CONSISTENCY AND THE THEORY OF TRUTH

RICHARD G HECK JR.

ABSTRACT. What is the logical strength of theories of truth? That is: If you take a theory $\mathcal{T}$ and add a theory of truth to it, how strong is the resulting theory, as compared to $\mathcal{T}$? Once the question has been properly formulated, the answer turns out to be about as elegant as one could want: At least when $\mathcal{T}$ is finitely axiomatized theory, theories of truth act more or less as a kind of abstract consistency statement. To prove this result, however, we have to formulate truth-theories somewhat differently from how they have been and instead follow Tarski in 'disentangling' syntactic theories from object theories.

1. MOTIVATIONAL REMARKS

Tarski's classic paper "The Concept of Truth in Formalized Languages" is nicely representative of the state of logic in the 1930s: It is as much about what one cannot do as it is about what one can do. On the negative (or 'limitative') side, we have Tarski's celebrated theorem on the indefinability of truth. On the positive (or 'constructive') side, we have Tarski's demonstration that, for a wide range of theories $\mathcal{T}$, it is possible to add a theory of truth to $\mathcal{T}$ in such a way that the resulting theory is not only consistent (if $\mathcal{T}$ is) but also fruitful: Within it, we can prove the sorts of meta-mathematical results for which the notion of truth was then already being used. In particular, if we add a theory of truth to Peano arithmetic, PA—if, that is, we add axioms like "A conjunction is true iff both its conjuncts are true", and so forth—then we will be able to prove that PA is consistent by the following sort of argument: The axioms are all true; the rules of inference preserve truth; hence every theorem of PA is true; but some sentences, such as '$0 = 1$', are false; so some sentences are not theorems of PA; so PA is consistent.

My focus here will be on this positive part of Tarski's contribution, and the central question is what, precisely, a theory of truth buys us in terms of logical strength.[1] Since PA plus a truth-theory proves that PA is consistent, it follows from Gödel's second incompleteness theorem that the former is stronger than the latter. The same will be true of many

[1] Some of the results to be proven also have some historical significance, as relating to Tarski's infamous claim that the meta-language must be "essentially richer" than the object language (Tarski, 1944, §10). I discuss those results in a different paper (Heck, 2014c).

other theories $\mathcal{T}$. But can we say just how strong the resulting theory is, as compared to the original theory $\mathcal{T}$?

The question comes in another form, as well. As is well known, Tarski's argument for the consistency of PA depends not just upon the availability of a theory of truth but also upon our extending the induction axioms beyond those of PA to permit semantic vocabulary. If we do not allow 'semantic' induction, then the resulting theory is not only a conservative extension of PA[2] but, as we shall see, is interpretable in PA. That might well seem to suggest that truth-theories, on their own, have no logical strength whatsoever, but that the strength is provided by the new instances of induction.

It will emerge below, however, that PA is, in several respects, a very special case. What does or doesn't happen when we add a truth-theory to PA is not uninteresting, of course, but it is often very different from what happens when we add one to some other theory, in particular, to a finitely axiomatized theory. And it seems to me that, if we are interested in questions about the logical strength of theories of truth, then the right question to ask is not "What happens when you add a truth-theory to PA?" but: What happens when we add a truth-theory to an arbitrary theory $\mathcal{T}$?

It will turn out that theories of truth act much like abstract consistency statements, in a sense that is sufficiently captured by the following sort of result, specific versions of which will be proven below.

Theorem. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language $\mathcal{L}$. Then the result of adding to $\mathcal{T}$ a fully compositional theory of truth for $\mathcal{L}$, based upon some underlying theory of syntax $\mathcal{U}$, is equivalent to* $\mathcal{U} + \mathsf{Con}(\mathcal{T})$.

One might therefore think of a theory of truth for $\mathcal{L}$ as a kind of operator on theories: If you hand it a finitely axiomatized theory $\mathcal{T}$ in $\mathcal{L}$, then it hands you back a theory of the same logical strength as the underlying syntactic theory plus the statement that $\mathcal{T}$ is consistent. And this theory, as we shall see, is always stronger than $\mathcal{T}$.

It will take some effort to find the right framework in which to state and prove this kind of result, and many things will need to be clarified, e.g., what the right notion of 'equivalence' is. Most work on theories of truth, whether in the axiomatic or model-theorietic tradition, begins with a 'base theory', usually taken to be PA, to which a self-referential theory of truth is then added, with syntax being done via coding. Our initial investigations will be within this sort of framework, but we will see that it does not allow for a nice answer to the questions we want to study. The reason is that the 'base theory', in such treatments, plays two

[2] Model-theoretic proofs of this result have been available for some time (Kotlarski et al., 1981). The proof-theoretic argument given by Halbach (2011) contained a large lacuna, but it has since been filled by Leigh (2013).

very different roles, acting both as object theory and as syntactic theory, and it is impossible to strengthen one of these without simultaneously strengthening the other, sometimes in unexpected ways. What we need to do, then, is to 'disentangle the meta-theory from the object theory', as I shall put it, so that these can be varied independently. This idea, which revives Tarski's original approach. is perhaps the most important in this paper.

The plan for the paper is as follows. In an effort to make the discussion as accessible as possible, I will review, in Section 2, some of the central concepts from logic that we shall be using. Readers familiar with that material can skip to Section 3, where we'll discuss the usual way of 'adding a truth-theory' and prove some preliminary results. We will also discover some limitations of the approach we will have been pursuing to that point. Section 4 explores a different way of 'adding a truth-theory', the one that involves 'disentanglement'. That will allow us to prove all the nice results we had proven before plus more. I will close in by comparing the results proven here to some related results due to Visser.

2. Logical Preliminaries

In this section, I'll review some of the logical machinery to be used below. In Section 2.1, I'll make a few remarks about terminology. In Section 2.2, I introduce the notion of interpretability; Section 2.3 discusses the fragments of arithmetic in which we will be interested; Section 2.4 introduces the notion of a cut that is so fundamental to the study of weak fragments of PA and states some of the basic results about cuts; Section 2.5 presents a wonderful form of the second incompleteness theorem due to Pudlák.

Readers familiar with what is covered in the various sub-sections should be able to skip them.

