal{V}(x, T)$ be the formula:

$$\begin{aligned} \{ \quad & \exists v(\neg Pfor(v) \wedge x = \mathrm{neg}(v)) \vee \\ \vee \quad & \exists w_1((x = [T(w_1)] \wedge T(w_1)) \vee \\ & \vee(x = [\neg T(w_1)] \wedge \neg T(w_1))) \vee \\ \vee \quad & \exists w_2(x = \mathrm{neg}(\mathrm{neg}(w_2)) \wedge\ T(w_2)) \vee \\ \vee \quad & \exists w_3 \exists w_4((x = \mathrm{id}(w_3, w_4) \wedge\ w_3 = w_4) \vee \\ & \vee(x = \mathrm{neg}(\mathrm{id}(w_3, w_4)) \wedge w_3 \neq w_4)) \vee \\ \vee \quad & \exists w_5 \exists w_6(((x = \mathrm{and}(w_5, w_6)) \wedge\ T(w_5) \wedge T(w_6)) \vee \\ & \vee(x = \mathrm{neg}(\mathrm{and}(w_5, w_6)) \wedge\ (T(\mathrm{neg}(w_5)) \vee T(\mathrm{neg}(w_6))))) \vee \\ \vee \quad & \exists f((x = \mathrm{all}(f) \wedge \forall z T(\mathrm{pred}(f, z))) \vee \\ & \vee(x = \mathrm{neg}(\mathrm{all}(f)) \wedge \exists z T(neg(\mathrm{pred}(f, z)))))\} \end{aligned}$$

Remark 2.5 The formula is *stratified*: assign 1 to x, v, w_2, w_3, w_4, w_5, w_6, f; assign 0 to z and w_1; all occurrences of T are assigned type 2, except those acting on w_1.

[5] Henceforth IH stands for induction hypothesis in short.

Theorem 2.6 (Fixed Point Principle)

$$\forall x(T(x) \leftrightarrow \mathcal{V}(x,T))$$

Proof. $\Rightarrow$: let $T(x)$. Then $Pfor(x)$. If $x = \mathrm{neg}(y)$ and not $Pfor(y)$, clearly $\mathcal{V}(x,T)$. If $x = \mathrm{neg}(y)$ but $Pfor(y)$, we distinguish several cases and we apply the T-compositional axioms from left to right. E.g. if $x = \mathrm{neg}(\mathrm{id}(u,v))$, then $\neg u = v$ and we conclude $\mathcal{V}(x,T)$.
$\Leftarrow$: if $\mathcal{V}(x,T)$, we again argue by cases using T-axioms from right to left.

□

Corollary 2.7 (Internal Fixed Point Principle)

$$T[\forall x(T(x) \leftrightarrow \mathcal{V}(x,T))]$$

Proof. The statement of the fixed point theorem is stratified: hence apply the theorem and the stratified T-schema. □

3 Embedding the μ-calculus over arithmetic into ST_μ

3.1 ST_μ with numbers

Definition 3.1

1. $\mathrm{pair}(t,s) := (t,s)$, $\mathrm{left}(t) := (t)_0$, $\mathrm{right}(t) := (t)_1$
2. $N = \mu y.t(y)$, where $t(y) := [x|x = 0 \vee \exists z(z \in y \wedge x = \mathrm{suc}(z))]$
3. $<_N = \mu y.r(y)$, where

$$r(y) = [x|x = ((x)_0,(x)_1) \wedge ((x)_0 = 0 \wedge (x)_1 \neq 0) \vee$$
$$\vee((x)_1 = \mathrm{suc}((x)_0)) \vee \exists z(((x)_0,z) \in y \wedge (z,(x)_1) \in y))]$$

Proposition 3.2 (ST_μ)

1. $0 \in N \wedge \forall x(x \in N \rightarrow \mathrm{suc}(x) \in N)$
2. $x \in N \wedge x \neq 0 \rightarrow \exists y(y \in N \wedge x = \mathrm{suc}(y))$
3. if $A(x)$ is stratified, $A(0) \wedge (\forall x \in N)(A(x) \rightarrow A(\mathrm{suc}(x)) \rightarrow \forall x \in N.A(x)$
4. $<_N$ is irreflexive, transitive and connected on N; it satisfies, provably in ST_μ:

$$\forall a \forall b((0 <_N a) \wedge (a <_N \mathrm{suc}(a)) \wedge (a <_N \mathrm{suc}(b) \leftrightarrow a \leq_N b))$$

3.1.1 On the strength of ST_μ

Lubarsky ([12, p.296]) introduced a μ-calculus over Peano arithmetic, in which least fixed points rather than arbitrary fixed fixed points are required to exist. In the μ-calculus, it is possible to define sets of natural numbers, which occur higher up in the constructible hierarchy (see [12, p.292 and corollary, p.295]). On the proof-theoretic side, a formal system over Peano arithmetic $\mathsf{PA}(\mu)$, which embodies the μ-calculus, has been defined and investigated (see [7, Definition 1.3, p. 1463]). It turns out that $\mathsf{PA}(\mu)$ is proof-theoretically very strong, as made apparent by Theorem 1.6 in [7, p.1463]) (the paper refers to Möllerfeld's Ph.D. thesis [13] for proofs).

We now briefly outline $\mathsf{PA}(\mu)$.

Definition 3.3

(i) The language $\mathcal{L}_\mu$ of $\mathsf{PA}(\mu)$ is the standard language of second order arithmetic – as given in [17, p.2] and supplemented by the set constructor μ. It includes as individual terms the standard terms of the language of second order arithmetic, i.e. individual (number) variables, individual constants 0, 1 and terms inductively generated from variables and constants by application of the function symbols for plus and times. X, Y, Z, U, etc., are used as (meta) symbols for second order variables; they are intended to range over subsets of natural numbers.

