PANU RAATIKAINEN

THE CONCEPT OF TRUTH IN A FINITE UNIVERSE

Received 17 January 2000; revised version received on 4 April 2000

ABSTRACT. The prospects and limitations of defining truth in a finite model in the same language whose truth one is considering are thoroughly examined. It is shown that in contradistinction to Tarski's undefinability theorem for arithmetic, it is in a definite sense possible in this case to define truth in the very language whose truth is in question.

KEY WORDS: truth-definitions, finite models

1. INTRODUCTION

It has now become common wisdom that one cannot speak about the truth in a language in that same language, but that one has to use an essentially stronger metalanguage – at least, if one is not prepared to give up some basic laws of logic. This received picture is, of course, largely a consequence of Tarski's work, especially his justly celebrated theorem on the undefinability of truth.

My aim in this paper is to question this received picture slightly. My point of departure is the question: What if the universe is finite? For, although Tarski's undefinability theorem is known to hold for a wide range of theories, actually even the weakest such theories, for example Robinson Arithmetic Q, allow only infinite models: the classical methods of logic just require this (this remark will become more clear in what follows); no doubt one has assumed that theories allowing finite models are much too weak to have any use as a metatheory. Nevertheless, as I show below, certain more recent logical tools enable one to partly overcome this.

As long as one is dealing with the foundations of mathematics, it is indeed natural to consider theories that contain a reasonable amount of arithmetic, and consequently imply the infinity of their models – I am not trying to argue for strong finitism in the philosophy of mathematics. But when doing philosophy and theorizing about truth, the controversial infinite mathematical realm may not be the best possible starting point. And if we consider truth with respect to the material world, I think it is quite problematic if a theory of truth in itself rules out finite universes and implies that the world is infinite.[1]

Journal of Philosophical Logic **29:** 617–633, 2000.

Consequently, my approach below to the notion of truth will be a rather unusual. Figuratively speaking, I shall take an agnostic view concerning the size of the universe, that is, I shall allow the possibility that it is finite, although I make no claims about its size (in first-order logic, this allows the possibility that the universe may be infinite: one cannot then rule out this unless one determines the maximal size of the universe; this is a well known corollary of the compactness theorem). I then ask how the concept of truth can be approached in this framework. It turns out that the resulting picture differs in certain respects quite a lot from the standard view.

2. LOGIC AND FINITE MODELS

Recently, a great number of logicians have started to pay growing attention to the model theory of finite models, or briefly, finite model theory – not because of the concept of truth or any philosophical reason, but mainly because of its relevance for computer science.[2] Let us first review a couple of basic facts of this field that have been known for a long time.

The basic limitative result is Trakhtenbrot's Theorem (Trakhtenbrot, 1950), according to which the set of finitely valid first-order sentences (i.e. the sentences that are true in every finite model) is not recursively enumerable. This means that, in contradistinction to Gödel's completeness theorem, one cannot have a complete axioms and/or rules of inference for finitely valid sentences.

On the other hand, every finite model can be characterised up to isomorphism by a first-order sentence. It follows immediately that truth in any particular finite model can be completely axiomatized. For a model-theoretician, this may be a disappointing state of affairs (and consequently, finite model theory largely concentrates on different model classes). But viewing the issue philosophically, I don't think that there is any intrinsic value in the possibility to construct indefinitely many bizarre models of a theory.

As the notion of truth in a finite model is always decidable, it is (strongly) representable in any standard arithmetical theory. Hence, it is indeed easy to develop a complete theory of truth in finite models, say, in set theory or Peano Arithmetic (or, actually, even in Robinson Arithmetic). But in all these cases the metatheory is a theory of a very different level, as it claims that the universe is infinite, and this runs in the face of my finitist approach here. That is, all these theories are simply *false* in any finite model. This makes it impossible to compare theory and its metatheory in the way I would like to do. In what follows, I aim to investigate whether one can do any better. It turns out that one can.

Remarks. Before I proceed, it is in order to make more explicit what I exactly mean by "truth" here. Here I can be very short-spoken, for I shall simply follow the standard Tarskian inductive definition of truth in a model, which applies directly also to the finite models. Moreover, it may be best to emphasize that I shall commit myself to the classical logic, with bivalence and the law of the excluded middle (although in the context of finite domains, even intuitionists would not question their validity). So in these respects my approach is in complete accordance with Tarski's.

3. Varieties of definability

A central part of the received picture of truth is Tarski's Theorem, according to which, roughly speaking, truth is not definable (see Section 5). There are, however, various different kinds of notions of definability in logic (cf. Mostowski, 1965, pp. 29–30). Nowadays, Tarski's Theorem is often expressed by saying that arithmetical truth is not definable (by an arithmetical formula) in the standard model of arithmetic. This is the notion of *definability in a model*. However, one often considers rather *definability in a theory*; this general notion includes various sorts (weak, strong, etc.) of representability. (Tarski himself spoke of semantic definability and formal definability.) When considering finite models, there are good reasons to focus on the latter.