2.1. **Languages and Theories.**

The *languages* in which we'll be interested here are first-order languages, constructed from primitive expressions—terms, function-symbols, and predicates of one or more places—in the usual way. These languages will also be finite, in the sense that they have only finitely many primitives. It is convenient to identify a language with the set of its primitives, together with some indication of their logical type, that is, with what is sometimes called the 'signature' of the language.

A *theory* here will always have a recursive set of axioms, and, following Feferman (1960), we understand the notion in an intensional sense: A theory is not a set of axioms but a 'presentation' of a set of axioms. When a theory has only finitely many axioms, the distinction between intensional and extensional conceptions all but lapses, since there is an obviously best way of specifying the axioms: as a list, i.e., as a disjunction:

$x = \ulcorner A_1 \urcorner \vee \cdots \vee x = \ulcorner A_n \urcorner$.[3] But the distinction does matter, in general, and it will matter at certain points below.[4]

A theory is 'stated in' a language.

2.2. **Interpretability.** There are a number of ways of comparing the logical strength of theories. If the theories are stated in the same language, then the obvious question is whether one proves all the results the other proves. Comparison is more difficult when the theories are stated in different languages. In that case, the theories will trivially prove different theorems: If A is in the language of the one but not of the other, then $\ulcorner A \vee \neg A \urcorner$ will be a theorem of the one but not of the other; this is true even if the (non-logical) axioms of the two theories are the same.[5]

If the language of one theory contains that of the other, then one way to compare them is to ask whether the first is a 'conservative extension' of the second, that is, whether the theory in the *extended* language proves any new theorems that can be stated in the *original* language. But even this fails if the theories are not so related. In that case, the usual method of comparison uses the notion of interpretation.

Let theories $\mathcal{B}$ (for 'base') and $\mathcal{T}$ (for 'target') be given, stated in languages $\mathcal{L}_{\mathcal{B}}$ and $\mathcal{L}_{\mathcal{T}}$, respectively. A *relative interpretation*[6] of $\mathcal{T}$ in $\mathcal{B}$ consists of two parts: a translation of $\mathcal{L}_{\mathcal{T}}$ into $\mathcal{L}_{\mathcal{B}}$, and proofs in $\mathcal{B}$ of the translations of the axioms of $\mathcal{T}$. The translation is compositional, in the sense that the only thing we actually need to do is define the (non-logical) primitive expressions of $\mathcal{L}_{\mathcal{T}}$ in terms of those of $\mathcal{L}_{\mathcal{B}}$ and specify a 'domain' for the interpretation in terms of a formula $\delta(x)$ of $\mathcal{L}_{\mathcal{B}}$. This can then be extended to a complete translation of $\mathcal{L}_{\mathcal{T}}$ into $\mathcal{L}_{B}$ in the obvious way, where quantifiers are 'relativized' to $\delta(x)$: $\forall x(\phi(x))$ is translated as: $\forall x(\delta(x) \rightarrow \phi^*(x))$, where $\phi^*(x)$ translates $\phi(x)$; $\exists x(\phi(x))$, as: $\exists x(\delta(x) \wedge \phi^*(x))$. As well as proofs of the translations of the axioms, we also need proofs of $\delta(t^*)$, for each primitive term t of $\mathcal{L}_{\mathcal{T}}$,[7] and of the

[3] Alternatively, one could conjoin the axioms and specify the theory by a single identity: $x = \ulcorner A_1 \wedge \cdots \wedge A_n \urcorner$.

[4] See, in particular, notes 42, 53, and 76.

[5] Here's an illustration of why this sort of point needs to be kept in mind. It's well-known that the sub-theory of PA that excludes the axioms for multiplication—so-called Pressburger arithmetic—is decidable. But this is so only if the multiplication *symbol* is also excluded. If it is included, then the theory is undecidable. In fact, every theory in the language of arithmetic that is compatible with Q is undecidable (Tarski, 1953, Theorem 6).

[6] In fact, there are several different notions of interpretation. We shall only need this one.

[7] It is convenient to allow terms and function-symbols to be translated using descriptions, which can then be eliminated as Russell taught. In that case, we need $\mathcal{B}$ to prove that the descriptions are proper.

closure condition

$$\forall x_1 \cdots x_n[\delta(x_1) \wedge \cdots \wedge \delta(x_n) \rightarrow \delta(f^*(x_1, \ldots, x_n))]$$

for each primitive function-symbol f, of however many places. We also need (if this isn't already covered) a proof that the domain is non-empty: $\exists x(\delta(x))$.

It follows that, if $\mathcal{B}$ is consistent, so is $\mathcal{T}$. If a contradiction could be derived from the axioms of $\mathcal{T}$, then that proof could be mimicked in $\mathcal{B}$: Just prove the translations of the axioms of $\mathcal{T}$ used in the proof of the contradiction, then append (a modified version of) the proof given in $\mathcal{T}$. Indeed, quite generally, if $\Sigma \vdash_{\mathcal{T}} A$, then $\Sigma^* \vdash_{\mathcal{B}} A^*$, where, again, the asterisk means: translation of. Moreover, if $\mathcal{B}$ and $\mathcal{T}$ are not too terribly weak,[8] then all of this will be provable in $\mathcal{B}$ and $\mathcal{T}$ themselves. So, in particular, $\mathcal{T}$ will prove $\mathsf{Con}(\mathcal{B}) \rightarrow \mathsf{Con}(\mathcal{T})$ and so cannot prove $\mathsf{Con}(\mathcal{B})$, though $\mathcal{B}$ might well prove $\mathsf{Con}(\mathcal{T})$.

Note that interpretability is transitive and reflexive and so is a preorder.

One way to give content to the idea that $\mathcal{B}$ is at least as strong as $\mathcal{T}$ is therefore to take it to mean: $\mathcal{T}$ is relatively interpretable in $\mathcal{B}$. That this really is a useful way to give content to the intuitive idea of relative strength emerged only after a good deal of hard work, beginning with Tarski, Mostowski, and Robinson (1953) and continuing through work by Feferman (1960) to the present day (e.g., Visser, 2006).