(ii) The set terms and formulas of $\mathcal{L}_\mu$, as well as the collections $\mathrm{POS}(X)$ and $\mathrm{NEG}(X)$ of X-positive and X-negative $\mathcal{L}_\mu$-set terms and formulas are simultaneously generated by the following inductive clauses:

1. Every set variable V is a set term of $\mathcal{L}_\mu$ and belongs to $\mathrm{POS}(X)$, for any X; moreover, it belongs to $\mathrm{NEG}(X)$, for all X different from V.
2. If t, s are arithmetical terms, $t = s$, $t < s$ are formulas, which belong to $\mathrm{POS}(X)$ and to $\mathrm{NEG}(X)$, for all X.
3. If S is a set term of $\mathcal{L}_\mu$ and r is a number term, then $r \in S$ is a formula of $\mathcal{L}_\mu$; if S belongs to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$], then $r \in S$ belongs to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$].
4. If A is a formula of $\mathcal{L}_\mu$, then $\neg A$ is a formula; if A belongs to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$], then $\neg A$ belongs to $\mathrm{NEG}(X)$ [$\mathrm{POS}(X)$].
5. If A and B are formulas of $\mathcal{L}_\mu$, then so also is $(A \wedge B)$. If A and B belong to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$], then $(A \wedge B)$ belongs to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$].
6. If A is a formula of $\mathcal{L}_\mu$, then $\forall x A$, $\forall Y A$ are formulas; if A belongs to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$], then $\forall x A$, $\forall Y A$ (with Y distinct from X) belong to $\mathrm{POS}(X)$ [$\mathrm{NEG}(X)$].
7. if $A(u, X)$ is a formula with *no bound set variable*, *positive in* X, then $\mu x X.A(x, X)$ is a set term of $\mathcal{L}_\mu$, in which all free occurrences of the variables X and x are bound by the set term constructor μ. If U is distinct from X and A belongs to $\mathrm{POS}(U)$ [$\mathrm{NEG}(U)$], then $\mu x X A(x, X)$ is a set term in $\mathrm{POS}(U)$ [$\mathrm{NEG}(U)$].

We assume that $\wedge$, $\neg$, $\forall$ as primitive logical constants and we let $A \vee B$, $\exists x A$, $\exists X A$ as abbreviations for $\neg(\neg A \wedge \neg B)$, $\neg\forall x \neg A$, $\neg\forall X \neg A$ (in the given order).

An *arithmetical formula* is a formula in the language without set variables; it belongs to $\mathrm{POS}(X)$ and $\mathrm{NEG}(X)$, for every X.

Definition 3.4 The axioms of $\mathsf{PA}(\mu)$ include, besides standard classical logic for the two sorts of natural numbers and subsets of natural numbers:

1. standard axioms for 0, successor, plus, times, natural ordering <, the axiom of induction (as given, say, in [17, Definition I.2.4]);
2. arithmetical comprehension ACA: if A is a formula of $\mathcal{L}_\mu$ with no bound set variables, $\exists X \forall u(u \in X \leftrightarrow A(u))$;
3. for each first-order $A(u, X)$ formula of $\mathcal{L}_\mu$ positive in X, the axiom stating that $I := \mu x X.A(x, X, Y)$ is the least fixed point of the operator defined by A:
   - $\forall x(x \in I \leftrightarrow A(x, I)$
   - $\forall Y(\forall x(x \in Y \leftrightarrow A(x, Y)) \rightarrow I \subseteq Y)$

Then we can prove:

Proposition 3.5 (Lower bound on ST_μ) $\mathsf{PA}(\mu)$ *is interpretable in* ST_μ.

Proof. The argument is straightforward, but we sketch it. Firstly, define a translation $A \mapsto A^\tau$ of the language of $\mathsf{PA}(\mu)$ into the language of ST_μ by simultaneous induction on the definition of individual term, set term and formula. Then choose the set N of Definition 3.1 as domain of first order variables and the subsets of

N as domain of second order variables. Formally, this means that, if X is a second order variable, we choose a fresh variable x and translate $t \in X$ as $T(\mathrm{pred}(x,t))$. Then $(X \subseteq N)$ is τ-translated into the language of ST_μas:

$$\forall u(T(\mathrm{pred}(x,u)) \to T(\mathrm{pred}(N,u))$$

Hence we inductively define: $(\forall X A(X))^\tau := \forall x(x \subseteq N \to A^\tau[X := x])$.[6]

Then apply μ-axioms in ST_μ, in order to interpret the μ-term I in the Definition 3.4; again the μ-axioms are required to define the translation of the operations $+$ and $\times$, as the least functional relations, satisfying the standard recursive equations. As to the translation of the arithmetical comprehension schema ACA, simply observe that, if A is an arithmetical formula of $\mathcal{L}_\mu$, the term $[x \mid A^\tau(u)]$ is stratified, and hence we can apply stratified β-conversion and the schema 2.2.

□

4 Background on NF

Let $\mathcal{L}_s$ be the elementary set theoretic language, which comprises the binary predicate symbols $\in$. $\mathcal{L}_s$–terms are simply individual variables $(x, y, z, \ldots)$; prime formulas (atoms) have the form $t \in s$, $t = s$ (t, s terms). $\mathcal{L}_s$–formulas are inductively generated from prime formulas by means of sentential connectives and quantifiers. The elementary set theoretic language $\mathcal{L}_s^+$ is obtained by adding to $\mathcal{L}_s$ the abstraction operator $\{-\mid -\}$; $\mathcal{L}_s^+$–terms and formulas are then simultaneously inductively generated. The clause for introducing class terms has the form: if φ is a formula, then $\{x \mid \varphi\}$ is a term where $FV(\{x \mid \varphi\}) = FV(\varphi - \{x\})$ ($FV(E)$ is the set of free variables occurring in the expression E).
Two terms (formulas) are called α–congruent, if they only differ by renaming of bound variables; we identify α–congruent terms (formulas).