First, one simply cannot define any infinite set, however simple, in a finite model – or even a finite set larger than the universe of a given model. Especially, it is therefore trivially impossible to define in a finite model the set of all sentences true in that model. For, there are always infinitely many true (and false) sentences; hence the question of definability of truth in a finite model does not even make sense. On the other hand, *in a theory* allowing finite models it may very well be possible to define, or represent, even infinite sets, as we will soon see.

Second, one may note that definability in a model is in fact a special case of representability, namely, it coincides with representability in the complete theory of the model. Note that in this case weak and strong representability coincide (cf. Mostowski, 1962). Consequently one may as well state Tarski's Theorem as saying that arithmetical truth is not weakly representable even in the (non-axiomatizable) complete theory of arithmetic (cf. Section 5).

Representability is normally a much weaker notion than definability in a model, and if we can establish representability in a manageable theory, we have established a stronger result – and on the other hand, definability in a model may just be too strong a notion for some purposes. Moreover, as every finite model can be completely described by a first-order sentence, definability in a finite model can be always turned to definability in the theory consisting of such a sentence. Thus, one loses little if one focuses on definability in a theory. Therefore, in what follows, I shall mainly discuss the notion of definability in a theory. More exactly, I shall concentrate mostly on the notion of weak representability, and occasionally consider also strong representability.

DEFINITION 3.1. A formalized theory F *weakly represents* a set S, if for some formula $\varphi(x)$ of the language of F, $n \in S \Leftrightarrow F \vdash \varphi(\bar{n})$.

If, moreover, $n \notin S \Leftrightarrow F \vdash \neg\varphi(\bar{n})$, one says that F *strongly represents* the set S.

It is obvious how one generalizes these notions for relations.

4. The classical approach: Robinson Arithmetic and its extensions

For comparison, let us first review "the classical approach", by which I mean roughly the approach that was initiated in Gödel's seminal 1931 paper on incompleteness, and which culminated in the classical book *Undecidable Theories* by Tarski, Mostowski and Robinson, from 1953, where it was shown that a very simple theory Q, now standardly called Robinson Arithmetic, is sufficient to strongly represent all recursive sets. It follows that Q also weakly represents all recursively enumerable sets. The axioms of Robinson Arithmetic Q are the following:

(Q1) $s(x) \neq 0$,
(Q2) $s(x) = s(y) \rightarrow x = y$,
(Q3) $x \neq 0 \rightarrow (\exists y)(x = s(y))$,
(Q4) $x + 0 = x$,
(Q5) $x + s(y) = s(x + y)$,
(Q6) $x \times 0 = 0$,
(Q7) $x \times s(y) = (x \times y) + x$.

This simple theory proves the so-called Diagonalization Lemma (also known as the Self-Referential Lemma, or Gödel's Fixed Point Lemma), fundamental to various "limitative results" of logic. Let us assume that an effective Gödel numbering of the language of arithmetic has been fixed, and let us denote the Gödel number of an expression φ by $\ulcorner\varphi\urcorner$.

THE DIAGONALIZATION LEMMA 4.1. *For every formula $\varphi(x)$ with one free variable there is a sentence ψ such that $Q \vdash \varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi$.*

This lemma enables one to prove that Q and all its consistent axiomatizable extensions are incomplete and undecidable. It is also the key tool in the proofs of the various theorems on the undefinability of truth, to which we next turn.

5. UNDEFINABILITY OF TRUTH: DIFFERENT VARIANTS OF TARSKI'S THEOREM

Let us first state Tarski's undefinability theorem in terms of weak representability. In fact, this is close to the form in which Gödel gave it (Gödel, 1934), see also (Tarski et al., 1953) – both Gödel and, later, Tarski, Mostowski and Robinson only stated it in terms of strong representability, but the theorem holds even for weak representability. One can prove the following two diferent variants:

THEOREM 5.1. *One cannot weakly represent arithmetical truth in any consistent recursively axiomatizable extension of Robinson Arithmetic.*

THEOREM 5.2. *One cannot weakly represent arithmetical truth in any sound extension (in the language of arithmetic) of Robinson Arithmetic.*

I shall refer to these formulations as the undefinability of truth, or as the non-existence of a *truth-definition*.

Now originally (in his pathbreaking (1933a)) Tarski did not present his undefinability result in this form.[3] Rather, his version proves the non-existence of a *truth-predicate*. That is, Tarski required that an adequate definition of truth should entail all instances of his famous Convention T, e.g.

The sentence "Snow is white" is true (in L)
if and only if snow is white

and similarly for every sentence of a given language L. More formally, this can be put as follows:

DEFINITION 5.3. A formula $T(x)$ is called a *truth-predicate* for a theory F if for every sentence φ of the language of F, $F \vdash T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$.