Though the notion of interpretation is particularly useful when we are dealing with theories stated in different languages, we can still ask whether $\mathcal{T}$ can be interpreted in $\mathcal{B}$ even when $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{\mathcal{B}}$ are the same: The interpretation of the primitives does not have to be the identity function. But of course it can be, and in that case the interpretation may take a very simple form, which we might call a *pure relativization*: The only substantial part of the interpretation is the relativization to a new domain. Many of the interpretations in which we shall be interested are of this kind.

Now, a couple definitions that apply (sensibly) only to non-finitely axiomatized theories $\mathcal{T}$.

Definition. $\mathcal{T}$ is said to be *locally interpretable* in $\mathcal{B}$ if every finite subset of $\mathcal{T}$ is interpretable in $\mathcal{B}$.

Local interpretability obviously follows from interpretability, which is also known as 'global' interpretability. The converse may fail. Local interpretability is also transitive and reflexive, and it relates to relative consistency just as global interpretability does: If $\mathcal{T}$ is locally interpretable in $\mathcal{B}$, then $\mathcal{T}$ is consistent if $\mathcal{B}$ is. The reason is that any proof of a contradiction in $\mathcal{T}$ will use only finitely many of $\mathcal{T}$'s axioms.

[8] Facts concerning interpretability can generally be verified in the theory known as $\mathsf{I}\Delta_0 + \Omega_1$, for which see below.

As said above, PA is going to turn out to be something of a special case. That is because PA is not only not finitely axiomatizable but is reflexive (Mostowski, 1952).[9]

Definition. $\mathcal{T}$ is *reflexive* if $\mathcal{T}$ proves the consistency of each of its finite sub-theories.

It is one of the central lessons of Feferman's classic paper "Arithmetization of Metamathematics in a General Setting" (Feferman, 1960) that PA's reflexivity can cause all sorts of unexpected phenomena as regards interpretability in PA. What will matter most to us here is the fact that reflexive theories collapse the distinction between local and global interpretability.

Theorem (Orey's Compactness Theorem)**.** *Suppose that $\mathcal{T}$ is locally interpretable in $\mathcal{B}$ and that $\mathcal{B}$ is reflexive. Then $\mathcal{T}$ is (globally) interpretable in $\mathcal{B}$.*

This result is due to Stephen Orey (1961)—hence the name—but it first appeared in the paper of Feferman's just mentioned, as Theorem 6.9.[10]

2.3. **Fragments of Arithmetic.** As mentioned earlier, we are going to be interested in the general question what happens when we add a truth-theory to some arbitrary theory $\mathcal{T}$. In practice, however, we shall mostly be concerned with PA and certain of its sub-theories.

Robinson arithmetic, or Q, is the theory whose axioms are the universal closures of the following eight formulae:

Q1 $\mathsf{S}x \neq 0$
Q2 $\mathsf{S}x = \mathsf{S}y \rightarrow x = y$
Q3 $x + 0 = x$
Q4 $x + \mathsf{S}y = \mathsf{S}(x + y)$
Q5 $x \times 0 = 0$
Q6 $x \times \mathsf{S}y = (x \times y) + x$
Q7 $x \neq 0 \rightarrow \exists y(x = \mathsf{S}y)$
Q8 $x < y \equiv \exists z(y = \mathsf{S}z + x)$

The last is often considered a definition of $<$; it is convenient in the present context to regard $<$ as just part of the language. The language

[9] A stonger notion of reflexivity restricts the complexity of proofs, as well. We shall not need that here.

[10] The version proved by Feferman and Orey is limited to reflexive theories $\mathcal{B}$ that extend PA. The more general theorem stated in the text results from later strengthenings, by various authors, of the ingredients used in the original proof, e.g., the arithmetical completeness theorem. Feferman (1960, Theorem 6.2) proves it only for PA. We shall later use a version of this result for $\mathsf{I}\Sigma_1$ (Theorem 4.15), and it can be proven even for $\mathsf{I}\Delta_0 + \Omega_1$ (Visser, 1991, §6) and then extended to Q by the method of cuts.

of Q, $\{0, \mathsf{S}, +, \times, <\}$, is what we shall call 'the language of arithmetic' and shall denote: $\mathcal{A}$. We regard $x \leq y$ as an abbreviation for: $x < y \vee x = y$.[11]

An important class of sub-theories of PA is characterized in terms of the induction axioms these theories contain. PA itself is Q plus the full induction scheme:

$$A(0) \wedge \forall x(A(x) \rightarrow A(\mathsf{S}x)) \rightarrow \forall x(A(x))$$

where $A(x)$ is any formula at all.[12]

A formula is said to be Δ_0 (a.k.a., Σ_0) if all quantifiers contained in it are 'bounded', that is, if all of its quantified subformulae are of the form $\forall x(x < t \rightarrow \cdots)$ or $\exists x(x < t \wedge \cdots)$, where t is a term. These are customarily abbreviated: $\forall x < t(\cdots)$ and $\exists x < t(\cdots)$. A formula is Σ_1 (resp., Π_1) if it is of the form $\exists v_1 \cdots \exists v_n(\phi)$ (resp., $\forall v_1 \cdots \forall v_n(\phi)$), where ϕ is Δ_0. A formula is Σ_n (resp., Π_n) if it is $\exists v_1 \cdots \exists v_n(\phi)$ (resp., $\forall v_1 \cdots \forall v_n(\phi)$), where ϕ is Π_{n-1} (resp., Σ_{n-1}).