4.1 Stratified comprehension

As usual for Quine's systems, we need *stratification*; we also define a restricted notion thereof, which is motivated by the consideration of "loosely predicative" class existence axioms (see Definition 1.1, [3], p.131).

(i) φ is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence [7] of φ in such a way that

- if $t \in s$ is a subformula of φ, the type of s is one greater than the type of t; if $t = s$ is a subformula of φ, the type of s is the same as the type of t;
- all free occurrences of the same variable in any subformula of φ have the same type;
- if x is free in ψ and $\forall x\psi$ is a subformula of φ, then the 'x' in $\forall x$ and the free occurrences of x in ψ receive the same type;
- if $t := \{x \mid \beta\}$ occurs in φ, x is free in β, then t is assigned a type one greater than the type assigned to x, and all the free occurrences of x in β receive the same type.

(ii) $\{x \mid \varphi\}$ is stratified if φ is stratified;

(iii) a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *loosely predicative* iff for some type $i \in \omega$, $\{x \mid \varphi(x, \vec{y})\}$ has type $i+1$, no (free or bound) variable of $\varphi(x, \vec{y})$ is assigned type greater than $i+1$; a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *predicative* iff $\{x \mid \varphi(x, \vec{y})\}$ is loosely predicative and in addition no quantified variable of $\varphi(x, \vec{y})$ is assigned the same type as $\{x \mid \varphi(x, \vec{y})\}$ itself.

(iv) φ is $n+1$–*stratified* iff φ is stratified by means of $0, \ldots, n$.

For instance, $\bigcup a = \{x \mid (\exists y \in a)(x \in y)\}$ is not loosely predicative, since it requires type 2, but $\bigcup a$ itself has type 1; $a \cap b = \{x \mid x \in a \wedge x \in b\}$ is predicative.

[6] $A^\tau[X := x]$ is the formula we obtain by replacing each occurence of formulas of the form $s \in X$ by $T\mathrm{pred}(x, s)$.

[7] Individual constants included; these can be given any type compatible with the clauses below.

Definition 4.1 The system NF comprises:

(i) predicate logic for the extended language [8];

(ii) extensionality: $\forall x \forall y(x =_e y \to x = y)$, where

$$t =_e s :\Leftrightarrow \forall x(x \in t \leftrightarrow x \in s)$$

(iii) explicit comprehension SCA: if φ is stratified, then

$$\forall u(u \in \{x \,|\, \varphi(x, \vec{y})\} \leftrightarrow \varphi(u, \vec{y}))$$

Other systems

(iv) NFP (NFI) is the subsystem of NF, where SCA is restricted to (loosely) predicative abstracts.

(v) NF_k (NFI_k, NFP_k) is the subsystem of NF (NFI, NFP), where (at most) k types are allowed for stratification.

Remark 4.2 By a theorem of Crabbé ([3, p.134, Theorem 1]), NFI is provably consistent, say, in third order arithmetic (at most). The details of the (different) consistency proofs for NFI can be found in [3] and in [8, Theorem 3, p.187].

In order to carry out a Kripke-like construction in the NF-systems and to represent the syntax, we shall essentially exploit Quine's *homogeneous pairing operation*, which *does require extensionality* and the existence of a copy of the natural numbers. But it is not difficult to check that Quine's pairing is indeed well-defined already in NFI. First of all, the collection of Fregean natural numbers is a set in NFI.

Definition 4.3

$$\begin{aligned}
&\emptyset = \{x \,|\, x \neq x\};\\
&V = \{x \,|\, x = x\}\\
&0 = \{\emptyset\};\\
&a + 1 = \{x \cup \{y\} \,|\, x \in a \wedge y \notin x\};\\
&\mathrm{Cl}_N(y) \Leftrightarrow 0 \in y \wedge \forall x(x \in y \to (x + 1) \in y);\\
&\mathcal{N} = \{x \,|\, \forall y(\mathrm{Cl}_N(y) \to x \in y)\}
\end{aligned}$$

NFI proves the existence of $\mathcal{N}$; in fact, by inspection, all sets in Definition 4.3 are loosely predicative. Furthermore, we have, provably in NFI:

Lemma 4.4 (NFI) *Let $\{x \,|\, \varphi(x)\}$ be loosely predicative.*

$$\mathrm{Cl}_N(\{x \,|\, \varphi(x)\}) \to \mathcal{N} \subseteq \{x \,|\, \varphi(x)\}; \tag{1}$$

$$(\forall x)(x \in \mathcal{N} \leftrightarrow x = 0 \vee (\exists y \in \mathcal{N})(x = y + 1)); \tag{2}$$

$$\emptyset \notin \mathcal{N} \wedge (\forall x \in \mathcal{N})(V \notin x); \tag{3}$$

$$(\forall x \in \mathcal{N})(x + 1 \neq 0); \tag{4}$$

$$(\forall x \in \mathcal{N})(\forall y \in \mathcal{N})(x + 1 = y + 1 \to x = y) \tag{5}$$

Clearly $\mathcal{N}$ is infinite by (3) above. As to the proof, (3) holds in NFI + Union, as NFI+Union $\equiv$ NF by [3, p.131, Lemma 1], and NF proves (3) according to a famous result of Specker (see [18], [4, pp.48-49]). On the other hand, NFI + $\neg$Union implies (3) by [3, p.135]. The claims (2), (1) with the Peano axioms are provable in NFI ((5) requires the second part of (3)).