(Often what I call here a truth-predicate is also called a truth-definition, but I prefer to use "truth-predicate" in order to clearly separate the two notions, for the reasons that will become evident below.) By simple application of the Diagonalization Lemma, one can easily prove the following two version of Tarski's theorem (cf. Smullyan, 1992, pp. 194–195):

THEOREM 5.4. *Let F be a sound formalized theory (i.e. F proves no theorem that is false in the standard model of arithmetic). Then there is no truth-predicate for F.*

THEOREM 5.5. *Let F be a consistent formalized theory which proves the diagonalization lemma. Then there is no truth-predicate for F.*

Discussion

Now, the truth in a particular *finite model* is decidable, and hence Q can even strongly represent truth of any first-order language in any finite model. In particular, it can deal with the truth of its *own language* in a finite model (although it clearly cannot have a truth-predicate, by Theorems 5.4 and 5.5). This is perhaps worth pointing out, for some passages of Tarski as well as many popular expositions may seem to suggest that such a situation is not at all possible.

Nevertheless, as I have already noted, this is not in all respects satisfactory, for Q itself allows only infinite models, is thus itself false in every finite model. (One might say in defence that although the universe is finite there is, from the metaperspective, something with infinite number, namely the true sentences; however, I shall not pursue this line further here.)

Our first two variants of Tarski's theorem (Theorems 5.1 and 5.2) concern the set of sentences true in the standard model, i.e. they are about the truth of a language in a specific model. The latter variants (5.4 and 5.5) on truth-predicates, on the other hand, say nothing about any particular model, but concern truth in a language – in any model. Now certainly the existence of a truth-predicate is sufficient to have a truth definition (in a model, or in *some* theory of the language); however, as we will shortly see, it is not a necessary condition: one may have a truth-definition without having a truth-predicate.

The literature has not always clearly distinguished between the above two variants of Tarski's theorem, that is, between the non-existence of truth-definition and the non-existence of truth-predicate; both follow easily from the Diagonalization Lemma, and utilize the Liar paradox – and the former implies the latter. However, in my finitary setting, it is of fundamental importance to clearly distinguish between these two (as will become clear later).

6. TRUNCATED ARITHMETIC TA

Let us next look how one could suitably weaken Robinson Arithmetic so that it would allow finite models, without losing all its proof-theoretical power. For this purpose, I have formulated the following finitary arithmetic that I call Truncated Arithmetic[4]; let us abbreviate it by TA. It is a theory in which every function symbol has the standard interpretation whenever its arguments and value are within the given finite model, and in the other case they are interpreted as the largest element of the model.[5]

Let us consider the language $L(\mathrm{TA}) = \{0, <, s, +, \times\}$ where 0 is a constant, $<$ a binary relation symbol, s unary function symbol and $+, \times$ are binary function symbols. Terms $0, s(0), s(s(0)), s(s(s(0)))\ldots$ are called *numerals*. The n-th numeral, whose standard interpretation is n (that is, which contains n occurrences of s), is denoted by $\bar{n}$. Now the axioms of TA are (the universal closures of) the following:

(A1) $\neg(x < x)$,
(A2) $x < y \vee y < x \vee x = y$,
(A3) $(x < y \wedge y < z) \rightarrow x < z$,
(A4) $\neg(x < 0)$,
(A5) $s(x) = y \rightarrow (\forall z)\neg(x < z \wedge z < y)$,
(A6) $[x < s(x)] \vee (\forall y)[(y < x \vee y = x) \wedge x = s(x)]$,
(A7) $x + 0 = x$,
(A8) $x + s(y) = s(x + y)$,
(A9) $x \times 0 = 0$,
(A10) $x \times s(y) = (x \times y) + x$.

The content of these axioms is the following: axioms (A1)–(A3) just state that $<$ is a strict total order (note that in TA '$<$' is not, in contradistinction to Q, definable in terms of '$+$'); (A4) that 0 is always the first element; (A5) says that the successor is always the immediate successor; (A6) guarantees that below the (possible) last element s behaves just like in the standard case, and beginning from it it collapses to the identity; (A7)–(A10) are just the familiar defining axioms for addition and multiplication (i.e. they are identical to the axioms (Q4)–(Q7)); however, whenever the model is finite and successor behaves non-standardly, also addition and multiplication collapse; their value is the last element whenever the "standard" value is larger than it.[6]

Clearly this theory is satisfied by a model with just one object, and has a model with each finite cardinality; obviously, it has also infinite models, and especially, it is satisfied by the standard model of arithmetic (thus it is sound). Hence, TA is semantically a very weak theory (and consequently, the incompleteness of TA can be shown by a simple semantical argument). However, it is surprisingly strong in another respect.

First, note that TA is sufficient to calculate correctly the value of any closed term as a numeral. Moreover, TA is complete with respect to all sentences with the form $(\exists x_1) \dots (\exists x_n)(t = t')$, where t, t' are any terms (these are called ∃-sentences), i.e. TA proves all sentences of this form that are true in the standard model $\mathcal{N}$ (these facts can be proved by simple inductions).[7] It follows that if a Diophantine equation has a solution, this fact can be proved in TA.