The theory $\mathsf{I}\Theta$ is Q plus induction for formulae in the set Θ: So $A(x)$ has to be in Θ. Thus, $\mathsf{I}\Delta_0$ is Q plus induction for Δ_0 formulae, and $\mathsf{I}\Sigma_1$ is Q plus induction for Σ_1 formulae. $\mathsf{I}\Delta_0$ is in one sense clearly stronger than Q: It proves lots of important generalizations about the natural numbers that Q does not. But in another sense it is still a very weak theory: It is interpretable in Q.[13] Another respect in which $\mathsf{I}\Delta_0$ is weak is that, although one can define the relation $y = 2^x$ by means of a Δ_0 formula $\mathsf{exp}(x, y)$, we cannot prove in $\mathsf{I}\Delta_0$ that exponentiation is total; that is, we cannot prove: $\forall x \exists y(\mathsf{exp}(x, y))$. The obvious proof uses induction on $\exists y(\mathsf{exp}(x, y))$, which is Σ_1. But for that very reason, the totality of exponentiation is provable in $\mathsf{I}\Sigma_1$, as is the totality of every other primitive recursive function. So $\mathsf{I}\Sigma_1$ is much stronger than $\mathsf{I}\Delta_0$: Indeed, $\mathsf{I}\Sigma_1$ proves $\mathsf{Con}(\mathsf{I}\Delta_0)$.[14]

There is one final theory, known as $\mathsf{I}\Delta_0 + \Omega_1$, that we shall need. It extends $\mathsf{I}\Delta_0$ by asserting the totality of a certain function $\omega_1(x)$ that, like 2^x, is Δ_0-definable but not $\mathsf{I}\Delta_0$-provably total. The precise definition varies between authors, but one definition is:

$$\omega_1(x) = 2^{|x|^2}$$

where $|x|$ is the least y such that $2^{\mathsf{S}y} > \mathsf{S}x$ (Visser, 1991, p. 83). As said, the relation $y = \omega_1(x)$ can be defined by a Δ_0 formula $\Omega_1(x, y)$,

[11] Note that, in Q, we do not necessarily have: $x \leq y \equiv \exists z(y = z + x)$. The left to right direction is easy, but the other needs $Sz + x = z + Sx$. So we do have this equivalence in $\mathsf{I}\Delta_0$ and even in the theory known as I_{open}.

[12] It is customary then to drop Q7, which has a trivial inductive proof.

[13] That $\mathsf{I}\Delta_0$ is *locally* interpretable in Q was first proven by Nelson (1986). That it is globally interpretable was proven by Wilkie (Wilkie and Paris, 1987). The proof is discussed both by Hájek and Pudlák (1993, pp. 366–70) and by Burgess (2005, §2.2). The techniques used are those we shall discuss in Section 2.4.

[14] This is already provable in $\mathsf{I}\Delta_0 + \mathsf{superexp}$, in fact.

and $\mathsf{I}\Delta_0 + \Omega_1$ is then $\mathsf{I}\Delta_0$ plus the formula asserting that this relation is total: $\forall x \exists y(\Omega_1(x, y))$. The interest of this theory lies in the fact that it is, as Visser puts it, "just right for treating syntax".[15] And, like $\mathsf{I}\Delta_0$, it is interpretable in Q (Hájek and Pudlák, 1993, p. 367).

It will be important below that, although $\mathsf{I}\Sigma_n$ so described is not finitely axiomatized, there is a finite axiomatization to be had if $n > 0$ (Hájek and Pudlák, 1993, pp. 77ff). We shall assume such an axiomatization. It is not presently known whether $\mathsf{I}\Delta_0$ or $\mathsf{I}\Delta_0 + \Omega_1$ is finitely axiomatizable.

As we shall see later, it is sometimes extremely helpful if our language contains no terms other than variables. We shall therefore also want to use what we might call the language of *relational* arithmetic. This language contains predicate letters Z,[16] P, A, and M in place of 0, S, $+$, and $\times$. And, whereas, in the usual language of arithmetic, the totality and functionality of S, $+$, and $\times$ are truths of logic, here these facts are explicitly recorded as non-logical axioms:[17]

Z $\exists x(\mathsf{Z}x \wedge \forall y(\mathsf{Z}y \rightarrow x = y))$
P $\forall x \exists y(\mathsf{P}xy \wedge \forall z(\mathsf{P}xz \rightarrow y = z))$
A $\forall x \forall y \exists z(\mathsf{A}xyz \wedge \forall w(\mathsf{A}xyw \rightarrow z = w))$
M $\forall x \forall y \exists z(\mathsf{M}xyz \wedge \forall w(\mathsf{M}xyw \rightarrow z = w))$

It should be clear that theories in the usual language of arithmetic have natural correlates in the language of relational arithmetic. We can thus state a theory Q_R in this language, with much the same content as Q, by simply adapting the axioms of Q itself. The first four axioms, for example, would be:

QR1 $\neg \mathsf{P}x0$
QR2 $\mathsf{P}xz \wedge \mathsf{P}yz \rightarrow x = y$
QR3 $\mathsf{A}x0x$
QR4 $(\mathsf{P}yz \wedge \mathsf{A}xzu) \wedge (\mathsf{A}xyw \wedge \mathsf{P}wv) \rightarrow u = v$

The first two conjuncts of QR4 say, in effect, that $u = x + \mathsf{S}y$; the next two, that $v = \mathsf{S}(x + y)$.

[15] Wilkie and Paris (1987) seem to have been the first to recognize the importance of $\mathsf{I}\Delta_0 + \Omega_1$. One has to use a more "efficient" coding than is customary, however, to get things to work. Hajék and Pudlák (1993, pp. 303ff) give the details. We could also use, as Nicolai (2014) does, the theory S^1_2 introduced by Buss (1986), which would be an advantage in some ways, since it is finitely axiomatized but has the same interpretability strength as $\mathsf{I}\Delta_0 + \Omega_1$.

[16] For our purposes, it would be fine to allow a term for zero. It's function symbols that will cause problems below, since they give rise to complex terms. But it's smoother just to do away with terms that aren't variables.

[17] One does not *have* to assume such axioms. There are relational versions of Q that do not assume the functionality of P, A, and M and yet that are still essentially undecidable and, in fact, interpret Q itself (Hájek, 2007; Švejdar, 2007; Heck, 2014a).

It should be clear that Q and Q_R are interpretable in one another, in a very straightforward way.[18] Similar things can be said about relational versions of the other theories mentioned.

2.4. **The Method of Cuts.** The proof of the main result below uses a technique called 'shortening of cuts' that is due to Robert Solovay. Since this method is not widely known among philosophers, I shall spend some time introducing it.