[8] If the abstraction operator is assumed as primitive, the extended logic contains the schema

$$\forall u(\varphi(u) \leftrightarrow \psi(u)) \to \{x \,|\, \varphi(x)\} = \{x \,|\, \psi(x)\}$$

Definition 4.5 (Quine's homogeneous pairing; see [15, Chap.10, pp.281–284], [16, p.380])

$$\begin{aligned}
\phi(a) &= \{y \,|\, y \in a \wedge y \notin \mathcal{N}\} \cup \{y+1 | y \in a \wedge y \in \mathcal{N}\};\\
\theta_1(a) &= \{\phi(x) \,|\, x \in a\};\\
\theta_2(a) &= \{\phi(x) \cup \{0\} \,|\, x \in a\};\\
(a,b) &= \theta_1(a) \cup \theta_2(b);\\
Q_1(a) &= \{z \,|\, \phi(z) \in a\};\\
Q_2(a) &= \{z \,|\, \phi(z) \cup \{0\} \in a\}
\end{aligned}$$

The definitions above are (at most) loosely predicative and hence the universe of sets is closed under the corresponding operations, provably in NFI.

We below exploit the fact that Quine's pairing operation and its projections are $\subseteq$-monotone in both arguments: indeed, the definitions of (a,b), $Q_1(a)$, $Q_2(a)$ are positive in a, b.[9]

Lemma 4.6 *We have, provably in NFI:*

1. $\phi(a) = \phi(b) \to a = b$ *;*
2. $0 \notin \phi(a)$*;*
3. $\theta_i(a) = \theta_i(b) \to a = b$*, where* $i = 1, 2$*;*
4. $(x,y) = (u,v) \to x = u \wedge y = v$.
5. *each projection* Q_i *(*$i = 1, 2$*) is* $\subseteq$*–monotone; moreover the map* $x, y \longmapsto (x,y)$ *is surjective and* $\subseteq$*–monotone in each variable:*

$$\begin{aligned}
&x \subseteq u \to Q_1(x) \subseteq Q_1(u) \wedge Q_2(x) \subseteq Q_2(u)\\
&x = (Q_1(x), Q_2(x))\\
&x \subseteq u \wedge y \subseteq w \to (x,y) \subseteq (u,w)
\end{aligned}$$

The proof hinges upon the properties of $\mathcal{N}$ and the successor operation (see [15, Chap.10, pp.280]).

Lemma 4.7 (Fixed point) *Let* $A(x,a)$ *be a formula which is positive in* a*. Assume that*

$$\Gamma_A(a) = \{x \,|\, A(x,a)\}$$

is loosely predicative, where x*,* a *are given types* i*,* $i+1$ *respectively. Then* NFI *proves the existence of a set* c *of type* $i+1$*, such that*:

(i) $\Gamma_A(c) \subseteq c$;

(ii) $\Gamma_A(a) \subseteq a \Rightarrow c \subseteq a$.

The proof is standard: observe that the set

$$c := \{x \,|\, \forall d(\Gamma_A(d) \subseteq d \to x \in d)\}$$

is loosely predicative.

[9] Reminder: a formula $A(x,a)$ is positive in a if every free occurrence of a in the negation normal form of A is located in atoms of the form $t \in a$, which are prefixed by an even number of negations and where $a \notin FV(t)$.

5 Embedding stratified truth in NF

5.1 Generating truth

We use Quine's pairing for representing logical constructors, and Lemma 4.7 to interpret the truth predicate.

Definition 5.1

$$\begin{aligned} \dot{\neg}x &:= (0,x);\\ x\dot{\wedge}y &:= (1,(x,y));\\ \dot{\forall}f &:= (2,f);\\ \dot{\in}xy &:= (3,(x,y));\\ \dot{=}xy &:= (4,(x,y)) \end{aligned}$$

We also write $[x=y]$ for $(\dot{=}xy)$. If $\{x\}$ denotes the singleton, we let

$$[x\in y] := \dot{\in}\{x\}y = y\cdot x$$

Under the dot-application, the universe of sets becomes an applicative structure. $y\cdot x$ is stratified only if y and x are given the types $i+1$ and i (respectively), and *the result of applying y to x is one greater than the type of x.*

By Lemma 4.6:

Fact 5.2

(i) $\dot{\neg}x$, $x\dot{\wedge}y$, $\dot{\forall}f$, $\dot{=}xy$ are $\subseteq$-monotone in x, y.

(ii) $y\cdot x$ is $\subseteq$-monotone in y.

We now model the Kripke-Feferman notion of self-referential truth within the abstract framework of Quine's set theory. First of all, in analogy with the notion of P-form (Definition 1.8), let

$$\begin{aligned} Pfr(x) \quad\Leftrightarrow\quad & \exists u\exists v(x=[u\in v])\vee\exists z(x=\dot{\neg}z\vee x=\dot{\forall}z)\vee && (6)\\ & \vee\exists w_1\exists w_2((x=[w_1=w_2])\vee(x=w_1\dot{\wedge}w_2)) && (7) \end{aligned}$$