Now the logical tool than enables me (at least partly) to overcome the infinity assumption of the classical approach is the following: In 1970 Yuri Matiyasevich proved that Hilbert's tenth problem is unsolvable (Matiyasevich, 1970); but although he is most famous for this devastating negative result, his strong basic result was actually that every recursively enumerable set can be given a Diophantine representation, i.e. they all can be defined by a ∃-formula. Matiyasevich's result built essentially on earlier work by Julia Robinson, Martin Davis and Hilary Putnam (Davis, Putnam and Robinson, 1961). Hence it is often called the MRDP-theorem.

MRDP-THEOREM 6.1. *Every recursively enumerable set is definable by a ∃-formula.*

Now this entails, together with the above considerations, the following fact that will play a crucial role below:

THEOREM 6.2. *For any recursively enumerable set S one can find a ∃-formula (having the form* $(\exists x_1) \dots (\exists x_n)(t = t')$*) that weakly represents S in* TA.

In fact, if one added to TA the defining axioms of exponentiation, even the older result by Robinson, Davis and Putnam, from 1961, would be sufficient.

It follows that TA is what Smullyan (1961, 1993) has called a Gödel Theory: it is incomplete and undecidable; however, it is *not* a Rosser Theory, and hence, in Tarski's terminology, *not* essentially undecidable; that is, its every consistent extension is not undecidable. For one may add to it, for example, the sentence $(\forall x)(x = 0)$ – or, any sentence that determines the model's size to be some finite cardinality – and the resulting theory is complete and decidable.

Moreover, TA provides a particularly weak (even finitely satifiable) finitely axiomatizable theory that is undecidable, and hence one can use TA, instead of "infinitary" Q, to prove also *the undecidability of first-order logic*. That is, let φ_{TA} be a conjunction of the axioms of TA. Then a sentence ψ is a theorem of TA if and only if $(\varphi_{\mathrm{TA}} \rightarrow \psi)$ is a theorem of first-order logic. Hence a decision method for the latter would yield a decision method for the former, which is impossible.

7. ON THE DEFINABILITY OF TRUTH

Let us now return to our main subject, that is, the definability of truth (in finite models). Recall again that truth and falsity in a particular finite model are decidable, or recursive, and hence trivially recursively enumerable; therefore my theory TA can, by Theorem 6.2, weakly represent truth of any first-order language in any finite model. Especially, it can weakly represent truth-in-its-own-language in any finite model!

More exactly, assume that, in addition to a Gödel numbering of sentences (from now on, I shall assume that our numbering is defined for *all* first-order expressions, and not just for $L(\mathrm{TA})$), one has fixed a coding of finite models, i.e. an effective mapping from the set of finite models to the set of natural numbers. The code number of a finite model $\mathcal{M}$ is denoted by $\ulcorner\mathcal{M}\urcorner$. Then the following two facts follow directly from Theorem 6.2:

THEOREM 7.1. *There is a formula $Sat(x, y)$ of the language of* TA *such that* TA $\vdash Sat(\ulcorner\mathcal{M}\urcorner, \ulcorner\varphi\urcorner) \;\Leftrightarrow\; \mathcal{M} \models \varphi$.

THEOREM 7.2. *For any fixed finite model $\mathcal{M}$, there is a formula $Tr_{\mathcal{M}}(x)$ of $L(\mathrm{TA})$ such that* TA $\vdash Tr_{\mathcal{M}}(\ulcorner\varphi\urcorner) \;\Leftrightarrow\; \mathcal{M} \models \varphi$.

Moreover, let us again denote a conjunction of the axioms of TA by φ_{TA}. As $\varphi_{\mathrm{TA}} \vdash Tr_{\mathcal{M}}(\ulcorner\psi\urcorner)$ for every true ψ, by the Deduction Theorem (or, the introduction rule for implication), $\varphi_{\mathrm{TA}} \rightarrow Tr_{\mathcal{M}}(\ulcorner\psi\urcorner)$ is *a theorem of logic* if and only if $\mathcal{M} \models \psi$.

Discussion

As Theorems 7.1 and 7.2 apply in particular to $L(\mathrm{TA})$, one may thus in TA (and its sound extensions) obtain, in an exact sense, a complete truth-definition in the very same language whose truth one is considering. The state of affairs here is thus the direct opposite to the infinitary case, i.e. Theorems 5.1 and 5.2. The situation is this comfortable as long as one considers the issue purely in the proof-theoretical level, from the point of view of TA, as I have done. The complications begin to occur when one considers these truth-definitions as interpreted in a particular *finite* model of TA. Then not only all true (in the intended sense) instances of $Tr_{\mathcal{M}}(\ulcorner\psi\urcorner)$, but also some false ones, become true. (This is because in “too small" models, some true sentences of the form $(\forall x_1)\ldots(\forall x_n)(t \neq t')$ become false.) Only the true ones are satisfied only by the *infinite* standard model of arithmetic $\mathcal{N}$! Similar qualifications must be made, *mutatis mutandis*, with respect to $Sat(x, y)$. This is a definite limitation in my approach, but unavoidable as long as there are infinitely many truths (cf. Section 10) – and, in the case of $Sat(x, y)$, infinitely many finite models. Nevertheless, I think that the above partial positive results are philosophically rather interesting.