Let $\mathcal{T}$ be an arithmetical theory that does not have full induction, in the sense that there are formulae with the form of induction axioms that are not theorems of $\mathcal{T}$. Then there will ordinarily be formulae $\phi(x)$ for which $\mathcal{T}$ proves the hypotheses of the relevant induction axiom, $\phi(0)$ and $\forall x(\phi(x) \rightarrow \phi(\mathsf{S}x))$, but for which $\mathcal{T}$ does *not* prove its conclusion: $\forall x(\phi(x))$.[19] Obviously, $\mathcal{T}$ will therefore prove $\phi(0)$, $\phi(1)$, $\phi(2)$, and so forth. So, from the point of view of $\mathcal{T}$, $\phi(x)$ is a formula that is true of 0, 1, 2, and so on, but that is, for all $\mathcal{T}$ knows, false of some natural numbers. And, by the completeness theorem, there will be models of $\mathcal{T}$ in which $\phi(x)$ is *not* true of all of the 'natural numbers'.

For example, as is well-known, Q does not prove that no number is its own successor. But Q does prove both $0 \neq \mathsf{S}0$, which follows immediately from the first axiom of Q, and $x \neq \mathsf{S}x \rightarrow \mathsf{S}x \neq \mathsf{SS}x$, which follows just as immediately from the second. So $x \neq \mathsf{S}x$ is the kind of formula Russell called 'inductive', and that terminology has been adapted to the present context.

Definition. A formula $\iota(x)$ is said to be *inductive* in $\mathcal{T}$ if

(1) $\mathcal{T} \vdash \iota(0)$
(2) $\mathcal{T} \vdash \forall x(\iota(x) \rightarrow \iota(\mathsf{S}x))$

And the important thing about inductive formulas, for our purposes, is that they can be used to construct relative interpretations. The crucial result is this one.

Theorem 2.1. *Let $\iota(x)$ be a formula that is inductive in $\mathcal{T} \supseteq \mathsf{Q}$ and that is no worse than Π_1. Then $\mathcal{T}$ interprets $\mathsf{Q} + \forall x(\iota(x))$.*

It's not essential for what follows that the reader understand the proof of Theorem 2.1. But the method used in its proof—the shortening of cuts—is essential for the work we shall be doing below, so it is worth having some sense for how it works in a simple case. I shall therefore explain the ideas behind the proof of Theorem 2.1 by continuing to discuss the

[18] Terms in the language of Q will be translated by definite descriptions, as mentioned in note 7. And the equivalence is actually much stronger: The theories are, in a precise technical sense, 'synonymous'.

[19] In the case of $\mathsf{I}\Sigma_n$, one can actually exhibit such formulae (Hájek and Pudlák, 1993, p. 172).

example already mentioned: We'll see how to prove that Q interprets $\mathsf{Q} + \forall x(x \neq \mathsf{S}x)$.

The basic idea is simply to restrict the domain to the numbers that satisfy $x \neq \mathsf{S}x$—which, one might say, might as well *be* the natural numbers, so far as Q is concerned. But that isn't quite right. The problem is that we do not, in general, know that the numbers satisfying an inductive formula constitute an initial segment of all the numbers there are. The standard numbers will all satisfy $\iota(x)$, but then there may be some that don't and then some more that do after the ones that don't. So if we want a formula that might play the role of a 'new domain', then we need a slightly different notion, the notion of a *cut*.

Definition. A formula $\iota(x)$ is a *cut* in a theory $\mathcal{T}$ if $\iota(x)$ is inductive in $\mathcal{T}$ and is $\mathcal{T}$-provably closed downwards, i.e.:

$$\mathcal{T} \vdash \forall x[\iota(x) \to \forall y < x(\iota(y))]$$

If $\mathcal{T}$ does not prove $\forall x(\iota(x))$, then $\iota(x)$ is said to be a *proper cut* in $\mathcal{T}$.

The numbers satisfying a formula that is a cut in $\mathcal{T}$ will constitute an initial segment of $\mathcal{T}$'s natural numbers, and if the cut is proper, there will be models in which they constitute a proper initial segment.

The key result relating inductive formulas and cuts is this one.

Lemma 2.2 (Hájek and Pudlák 1993, pp. 368–9)**.** *Let $\iota(x)$ be inductive in $\mathcal{T} \supseteq \mathsf{Q}$. Then there is a cut $\kappa(x)$ in $\mathcal{T}$ for which $\mathcal{T} \vdash \forall x(\kappa(x) \to \iota(x))$. That is: Every $\mathcal{T}$-inductive formula can be shortened to a $\mathcal{T}$-cut.*

Proof. The obvious idea is to consider $\forall y \leq x(\iota(x))$ and to show that it defines a cut. Unfortunately, this doesn't quite work. The problem is that the proof that the formula in question defines a cut needs the transitivity of $\leq$, and Q does not prove that $\leq$ is transitive.[20]

This obstacle can be overcome, however, and the way in which this is done is a nice illustration of how the shortening of cuts works: We can simply restrict our attention to numbers for which $\leq$ *is* transitive. Consider the formula:

$$\chi(x) \overset{df}{\equiv} \iota(x) \wedge \forall y \forall z(y \leq x \wedge z \leq y \to z \leq x)$$

$\chi(x)$ says, more or less, that x satisfies $\iota(x)$ and that $\leq$ is transitive below x. It can be proven, in Q, that, if $\iota(x)$ is inductive, then so is $\chi(x)$. The proof uses the following theorems of Q:

(i) $x \leq 0 \to x = 0$
(ii) $x \leq y \equiv \mathsf{S}x \leq \mathsf{S}y$
(iii) $0 \leq x$

We get $\chi(0)$ from $\iota(0)$ and (i). So suppose $\chi(a)$. We want to show that $\chi(\mathsf{S}a)$. Certainly $\iota(\mathsf{S}a)$, since $\iota(a)$, and $\iota(x)$ is inductive. So suppose $y \leq \mathsf{S}a$

[20] This can be shown by constructing a simple counter-model.

and $z \leq y$. We want to show that $z \leq \mathsf{S}a$. If $z = 0$, then $0 \leq \mathsf{S}a$, by (iii). So suppose $z \neq 0$. Then $z = \mathsf{S}b$, for some b, and since $z \leq y$ and $z \neq 0$, we have $y \neq 0$, by (i), so $y = \mathsf{S}c$, for some c. Hence, $\mathsf{S}b \leq \mathsf{S}c$, so by (ii), $b \leq c$. Moreover, $\mathsf{S}c \leq \mathsf{S}a$, so $c \leq a$. By the induction hypothesis, then, $b \leq a$, so, by (ii) again, $\mathsf{S}b \leq \mathsf{S}a$, i.e., $z \leq \mathsf{S}a$, and we are done.