$Pfr(x)$ is stratified (assign 1 to x, v, w_1, w_2 and 0 to u, z). The truth predicate W is introduced as the fixed point of a stratified positive (in a) operator $\mathcal{T}(x,a)$, which encodes the recursive clauses for partial self-referential truth and is given by the formula

$$\begin{aligned} \exists y \quad (\quad & x=\dot{\neg}y\wedge\neg Pfr(y))\vee\\ \exists u\exists v\exists w \quad [\quad & (x=[u\in v]\wedge u\in v)\vee\\ \vee \quad & (x=\dot{\neg}[u\in v]\wedge\neg u\in v)\vee\\ \vee \quad & (x=[v=w]\wedge\ v=w)\vee\\ \vee \quad & (x=[\neg v=w]\wedge\ \neg v=w)\vee\\ \vee \quad & (x=\dot{\neg}\dot{\neg}v\wedge\ v\in a)\vee\\ \vee \quad & (x=v\dot{\wedge}w\wedge\ v\in a\wedge w\in a)\vee\\ \vee \quad & (x=\dot{\neg}(v\dot{\wedge}w)\wedge\ (\dot{\neg}v\in a\vee\ \dot{\neg}w\in a))\vee\\ \vee \quad & (x=\dot{\forall}v\wedge\forall z(v\cdot z\ \in a))\vee\\ \vee \quad & (x=\dot{\neg}\dot{\forall}v\wedge\exists z(\dot{\neg}v\cdot z\in a))] \end{aligned}$$

Clearly $\Psi(a) := \{x\,|\,\mathcal{T}(x,a)\}$ is $\subseteq$-monotone in a and is *predicative* in the sense of 4.1 (iii): it receives type 2 once we assign type 0 to u, z, type 1 to x, y, v, w, type 2 to a, no quantifier acts on variables of type 2.

Definition 5.3

$$\begin{aligned} \mathrm{Cl}_T(a) &:= \forall x(\mathcal{T}(x,a)\to x\in a)\\ W &:= \{x\,|\,\forall a(\mathrm{Cl}_T(a)\to x\in a)\} \end{aligned}$$

The Fixed Point Lemma 4.7 immediately implies:

Proposition 5.4 NFI *proves:*

1. $\exists y(y = W)$;
2. $\forall a(\mathcal{T}(a, W) \to a \in W)$;
3. $\mathrm{Cl}_T(a) \to W \subseteq a$.

Remark 5.5 The interpretation of the truth predicate requires *an inductive definition over the universe which still yields a set, i.e. an object of the universe*. This makes essential use of the *peculiar impredicative features* of NFI. If we should try to carry out such definition over the standard set theoretic universe of ZFC, T would result in a *proper* class, and hence we should be forced to apply an impredicative theory of classes à la Morse-Kelley.

Definition 5.6 We inductively (and simultaneously) specify a translation $(-) \mapsto (-)^\sigma$ of terms and formulas of ST_μ into NF; below we use $(-,-)$, Q_1, Q_2, W for the corresponding set theoretic notions of Definitions 4.5, 5.3, 5.1:

$$\begin{array}{rcl}
x^\sigma & := & x \\
0^\sigma & = & 0 \\
(\mathrm{suc}(t))^\sigma & = & (t^\sigma) + 1 \\
(\mathrm{pair}(t,s))^\sigma & = & (t^\sigma, s^\sigma) \\
(\mathrm{left}(t))^\sigma & = & Q_1(t^\sigma) \\
(\mathrm{right}(t))^\sigma & = & Q_2(t^\sigma) \\
(\mathrm{pred}(t,s))^\sigma & = & [s^\sigma \in t^\sigma] \\
(T(t))^\sigma & = & t^\sigma \in W \\
(\mathrm{tr}(t))^\sigma & = & [t^\sigma \in W] \\
(t = s)^\sigma & = & (t^\sigma = s^\sigma) \\
(\mathrm{id}(t,s))^\sigma & = & [t^\sigma = s^\sigma] \\
(A \wedge B)^\sigma & = & A^\sigma \wedge B^\sigma \\
(\mathrm{and}(t,s))^\sigma & = & t^\sigma \dot{\wedge} s^\sigma \\
(\neg A)^\sigma & = & \neg A^\sigma \\
(\mathrm{neg}(t))^\sigma & = & \dot{\neg} t^\sigma \\
(\forall x A)^\sigma & = & \forall x (A)^\sigma \\
(\mathrm{all}(t))^\sigma & = & \dot{\forall} t^\sigma \\
[x|A]^\sigma & = & \{x|A^\sigma\} \\
(\mu y.t(y,\vec{x}))^\sigma & = & \{u|\forall z(t^\sigma(z,\vec{x}) \subseteq z \to u \in z)\}
\end{array}$$

Lemma 5.7 (Preservation Lemma)

(i) If A (respectively t) is a stratified formula (term) of ST_μ, then A^σ (t^σ) is a formula (term), which is stratified in the sense of NF,[10] such that $FV(A^\sigma) = FV(A)$ ($FV(t^\sigma) = FV(t)$).

(ii) Furthermore, if $t(y,\vec{x})$ ($A(y,\vec{x})$) is stratified positive (negative) in y, then $t^\sigma(y,\vec{x})$ ($A^\sigma(y,\vec{x})$) is stratified $\subseteq$-monotone ($\subseteq$-anti-monotone)[11] in y.

[10] Explicit typing of $T(x)$ yields $T^1(x^0)$, while $(T(x))^\sigma = x^1 \in W^2$. But $T(x)$ can also be typed as $T^2(x^1)$ modulo type shifting. Hence the σ-map is type-preserving, i.e. *the same type assignment* to variables and terms of A is involved, provided we take into account type shifting.

[11] This means: if $a \subseteq b$, then $t^\sigma(a,\vec{x}) \subseteq t^\sigma(b,\vec{x})$ and $[u|A(u,a,\vec{x})]^\sigma \subseteq [u|A(u,b,\vec{x})]^\sigma$ (respectively $t^\sigma(b,\vec{x}) \subseteq t^\sigma(a,\vec{x})$ and $[u|A(u,b,\vec{x})]^\sigma \subseteq [u|A(u,a,\vec{x})]^\sigma$).

P r o o f. Proceed by simultaneous induction on the definition of *term* (*formula*), possibly positive (negative) in given parameters.

If $A := T(t)$ is stratified, then so is t; hence by IH t^σ is stratified as well as $t^\sigma \in W$.