It is important to note that weak representability, unlike strong representability, is not preserved in all consistent extensions. In particular, not every complete extension of TA weakly represent the truth; especially, a complete theory of a finite model does not weakly represent truth in that model. For the case is exactly the same as in the case of definability in a finite model, discussed in the very beginning of this paper. Namely, such a complete theory implies that the universe is finite, and again, the question of defining the set of infinitely many truths does not even make sense. For similar reasons, one can even less hope to strengthen the above results to strong representability of truth by extending TA to a complete theory of a finite model. Strong representability can only be obtained by adding axioms of infinity, i.e. by moving to Q or some stronger theory that has only infinite models.

8. FAILURE OF CONVENTION T

Thus far I have shown that in the case of finite models, it is possible to have a truth-definition for a language in that same language. On the negative side, one must recognize that one just cannot satisfy Tarski's adequacy condition, Convention T. (This was implicit already in my discussion above (in Section 5), when I noted that a truth-predicate is not relative to any model, but defines truth-in-a-language in any model.) This is a simple consequence of a variant of Tarski's Theorem, Theorem 5.4, for the axioms of TA are all true in the standard model of arithmetic and thus TA is sound in the relevant sense. Thus we have:

COROLLARY 8.1. *There is no truth-predicate for* TA.

In the case of our truth-defining formulas $Tr_{\mathcal{M}}(x)$ (or more generally, any formula that weakly represents truth), this fact can be seen also more directly. For assume that TA would, for some finite model $\mathcal{M}$, prove for $Tr_{\mathcal{M}}(x)$ all the instances of Convention T. Let then φ be a sentence that determines the (finite) cardinality of $\mathcal{M}$. We have assumed that

TA $\vdash Tr_{\mathcal{M}}(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$; further, as $Tr_{\mathcal{M}}(x)$ weakly represents truth-in-$\mathcal{M}$, and φ is true-in-$\mathcal{M}$, also TA $\vdash Tr_{\mathcal{M}}(\ulcorner\varphi\urcorner)$. Thus, by simple logic, TA $\vdash \varphi$. But this cannot be the case: φ is a sentence that fixes the cardinality of the model, whereas TA has models of each finite cardinality as well as infinite ones. Hence the assumption is false.

Discussion

The above consideration shows that Tarski's famous Convention T, i.e. the existence of a truth-predicate is, although sufficient, not necessary for having a truth definition.

That is, it is true that in the standard case, when one considers truth (in the language of arithmetic) in the (infinite) standard model $\mathcal{N}$ of arithmetic, there is neither a truth-predicate nor a truth definition (the latter implies the former), and there is consequently little need to sharply distinguish between these two notions. But if one focuses solely on the finite models, the situation is radically different. Although one still cannot have a truth-predicate (not in TA, not in Q or PA, not even in the highly non-effective and strong complete theory of arithmetic), it is possible to have a truth-definition in the sense of weak representability. Further, Q and its extensions can even strongly represent truth in a finite model, and this holds also for their own language, when interpreted such that it allows a finite model (as in the case of TA).

It thus begins to look as though the demand for a truth-predicate is an unnecessarily strong requirement when one considers exclusively finite models. In fact, the reason why Tarski originally approached truth via truth-predicates, i.e. why Tarski put forward his Convention T as the condition of material adequacy of a truth-definition, was that he was aware that in arithmetic and set theory, truth is highly non-constructive and inaccessible, and needed a way to define truth independently of any "criterion" for recognizing truths, the latter being impossible in these cases (cf. Etchemendy, 1988). In our case of finite models no such problems arise, for truth is decidable, and the detour via truth-predicates becomes unnecessary.

I have emphasized above that a truth-predicate has nothing to do with truth in a particular model, but fixes truth-in-a-language in any model. But it is interesting to note that so does *in the class of finite models*, although in a different way, also my satisfaction relation $Sat(x, y)$ (see Theorem 7.1).

9. The Liar paradox?

It is immediate from our above consideration that TA can also define (i.e. weakly represent) falsity (in a finite model) in its own language. At this point, one is probably wondering whether the Liar paradox and other semantical anomalies threaten, and whether my approach is even consistent. The short but uninformative answer is: TA is obviously consistent, as it has a model, and even the trivial model of one object. In fact, one may note in passing that in a sense, TA can even prove its own consistency. That is, Theorem 7.1 trivially implies the following (where φ_{TA} is, again, a conjunction of the axioms of TA):

THEOREM 9.1. $\mathrm{TA} \vdash (\exists x) Sat(x, \ulcorner \varphi_{\mathrm{TA}} \urcorner)$.

Nevertheless, one should take this with a grain of salt and not read too much into it: our $Sat(x, y)$ is apparently not intensionally but only extensionally correct, and one cannot prove in TA many basic facts on satisfiability.