We can then pursue the original idea, but with $\chi(x)$ in place of $\iota(x)$:

$$\kappa(x) \overset{df}{\equiv} \forall w \leq x(\chi(w))$$

The verification that this defines a cut is left to the reader. □

So, although Q can't prove that $x \neq \mathsf{S}x$ is a cut, there is a 'subcut' $\kappa(x)$ of $x \neq \mathsf{S}x$ in Q. So we might now try simply restricting attention to $\kappa(x)$, the thought being that this will give us an interpretation in which $x \neq \mathsf{S}x$ holds and in which the axioms of Q just keep right on holding. But this doesn't quite work, either, the reason being that we need to ensure that the domain of our interpretation is closed under the operations of succession, addition, and multiplication. That it is closed under S follows from the fact that $\kappa(x)$ is inductive. But we have no reason at this point to think we can prove either of these:

$$\forall x \forall y(\kappa(x) \wedge \kappa(y) \rightarrow \kappa(x+y))$$
$$\forall x \forall y(\kappa(x) \wedge \kappa(y) \rightarrow \kappa(x \times y))$$

What to do?

The answer is to use the method of shortening cuts to restrict attention to numbers that *do* have sums and products inside the cut we are defining, much as we just restricted attention to the numbers for which $\leq$ is transitive.[21] Doing so allows us to prove the following.

Lemma 2.3. *If $\mathcal{T} \supseteq \mathsf{Q}$, then every $\mathcal{T}$-inductive formula $\iota(x)$ can be shortened to a $\mathcal{T}$-cut $\kappa(x)$ on which $\mathcal{T}$ proves the relativizations of the axioms of* Q.

That is, $\mathcal{T}$ will prove the relativization of Q4:

$$\forall x[\kappa(x) \rightarrow \forall y(\kappa(y) \rightarrow \mathsf{S}(x+y) = x + \mathsf{S}y)]$$

and similarly for the other axioms.

We can now see how to prove that Q interprets $\mathsf{Q} + \forall x(x \neq \mathsf{S}x)$. Since $x \neq \mathsf{S}x$ is inductive in Q, there is, by Lemma 2.3, a subcut $\kappa(x)$ of $x \neq \mathsf{S}x$ on which Q proves the relativizations of the axioms of Q. So, if we take as our interpretation the 'pure relativization' to $\kappa(x)$, that gives us an interpretation of the axioms of Q in $\mathcal{T}$. So we need only show that Q proves

$$\forall x(\kappa(x) \rightarrow x \neq \mathsf{S}x)$$

But of course it does, since that says, precisely, that $\kappa(x)$ is a subcut of $x \neq \mathsf{S}x$.

[21] Burgess (2005, §2.2) gives an accessible treatment of this part of the construction.

Similar reasoning shows that, if $\iota(x)$ is any $\mathcal{T}$-inductive, quantifier-free formula, then $\mathcal{T}$ interprets $\mathsf{Q} + \forall x(\iota(x))$. To prove Theorem 2.1, we need only extend from there to the case of Δ_0 formulae,[22] and from there to the case of Π_1 formulae.[23] But it is *not* in general true that, if $\iota(x)$ is a Σ_1 cut in $\mathcal{T} \supseteq \mathsf{Q}$, then $\mathcal{T}$ interprets $\mathsf{Q} + \forall x(\iota(x))$. The standard counterexample is $\exists y(\mathsf{exp}(x, y))$. This is inductive even in Q, but Q does not interpret $\mathsf{Q} + \forall x \exists y(\mathsf{exp}(x, y))$ (Hájek and Pudlák, 1993, p. 391).

As it happens, shortening of cuts can be used to prove stronger forms of Lemma 2.3 and so of Theorem 2.1.

Lemma 2.4. *If $\mathcal{T} \supseteq \mathsf{Q}$, then every $\mathcal{T}$-inductive formula $\iota(x)$ can be shortened to a $\mathcal{T}$-cut $\kappa(x)$ on which $\mathcal{T}$ proves the relativizations of the axioms of $\mathsf{I}\Delta_0 + \Omega_1$.*

Corollary 2.5. *Let $\iota(x)$ be a formula that is inductive in $\mathcal{T} \supseteq \mathsf{Q}$ and that is no worse than Π_1. Then $\mathcal{T}$ interprets $\mathsf{I}\Delta_0 + \Omega_1 + \forall x(\iota(x))$.*

We'll need Lemma 2.4 below.

2.5. **The Unprovability of 'Small' Consistency.** Gödel's second incompleteness theorem tells us that no 'sufficiently strong' and 'sufficiently expressive' theory proves its own consistency. We will need here a beautiful strengthening of Gödel's result that was proved in the mid-1980s by Pudlák. If we think of the numbers satisfying a cut as 'small' numbers,[24] then what Pudlák's result says is that no theory containing Q can even prove that there are no 'small' proofs of contradictions from its axioms. More formally, what Pudlák's theorem says is that no theory containing Q proves its own consistency on a cut.[25]

[22] The basic point is that, if $\iota(x)$ is Δ_0 and $\kappa(x)$ is a cut in $\mathcal{T}$, then $\forall x(\kappa(x) \rightarrow \iota(x))$ is going to be $\mathcal{T}$-provably equivalent to $\forall x(\kappa(x) \rightarrow \iota^{\kappa}(x))$, where $\iota^{\kappa}(x)$ is the relativization of $\iota(x)$ to $\kappa(x)$. The argument is by induction on the complexity of expressions and is straightforward. Burgess (2005, pp. 101–4), again, gives an accessible treatment.