Let $A := t = s$ be stratified. Then so are t^σ and s^σ. But this implies that $(t = s)^\sigma \equiv t^\sigma = s^\sigma$ is stratified.

Let $t^\sigma := (\mathrm{pred}(s, r))^\sigma$ be stratified. Then $t^\sigma = [r^\sigma \in s^\sigma]$ is stratified too, since by IH s^σ is assigned a type which is one greater than the type assigned to r^σ. The cases where t (A) is built up by means of and, id, all, tr ($\wedge$, $\forall$) are straightforward by IH.

If $t := [x|A]$, then A^σ is stratified, whence $t^\sigma := [x|A^\sigma]$ is stratified.

Let $t := \mu y.r(y, x)$ be stratified. Then by IH $r(y, \vec{x})$ is stratified and positive (negative) in y, x. Hence $(\mu y r(y, x))^\sigma = \{u|\forall z(r^\sigma(z, x) \subseteq z \to u \in z)\}$ is stratified and $r^\sigma(y, x)$ is positive in y and x. Hence $(\mu y r(y, x))^\sigma$ is positive in x; thus, if $a \subseteq b$,

$$(\mu y r(y, a))^\sigma \subseteq (\mu y r(y, b))^\sigma.$$

If $t(y, \vec{x}) := \mathrm{pred}(r, s)$ is stratified positive (negative) in y, r has type one greater than the type of s and r is positive (negative) in y. Hence by IH r^σ is stratified monotone in y with type one greater than the type of s^σ, which is also stratified. It follows by definition of the σ-translation, the property of monotonicity of Lemma 4.6 and Quine's pairing that $t(y, \vec{x})^\sigma$ is stratified and monotone in y.

Let us check the case where

$$t(y, \vec{x}) = \mathrm{id}(s(y, \vec{x}), r(y, \vec{x}))$$

Then we have to show that, if $v \subseteq w$, $t(v, \vec{x})^\sigma \subseteq t(w, \vec{x})^\sigma$. By IH we have

$$\begin{array}{lll} u \subseteq w & \to & s(u, \vec{x}) \subseteq s(w, \vec{x}) \\ u \subseteq w & \to & r(u, \vec{x}) \subseteq r(w, \vec{x}) \end{array}$$

The conclusion again follows by item 5 of the Lemma 4.6, definition of id and Quine's pairing. The remaining cases when t is built up by means of neg, all are similar. □

Lemma 5.8 If A is a stratified formula of ST_μ, NF proves:

$$A^\sigma[x := u] \leftrightarrow [A^\sigma[x := u]] \in W$$

P r o o f. By induction on A, applying the previous lemma and Proposition 5.4 on W. We only consider the case of the universal quantifier. Then by applying stratified comprehension in the last step:

$$\begin{array}{lll} [\forall x A]^\sigma \in W & \leftrightarrow & \forall u([u \in [x|A^\sigma]] \in W) \\ & \leftrightarrow & \forall u(u \in [x|A^\sigma]) \\ & \leftrightarrow & \forall u A^\sigma[x := u] \equiv (\forall u A)^\sigma[x := u] \end{array}$$

□

Lemma 5.9 NF proves the σ-translation of extensionality:

$$(\forall x(x \in a \leftrightarrow x \in b) \to a = b)^\sigma$$

P r o o f. Assume the σ-translation of the antecedent of extensionality, i.e.

$$(\forall x(x \in a \leftrightarrow x \in b))^\sigma$$

Then apply Proposition 5.4. □

The σ-translation induces an interpretation into NF.

Theorem 5.10 If $\mathsf{ST}_\mu \vdash A$, then $\mathsf{NF} \vdash A^\sigma$.

P r o o f. It is enough to prove the σ-translation of the ST_μ-axioms. We repeatedly use Proposition 5.4 and the independence (or injectivity) of the chosen representation for the logical and descriptive symbols (see Definition 5.1).

(i) T is total on predication. Consider e.g. the σ-translation of

$$T(\mathrm{pred}(y,x)) \vee T(\mathrm{neg}(\mathrm{pred}(y,x))$$

This amounts to verify

$$([x \in y] \in W \leftrightarrow x \in y) \wedge ([\neg x \in y] \in W \leftrightarrow \neg(x \in y)), \tag{8}$$

which in turn follows from the second and third clauses of the inductive definition of W. By (8) also

$$([x \in y] \in W \vee [\neg x \in y] \in W \tag{9}$$

for every x, y. Hence if we choose $y := W$, we obtain:

$$([x \in W] \in W \vee [\neg x \in W] \in W \tag{10}$$

But (10) implies the σ-translation of the T-axioms involving tr. The verification of the extant cases ($=$, $\wedge$, $\forall$) is also routine.

(ii) T-consistency. Then we must prove the corresponding σ-translation, i.e.

$$\neg(x \in W \wedge (\dot{\neg}x) \in W)$$

Choose $\psi(x) := \neg((\dot{\neg}x) \in W)$. Then $\{x|\psi(x)\}$ is a set in NFI and it is easy to check:

$$\forall x(\mathcal{T}(x, \{x|\psi(x)\}) \rightarrow \psi(x))$$

The conclusion is a consequence of Proposition 5.4, item 3.

(iii) Stratified β-conversion: we want, if A is stratified,

$$(T(\mathrm{pred}([x|A], u))^\sigma \leftrightarrow T[A[x := u]])^\sigma$$

But by definition of W with Proposition 5.4, stratified comprehension and Lemma 5.7, we have:

$$\begin{array}{lcl} (T(\mathrm{pred}([x|A], u))^\sigma & \leftrightarrow & [u \in [x|A^\sigma]] \in W) \\ & \leftrightarrow & u \in [x|A^\sigma] \\ & \leftrightarrow & A^\sigma[x := u] \\ & \leftrightarrow & [A^\sigma[x := u]] \in W \\ & \leftrightarrow & (T[A[x := u]])^\sigma \end{array}$$

The remaining β conversion schema is similar.