But be that as it may, the closer analysis of the reasons for the failure of constructing the Liar paradox for TA is actually quite interesting. Let us next examine in detail what happens.

First, recall that the infinite standard model of arithmetic $\mathcal{N}$ is also a model of TA. Consider then a $\exists$-sentence $F(x)$ that defines Falsity-in-$\mathcal{M}$ (for a fixed finite model $\mathcal{M}$). Now by the Diagonalization Lemma, there is a sentence ψ such that $\mathrm{Q} \vdash F(\ulcorner \psi \urcorner) \leftrightarrow \psi$, and hence the equivalence holds in particular in the infinite standard model of arithmetic $\mathcal{N}$, i.e.

$$(*) \qquad \mathcal{N} \models F(\ulcorner \psi \urcorner) \leftrightarrow \psi.$$

But what is the truth-value of ψ in $\mathcal{M}$, and in $\mathcal{N}$?

(i) Assume first that ψ is false in $\mathcal{M}$. Then by weak representability of falsity, $\mathrm{TA} \vdash F(\ulcorner \psi \urcorner)$, which is thus true in every model of TA, and in particular in $\mathcal{N}$. By the above equivalence $(*)$, also ψ must be true in $\mathcal{N}$. In sum, if ψ is false in $\mathcal{M}$, it is true in $\mathcal{N}$.

(ii) Assume then that ψ is true in $\mathcal{M}$. Thus $F(\ulcorner \psi \urcorner)$ must be false in $\mathcal{N}$, and ψ too must be (by $(*)$) false in $\mathcal{N}$.

Thus, independently of the truth-value of ψ, it must necessarily have a different truth-value in $\mathcal{M}$ and in $\mathcal{N}$. It follows that neither ψ nor $\neg\psi$ is provable in TA. (One thus obtains an undecidable sentence by an application of the Liar paradox somewhat different from the standard Gödelian reasoning.) Consequently, the attempt to construct the Liar paradox fails because the “paradoxical” sentence has a non-standard interpretation in the finite model $\mathcal{M}$ and fails to have the intended anomalous meaning;

it, so to say, necessarily goes beyond the finite model in question. It thus turns out that a genuine self-reference requires, in the formalized first-order languages, axioms of infinity![8]

10. Finitely many truths?

Thus far, I have considered the whole infinite totality of sentences true in a finite model; and this has also led to certain complications. But one may note that the "whole truth" of a finite model can be expressed by one sentence, and that there are only finitely many truths "simpler" than this sentence. Consequently, it would not be terribly restrictive to focus ones attention to a finitely limited set of sentences. That is, one may fix a natural (e.g., lexicographical) ordering of the sentences, and beginning from the simplest, focus on a finite number of truths (in a fixed finite model) including a complete description of the model in question.

If this line is chosen, even a list-like definition of truth, discussed in passing by Tarski (1933a, 1969) (cf. Etchemendy, 1988, Heck, 1997, Soames, 1998), is not that inappropriate. Let us thus assume that in a particular model $\mathcal{M}$ the relevant true sentences are among $\varphi_1, \varphi_2, \ldots, \varphi_n$ (the first n sentences in the chosen Gödel numbering). Now consider a formula of L(TA) having the following form:

$$(D) \qquad (x = \ulcorner\varphi_1\urcorner \wedge \varphi_1) \vee (x = \ulcorner\varphi_2\urcorner \wedge \varphi_2) \vee \cdots \vee (x = \ulcorner\varphi_n\urcorner \wedge \varphi_n)$$

Such a formula would then be a perfectly satisfactory truth-definition. However, unlike in Tarski's case, with TA as a metatheory (D) does not satisfy Tarski's criterion of material adequacy, i.e. Convention T. This is because (D) implies all the instances of Convention T (for $\varphi_1, \varphi_2, \ldots, \varphi_n$) only if the background metatheory proves $\ulcorner\varphi_i\urcorner \neq \ulcorner\varphi_j\urcorner$, when $i \neq j$ (Tarski makes this assumption explicitly in (1933a)). But clearly TA proves no such things. However, such inequalities do hold in a large enough finite model $\mathcal{M}$ (i.e. $card(\mathcal{M}) \geq max(i, j)$), and are provable in a theory extending TA that implies that there are enough different entities, e.g., a complete theory of such a model $\mathcal{M}$; such a theory even proves all the instances of Convention T, for these sentences.

Consequently, in a favourable case it is in a sense even possible to define truth in a finite model in the very model. Namely, one may naturally distinguish simple and complex models, according to the complexity (i.e. its place in the chosen natural numbering) of the simplest complete description of the model. If a model is simple and relatively large, it may indeed be possible to define in the model "the whole truth" in that model. The same fact holds, *mutatis mutandis*, with respect to strong representability in a complete theory of a finite model. Thus definability of truth in a model becomes a somewhat contingent matter depending on the size and complexity of the model. (Moreover, it clearly depends on the Gödel numbering of the sentences. The most natural choice would perhaps be to enumerate sentences in their lexicographical order.)