[23] Since the extension to Π_1 formulae is not often stated, it is worth sketching the proof. Suppose that $\iota(x)$ is Π_1, say $\forall y(\phi(x, y))$, where $\phi(x, y)$ is Δ_0. Then what we need to show is that $\mathcal{T}$ proves

$$\forall x[\kappa(x) \rightarrow \forall y(\kappa(y) \rightarrow \phi^{\kappa}(x, y))]$$

As mentioned in the previous note, $\mathcal{T}$ will prove

$$\forall y(\kappa(y) \rightarrow \phi(x, y)) \equiv \forall y(\kappa(y) \rightarrow \phi^{\kappa}(x, y)),$$

so we need only show that $\mathcal{T}$ proves

$$\forall x[\kappa(x) \rightarrow \forall y(\kappa(y) \rightarrow \phi(x, y))].$$

But we already know that $\mathcal{T}$ proves the stronger: $\forall x[\kappa(x) \rightarrow \forall y(\phi(x, y))]$, since $\kappa(x)$ is a subcut of $\forall y(\phi(x, y))$ in $\mathcal{T}$.

[24] If it sounds as if there are connections here with Wang's Paradox, there are.

[25] Here, $\mathsf{Bew}_{\mathcal{T}}(x, y)$ is an appropriate (i.e., intensionally correct) formalization of 'x is a $\mathcal{T}$-proof of y'.

Theorem 2.6 (Pudlák 1985, Theorem 2.1)**.** *Suppose $\mathcal{T} \supseteq \mathsf{Q}$ is consistent, and let $\kappa(x)$ be a $\mathcal{T}$-cut. Then $\mathcal{T}$ does not prove:*

$$\forall x(\kappa(x) \rightarrow \neg \mathsf{Bew}_{\mathcal{T}}(x, \ulcorner 0 = S0 \urcorner))$$

This is a substantial strengthening of Gödel's result, in three respects. First, the usual form of the second incompleteness theorem applies only to theories containing enough induction to prove the Hilbert-Bernays-Gödel-Löb derivability conditions. Pudlák's version, by contrast, applies to any theory containing Q, which certainly does not prove the derivability conditions.[26] Second, Gödel's result tells us only that $\mathcal{T}$ cannot show that there are *no* proofs of contradictions, and this is compatible with $\mathcal{T}$'s being able to show that there are no 'small' proofs of contradictions.

The third respect in which Pudlák's result is an improvement emerges from the following consequence of Theorem 2.6.[27]

Theorem 2.7 (Pudlák 1985, Corollary 3.5)**.** *Suppose $\mathcal{T}$ is finitely axiomatized, sequential,[28] and consistent. Then $\mathcal{T}$ does not interpret $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

Whereas Gödel tells us that a (sufficiently strong and expressive) consistent theory $\mathcal{T}$ cannot *prove* $\mathsf{Con}(\mathcal{T})$, Pudlák tells us that, if $\mathcal{T}$ is finitely axiomatized and sequential, it cannot even *interpret* $\mathsf{Q}+\mathsf{Con}(\mathcal{T})$, let alone interpret $\mathcal{T} + \mathsf{Con}(\mathcal{T})$, let alone prove $\mathsf{Con}(\mathcal{T})$.[29]

The proofs of these two results are (well) beyond the scope of the present discussion.[30]

Proposition 2.8. *Suppose $\mathcal{S} \supseteq \mathsf{Q}$ proves $\mathsf{Con}(\mathcal{T})$ on a cut. Then $\mathcal{S}$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and even $\mathsf{I}\Delta_0 + \Omega_1 + \mathsf{Con}(\mathcal{T})$.*

Proof. To say that $\mathcal{S}$ proves $\mathsf{Con}(\mathcal{T})$ on a cut is to say that there is an $\mathcal{S}$-cut $\kappa(x)$ such that $\mathcal{S}$ proves $\forall x(\kappa(x) \rightarrow \neg \mathsf{Bew}_{\mathcal{T}}(x, \ulcorner 0 = 1 \urcorner))$. That cut can be shortened to one on which the axioms of $\mathsf{I}\Delta_0 + \Omega_1$ are available. Relativizing to that cut then gives us an interpretation of $\mathsf{I}\Delta_0 + \Omega_1 + \mathsf{Con}(\mathcal{T})$. □

[26] Wilkie and Paris (1987) show, however, that, with appropriate formulations, $\mathsf{I}\Delta_0+\Omega_1$ will do so.

[27] Although Q is not sequential, it does still follow from this result that Q does not interpret $\mathsf{Q} + \mathsf{Con}(\mathsf{Q})$. The reason is that, in $\mathsf{Q} + \mathsf{Con}(\mathsf{Q})$, we can construct a cut on which the axioms of Buss's S^1_2 are true; the relativization of $\mathsf{Con}(\mathsf{Q})$ to this cut is then trivially provable. So $\mathsf{Q} + \mathsf{Con}(\mathsf{Q})$ interprets $\mathsf{S}^1_2 + \mathsf{Con}(\mathsf{Q})$. So if Q interpreted $\mathsf{Q} + \mathsf{Con}(\mathsf{Q})$, it would also interpret $\mathsf{S}^1_2 + \mathsf{Con}(\mathsf{Q})$. But S^1_2 contains Q, so it would interpret $\mathsf{S}^1_2 + \mathsf{Con}(\mathsf{Q})$. Since S^1_2 is finitely axiomatized, however, that contradicts Theorem 2.7.

[28] See Section 3.1 for the definition of a sequential theory.

[29] Feferman (1960, p. 76, theorem 6.5) proved an antecedent of Pudlák's result: If $\mathsf{PA} \subseteq \mathcal{T}$, then $\mathcal{T}$ does not interpret $\mathcal{T} + \mathsf{Con}(\mathcal{T})$, assuming that the axioms of $\mathcal{T}$ are represented by a Σ_1 formula.

[30] As well as the paper of Pudlák's already cited, the interested reader may consult Hajék and Pudlák (1993, pp. 173ff) and Visser (2009a).

Putting these together, we have:[31]

Corollary 2.9. *Suppose that $\mathcal{T} \supseteq \mathsf{Q}$ is a consistent, finitely axiomatized, sequential theory, and that $\mathcal{S}$ proves $\mathsf{Con}(\mathcal{T})$ on a cut. Then $\mathcal{S}$ is not interpretable in $\mathcal{T}$.*

Proof. If $\mathcal{S}$ proves the consistency of $\mathcal{T}$ on a cut, then by Proposition 2.8 it will interpret $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. But if $\mathcal{S}$ were interpretable in $\mathcal{T}$, then $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ would be interpretable in $\mathcal{T}$, contradicting Theorem 2.7. □

It is Corollary 2.9 that will do much of the work below.