(iv) Self-reference: let $t(y, \vec{x})$ be stratified positive (negative) in y. Then by Lemma 5.7, $t^\sigma(y, \vec{x})$ is stratified positive in y. Hence $\mu y t(y, \vec{x})^\sigma = \{u|\forall z(t^\sigma(z, \vec{x}) \subseteq z \rightarrow u \in z)\}$ satisfies the due fixed point equation by Lemma 4.7.

(v) μ-extensionality: straightforward.

(vi) Logical operators are injective: the σ-translation of the corresponding axioms is sound, simply because the logical operators act as ordered sequence operators, built upon Quine's ordered pair.

(vii) Logical operators are not surjective: in fact there are objects (e.g.$(4, a)$) which differ from $\dot{\neg}x$, $x\dot{\wedge}y$, $\dot{\forall}$ and $\dot{\in}xy$ (use Lemma 4.4). The images of the logical operators are trivially disjoint (we use distinct Fregean numbers as labels), and there are objects which are not P-forms, e.g. any ordered pair $(\emptyset, a)$.

□

Remark 5.11 The full strength of stratified comprehension is exploited in interpreting predication. Once predication is restricted to loosely stratified (or predicative) formulas, the resulting version of ST_μ becomes reducible to a consistent subsystem of NF.

6 Conclusion: stratified truth?

Let us try to assess some limits of the theory.

Why stratification? On one hand, that there is a type raising when we move from the mere claim of A to the claim of $T[A]$, can on intuitive grounds be conceded (at least according to the present author). And this is a good reason to pursue the typed theories of truth[12]. On the other hand, our practice with natural language tends to support the idea that we have to deal with *the* truth predicate, without any further type qualification; hence types ought to be left implicit or possibly avoided. Stratification can be regarded as a way to make both sides coexist. Of course, the awkward aspect is that we do appeal to a theory, which has certain unnatural features. Moreover, the consistency of ST_μ relies in its full strength upon a discipline – stratification – which is not fully understood, as shown by the yet unsolved (?) problem whether NF be consistent or not.

An additional unsatisfactory point is that the syntactical apparatus of ST_μ is, at the present stage of formalization, rather complex, and the stratification device is not so transparent as the corresponding explicit typed versions of truth.

Nevertheless, though type-theoretic in essence, ST_μ allows limited, yet non-trivial forms of self-reference, which are based after all on a semantical construction. And these limitations are apparently essential, in order to preserve consistency. In contrast with usual formal theories of truth, a distinctive feature of ST_μ is that it allows forms of *direct* self-reference (to make this clear with an example from recursion theory, the second recursion theorem instantiates indirect self-reference, while the first recursion theorem typically supports direct self-reference).

A positive interesting point might be that the compositional axioms of truth receive an unrestricted formalization in ST_μ, and the truth predicate believes that they are true (in sharp contrast, say, with Kripke-like systems). If we compare ST_μ with other strong axiomatic systems of truth, we must stress that a high degree of impredicativity is gained. The ground for it is the idea that the basic membership relation is well-defined and given, as made clear by the axiom of well-definedness for truth, and by the semantical clauses in NF governing the operator for inductively defining truth.

As to the relation with the literature, Holmes [9] explores the possibility that formal semantics is expressed in Quine's NFU, i.e. NFU with urelemente. In particular he shows that the reason why Tarski's argument fails, is not the undefinability of truth, but that the quotation operation becomes type-raising, causing the predicate needed for the 'Tarski sentence' to be unstratified and blocking diagonalization. Now, as already seen, something related happens in our case: the operation for encoding formulas of the form $T(x)$, $\neg T(x)$ is also type raising, and this forbids a form of the Liar leading to inconsistency. Of course, this is at present only a surface analogy. Indeed, a comparative look at [9] makes clear a specific limitation of ST_μ: its truth predicate T is not intended for metamathematical applications, as it is not defined on the inductively defined set of (codes of) sentences of the given *formal language*, say, of NF itself. T can only be applied to objects of the intended universe, which stand for propositions, whatever this means. In other words, the truth notion of ST_μ is an ontological notion, and is alien to standard semantical arguments, which make use of truth or satisfaction for inductively testing some form of (partial) soundness of the provability tools.

Just as the study of axiomatic theories of truth over standard set theory ZFC has been recently developed (see [5]), the investigation of axiomatic notions of truth over non-standard set theories like NF might be the next reasonable step to the present work.

7 Appendix I: the largest fixed point operator

By analogy with the μ-calculus, we can consistently extend our language and theory with the binding operator ν and the corresponding axioms:

- if t is positive in the list y, $\vec{x}$ and stratified,

$$\forall \vec{x}(t(\nu y t(y, \vec{x}), \vec{x}) = \nu y t(y, \vec{x}))$$

[12] For a thorough critical discussion of the distinction between typed and type-free theories of truth, we send the reader to [6], especially part II, and chapters 10-11 in part III.

- ν-Extensionality: if two terms positive (negative) in $\vec{x}$, y and stratified, are pointwise equal, then

$$\forall\vec{x}\forall y(t(y,\vec{x}) = s(y,\vec{x})) \to \forall\vec{x}(\nu y t(y,\vec{x}) = \nu y s(y,\vec{x}))$$

- Maximality: if t is positive in the list y, $\vec{x}$ and stratified,

$$\forall\vec{x}(a \subseteq t(a,\vec{x}) \to a \subseteq \nu y t(y,\vec{x}))$$

If we extend the σ-translation 5.6 to ν-terms by stipulating

$$(\nu y.t(y,\vec{x}))^\sigma = \{u|\exists z(z \subseteq t^\sigma(z,\vec{x}) \wedge u \in z)\}$$

then the Theorem 5.10 naturally extends to the system ST_μ with the ν-operator and its axioms.