It is quite clear that a list-like definition such as (D) cannot lead to any contradiction. But it is still illuminating to note why it does not allow a construction of the Liar paradox. The explanation was clearly seen already by Tarski himself (1969, p. 65): such a truth definition, and obviously also its negation, contains all the sentences $\varphi_1, \varphi_2, \ldots, \varphi_n$ as proper parts, and cannot coincide with any of them.

Now if one were only interested in truth (in the present, restricted sense, i.e. as a finite set of truths) in a *particular* finite model, a list-like definition like the above (D) would be all that is needed. However, our earlier considerations allow a much more general defintion: our predicate $Sat(x, y)$ provides a definition of the truth in any finite model, and it picks up the set of finitely many relevant truths in any sufficiently large and simple model of TA. Hence the logical tools I developed earlier have their uses also in the present case where one concentrates on finite fragments of language.

11. CONCLUSIONS

One may conclude that concentrating one's focus on the finite models, on the one hand, makes logical truth, i.e. truth in every model, more difficult to manage (non-axiomatizable), but on the other hand, makes truth *simpliciter* quite accessible. In the case of finite models, one can, in a precise sense, give a complete definition of truth for a language in this same language – and moreover, for any other language and finite model. And this does not require one to give up any laws of classical logic.

No doubt there are numerous aspects of the above treatment that would deserve further elaboration. But I think that the considerations presented here already justify the conclusion that the study of the behaviour of truth-definitions in the case of finite models is, both logically and philosophically, rather interesting.

ACKNOWLEDGEMENTS

Earlier versions of this paper were read in "Philosophical Logic Seminar" at the Department of Philosophy, University of Helsinki, in September

1997, and in “Mathematical Logic Seminar”, at the Department of Mathematics, University of Helsinki, in December 1998. I am grateful for the comments of Ilkka Niiniluoto and Gabriel Sandu on the former, and Lauri Hella and Jouko Väänänen on the latter occasion.

I also thank Donald Smart for revising my English.

Finally, I am deeply indebted to the anonymous referee, whose comments to an earlier version of the paper were of high value.

NOTES

[1] I think it remarkable that Tarski – unlike, as far as I know, any of his later followers – did sense that *the infinite character* of his metatheory is philosophically problematic: “From the intuitive standpoint this may seem doubtful and hardly evident”, he wrote (1933a, p. 174). And again: “The intuitive obviousness of the axioms in not uncontested” (1933b, 282, n2). Further, Tarski confesses that if “expressions are regarded as the products of human activity”, then “the supposition that there are infinitely many expressions appears to be obviously nonsensical” (1933a, p. 174, n2). Tarski then considers another possible interpretation. “We could consider all physical bodies of particular form and size as expressions. The kernel of the problem is then transferred to the domain of physics. The assertion of the infinity of the number of expressions is then no longer senseless although it may not [*sic*!] conform to modern physical and cosmological theories” (1933b, p. 174, n2). Tarski leaves the question somewhat open.

Tarski even considered in passing the possibility I pursue in this paper: “The consequences mentioned could of course be avoided if the axioms were freed to a sufficient degree from existential assumptions.” (1933a, p. 175) Tarski adds, however, that such a weakening of axioms “would considerably increase the difficulties of constructing the metatheory, would render impossible a series of the most useful consequences and so introduce much complications into the formulation of definitions and theorems.” For these reasons, Tarski concludes that “it seems desirable, at least provisionally” to base his theory of truth on the unweakened infinitary axioms. My investigations in this whole paper may considered as a complementary commentary to these remarks by Tarski.

But be that as it may, I think that these careful qualifications by Tarski show his greatness and sensitivity as a philosophical thinker.

[2] See, e.g., Gurevich, 1984, Ebbinghaus and Flum, 1995.

[3] For historical details concerning Tarski’s Theorem, and Gödel’s independed discovery of it, as well as for different formulations they gave for it, see (Murawski, 1998).

[4] This apt name was suggested by Lauri Hella.

[5] Smorynski and Grohe have used theories that resemble somewhat TA, and their approaches have inspired my construction of TA; however, they both use only relation symbols, whereas I have used function symbols in order to have numerals and the standard notion of (weak) representability. Further, neither of them combine their theories with MRDP-Theorem in the essential way I do in Theorem 6.2. My theory TA resembles also the “Baby Arithmetic” of Bell and Machover; and Bell and Machover do combine their theory with MRDP-Theorem, if only to exhibit a weak theory that is undecidable. However, their theory has infinitely many axioms; whereas I have preferred to construct a finitely

axiomatizable theory (cf. note 7 below). Anyway, the treatment of Bell and Machover has importantly inspired my present approach.

[6] It is quite obvious that there is a certain amount of redundancy in my axioms of TA, and not all of them are really needed in the considerations that follow, but I think they form an intuitively clear theory that expresses nicely what is going on here.

[7] A sceptical reader may convince her/himself by noting that these facts hold already for the "Baby Arithmetic" of Bell and Machover (see pp. 335–336) which is clearly a subtheory of my TA (indeed, it is deducible already from my axioms (A7)–(A10); cf. Bell and Machover, p. 341). (See also Machover, 1996.)