8 Appendix II: on Yablo's paradox in a stratified frame

We consider a formalization of Yablo's paradox [19] in ST_μ. Define

$$t(f,x) = [\forall y^2(y^2 \in N^3 \wedge y^2 > x^2 \to \neg T^4(\mathrm{pred}(f^3, y^2))]^{13}$$

Proposition 8.1

$$\neg\exists f\forall x \in N(f \cdot x = t(f,x))$$

Indeed, the term t is stratified, but not positive in f. And the paradox shows that no solution to the equation above in f can exist.

8.1 Yablo's argument I

Recall that, if W is the truth set in NF, then

$$\begin{array}{rcl} \mathrm{pred}(f,t) & = & [t \in f] \\ & = & (\dot\in, (\{t\}, f) \\ \mathrm{tr}(t) & = & [t^1 \in W^2] \\ & = & (\dot\in, (\{t^1\}^2, W^2) \end{array}$$

Both expressions are positive (negative) in f, W, but NOT in t; W has type 2. Note that, if N is the set of natural numbers a la Frege in NF, then

$$[k^2 > n^2] = [\exists x^2(x^2 \in N^3)(k^2 = x^2 + n^2)]$$

8.2 Yablo's argument II

- Yablo sequences in NF. Consider the function

$$\begin{array}{rcl} f \cdot x & = & [\forall y(y \in N \wedge y > x \to \neg y \in f)] \\ & = & \mathrm{all}[y|y \in N \wedge y > x \to \neg y \in f] \\ & = & \dot\forall\{y^2|y^2 \in N^3 \wedge y^2 > x^2 \to \neg y^2 \in f^3\} \end{array}$$

Remark 8.2 This is not monotone in f! But it is stratified. On the other hand there exists f such that

$$\begin{array}{rcl} T(\mathrm{pred}(f,x)) & \leftrightarrow & T[\forall y(y \in N \wedge y > x \to T(\mathrm{pred}(f,y)))] \\ & \leftrightarrow & T(\mathrm{all}[y|y \in N \wedge y > x \to T(\mathrm{pred}(f,y))]) \\ & \leftrightarrow & \dot\forall\{y^2|y^2 \in N^3 \wedge y^2 > x^2 \to y^2 \in f^3\} \end{array}$$

[13] For the reader's sake, we have explicitly attached the type indices.

References

[1] A. Arnold and D.Niwinski, *Rudiments of μ-calculus*, North Holland, Amsterdam 2001.

[2] A. Cantini, On stratified truth, in: *Unifying the Philosophy of Truth* (T.Achourioti, H.Galinon, J.M.Fernandez, K.Fujimoto eds.), Springer, Cam 2015, 369-389.

[3] M. Crabbé, On the consistency of an impredicative subsystem of Quine's NF, *The Journal of Symbolic Logic*, 47, 1982, 131-136.

[4] T. Forster, *Set Theory with a Universal Set*, Oxford Logic Guides, no.31, Oxford, 1995.

[5] K.Fujimoto, Classes and Truths in Set Theory, *Annals of Pure and Applied Logic*, 163 (2012), no. 11, 1484-1523.

[6] V.Halbach, *Axiomatic Theories of Truth*, Cambridge University Press, Cambridge 2011.

[7] C. Heinatsch and M. Möllerfeld, The determinacy strength of Π^1_2-comprehension, *Annals of Pure and Applied Logic*, 161 (2010), 1462-1470.

[8] M.R.Holmes, The equivalence of NF-style set theories with "tangled" type theories: the construction of ω–models of predicative NF (and more), *The Journal of Symbolic Logic*, 60, 1995, 178–190.

[9] M.R.Holmes, Tarski's theorem and NFU, in: C.A.Anderson and M.Zelëny (eds.), *Logic, Meaning and Computation. Esssays in Memory of Alonzo Church*, Synthese Library, vol. 305, Kluwer Acad. Publ., Dordrecht, 2001, 469-478.

[10] G. Jäger, An intensional fixed point theory over first-order arithmetic, *Annals of Pure and Applied Logic*, 128 (2004), no. 1-3, 197–213.

[11] K.U. Kühnberger, B. Löwe, M. Möllerfeld, P. Welch: Comparing inductive and circular definitions: parameters, complexity and games. Studia Logica 81 (2005), no. 1, 79-98.

[12] R. S. Lubarsky, μ-definable sets of integers. J. Symbolic Logic 58 (1993), no. 1, 291–313.

[13] M. Möllerfeld, Generalized inductive definitions. The μ-calculus and Π^1_2-comprehension, Ph. D. thesis (Westfalische Wilhelms-Universität Münster, 2002).

[14] W.V.O. Quine, On ordered pairs, *The Journal of Symbolic Logic*, 10, 1945, 95-96.

[15] J. B. Rosser, *Logic for Mathematicians*, Mc Graw-Hill, New York 1953.

[16] D. Scott and D. McCarty, Reconsidering ordered pairs, *The Bulletin of Symbolic Logic*, Vol. 14, No. 3, 2008, 379-397

[17] S,G.Simpson, *Subsystems of Second Order Arithmetic*, Springer, Berlin-Heidelberg-New York, 1999.

[18] E.Specker, The axiom of choice in Quine's New Foundations for mathematical logic, *Proceedings of the National Academy of Sciences of the U.S.A.*, vol.39, 1953, 972-975.

[19] S.Yablo, Paradox Without Self-Reference, *Analysis* vol. 53, 1993, 251-252.