[8] I have here used only the standard fact that the Diagonalization Lemma is provable for Q and hence true in the standard model of arithmetic $\mathcal{N}$. One may wonder whether it is possible to improve this and prove the Diagonalization Lemma for TA. I prefer to refrain from a judgement: On the one hand, the standard proofs of the Diagonalization Lemma cannot be carried through in TA, for they require the diagonalization function to be strongly representable. On the other hand, I have a hunch that one could, with some extra labour, demonstrate a version of the Diagonalization Lemma even for TA. However, I saw little point in taking pains and attempting to do that. For, I have already shown that TA is undecidable and incomplete. The only consequence I can see in the present setting would be the following: if one had the Diagonalization Lemma even for TA, then one could conclude that the "paradoxical" Liar sentence ψ is (besides being undecidable in TA) true in $\mathcal{M}$. However, it is all too easy to find sentences that are true in a fixed finite model but unprovable in TA (e.g., cardinality claims).

REFERENCES

Bell, J. and Machover, M. (1977): *A Course in Mathematical Logic*, Elsevier, Amsterdam.

Davis, M., Putnam, H. and Robinson, J. (1961): The decision problem for exponential diophantine equations, *Ann. of Math. (2)* **74**(3), 425–436.

Ebbinghaus, H.-D. and Flum, J. (1995): *Finite Model Theory*, Springer, Berlin.

Gurevich, Y. (1984): Toward logic tailored for computational complexity, in M. M. Richter et al. (eds.), *Computation and Proof Theory*, Lecture Notes in Math. 1104, Springer-Verlag, Berlin, pp. 175–216.

Grohe, M. (1996): Some remarks on finite Löwenheim–Skolem Theorems, *Math. Logic Quart.* **42**, 569–571.

Gödel, K. (1931): Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monats. Math. Phys.* **38**, 173–198.

Gödel, K. (1934): On undecidable propositions of formal mathematical systems, Mimeographed notes, by S. Kleene and J. Rosser, published in M. Davis (ed.), *The Undecidable*, Raven Press, 1965, pp. 41–81.

Machover, M. (1996): *Set Theory, Logic and their Limitations*, Cambridge University Press, Cambridge.

Matiyasevich, Y. (1970): Diofantovost' perechislimykh mnozhestv, *Dokl. Akad. Nauk SSSR* **191**(2), 297–282 (Russian). (English translation, "Enumerable sets are Diophantine", *Soviet Math. Dokl.* **11**(2) (1970), 354–358.)

Matiyasevich, Y. (1993): *Hilbert's Tenth Problem*, MIT Press, Cambridge, Mass.

Mostowski, A. (1962): Representability of sets in formal systems, *Proc. Symp. Pure Appl. Math.* **5**, 29–49.

Mostowski, A. (1965): *Thirty Years of Foudational Studies*, Acta Philosophica Fennica, Fasc. XVII, Societas Philosophica Fennica, Helsinki.

Murawski, R. (1998): Undefinability of truth. The problem of priority: Tarski vs. Gödel, *History and Philosophy of Logic* **19**, 153–160.

Smorynski, C. (1991): *Logical Number Theory I*, Springer-Verlag, Berlin.

Smullyan, R. (1961): *Theory of Formal Systems*, Princeton University Press, Princeton.

Smullyan, R. (1992): *Gödel's Incompleteness Theorems*, Oxford University Press, Oxford.

Soames, S. (1998): *Understanding Truth*, Oxford University Press, New York.

Tarski, A. (1933a): *Pojęcie prawdy w językach dedukcyjnych*, Nakladem Towarzystwa Naukowego Warszawskiego, Warszawa. (English translation: "The Concept of Truth in Formalized Languages", in A. Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* (edited and translated by J. H. Woodger), Clarendon Press, Oxford, 1956. Page references are to the English translation.)

Tarski, A. (1933b): Einige Betrachtungen über die Begriffe ω-Widerspruchsfreiheit und der ω-Vollständigkeit, *Monats. Math. Phys.* **40**, 97–112. (English translation: "Some observations on the concepts of ω-consistency and ω-completeness", in A. Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* (edited and translated by J. H. Woodger), Clarendon Press, Oxford, 1956. Page references are to the English translation.)

Tarski, A. (1969): Truth and proof, *Scientific American* **220**(6), 63–77.

Tarski, A., Mostowski, A. and Robinson, R. M. (1953): *Undecidable Theories*, North-Holland, Amsterdam.

Trakhtenbrot, B. A. (1950): Nevozmoznost' algorifma dla problemy razresimosti na konecnyh klassah, ("Impossibility of an algorithm for the decision problem in finite classes"), *Dokl. Akad. Nauk SSSR* **70**, 569–572.

Department of Philosophy
P.O.Box 24
SF-00014 University of Helsinki
Finland
E-mail: panu.raatikainen@helsinki.fi