3. The Strength of Truth-theories

We are interested here in what happens when we 'add a theory of truth' to some object theory $\mathcal{T}$. What do we mean by adding a theory of truth? In this section, we will approach this question in the usual sort of way, where we begin with a certain 'base theory' $\mathcal{T}$, use $\mathcal{T}$ to formalize its own syntax, and then add various sorts of semantic axioms. It is going to turn out that, for our purposes, this is not the best approach. But to see that, we need to try it.

3.1. Formalizing Compositional Truth-theories.

Since the semantic axioms for the quantifiers, as Tarski bequeathed them to us, make use of sequences of elements from the domain, we shall need a nice theory of sequences if we're to formalize theories of truth. Technically, we'll need our base theory to be *sequential*.

Definition. Let $\mathcal{T}$ be a theory that contains Q, either straightforwardly or by interpretation. $\mathcal{T}$ is said to be *sequential* if, in short, it can code finite sequences of its elements. More precisely, $\mathcal{T}$ is sequential if there are formulae $\mathsf{seq}(s)$, $\mathsf{lh}(s,h)$ and $\mathsf{val}(s,n,x)$ for which $\mathcal{T}$ proves:

$$\forall s(\mathsf{seq}(s) \to \exists n(\mathsf{lh}(s,n)))$$
$$\exists s(\mathsf{seq}(s) \land \mathsf{lh}(s,0))$$
$$\forall s \forall n\{\mathsf{seq}(s) \land \mathsf{lh}(s,n) \to \forall m < n \exists x(\mathsf{val}(s,m,x))\}$$
$$\forall s \forall n\{\mathsf{seq}(s) \land \mathsf{lh}(s,n) \to \forall y \exists t[\mathsf{seq}(t) \land \mathsf{lh}(t,\mathsf{S}n) \land \mathsf{val}(t,n,y) \land \forall z \forall k < n(\mathsf{val}(s,k,z) \equiv \mathsf{val}(t,k,z))]\}$$

Here, $\mathsf{lh}(s,n)$ means: s is a sequence of length n; $\mathsf{val}(s,n,x)$ means: the $(n+1)$-st element of s is x. So the third principle says that every sequence of length n has an element at each position below n; the fourth, that each sequence can be extended by appending an arbitrary element of the domain; the second assures us that there is a 'null' sequence with

[31] Note that we do not need the hypothesis that $\mathcal{S}$ is consistent, since no consistent theory containing Q falsely proves the consistency of any (axiomatizable) theory. That is because Q already proves the inconsistency of every inconsistent (axiomatizable) theory.

which we can begin. We shall use '<>' as a term denoting one of[32] the null sequences whose existence is so guaranteed.

Q is not sequential, but there are lots of sequential theories that are interpretable in Q. For example, $\mathsf{I}\Delta_0$ is sequential, and it is interpretable in Q. More importantly, for our purposes, we can simply take the formulae $\mathsf{seq}(s)$, $\mathsf{lh}(s, h)$ and $\mathsf{val}(s, n, x)$ that make $\mathsf{I}\Delta_0$ sequential and add the principles that characterize sequential theories to Q as new axioms. This theory, which we might call $\mathsf{Q}_{\mathsf{seq}}$, is interpretable in Q, since it is obviously intepretable in any sequential theory. This fact will allow us to extend our main results to Q, even though they do not apply to Q directly. Note, morever, that every sequential theory interprets Q, so every such theory can 'do syntax' in the same sense in which Q can.[33]

It should be obvious that we can easily allow $\mathsf{val}(s, n, x)$ to have some fixed value, say, 0, if n is beyond the length of s. That is: A theory that contained an axiom to that effect would trivially be interpretable in one that did not. So we shall assume this principle, as well, since it allows us to pretend that our sequences are infinite, which simplifies matters considerably.

Although, officially, $\mathsf{lh}(s, h)$ and $\mathsf{val}(s, n, x)$ are relations, I'll also use the notation $\mathsf{lh}(s)$ and $\mathsf{val}(s, n)$ from time to time, in accord with Russell's theory of descriptions. The descriptions will always be (provably) proper in the cases that matter, and it makes many of the formulations much cleaner.

The theory of truth itself will consist of Tarski-style axioms for the logical and non-logical vocabulary. The axioms for the logical part of the language will always be the same:

(v) $\mathsf{Den}_\sigma(v_i, x) \equiv \mathsf{val}(\sigma, i, x)$, where v_i is the i^{th} variable

($=$) $\mathsf{Sat}_\sigma(\ulcorner t = u \urcorner) \equiv \exists x \exists y[\mathsf{Den}_\sigma(t, x) \wedge \mathsf{Den}_\sigma(u, y) \wedge x = y]$

($\neg$) $\mathsf{Sat}_\sigma(\ulcorner \neg A \urcorner) \equiv \neg \mathsf{Sat}_\sigma(A)$

($\wedge$) $\mathsf{Sat}_\sigma(\ulcorner A \wedge B \urcorner) \equiv \mathsf{Sat}_\sigma(A) \wedge \mathsf{Sat}_\sigma(B)$

($\forall$) $\mathsf{Sat}_\sigma(\ulcorner \forall v_i A(v_i) \urcorner) \equiv \forall \tau[\tau \overset{i}{\sim} \sigma \rightarrow \mathsf{Sat}_\sigma(\ulcorner A(v_i) \urcorner)]$

And similarly for the other logical constants.[34] Here, '$\mathsf{Den}_\sigma(t, x)$' means: t denotes x with respect to the sequence σ; '$\mathsf{Sat}_\sigma(A)$' means: σ satisfies A; and '$\tau \overset{i}{\sim} \sigma$' means that τ and σ agree on what they assign to each variable, with the possible exception of v_i, i.e.:

$$\forall k < \mathsf{lh}(\sigma)[k \neq i \rightarrow \forall x(\mathsf{val}(\sigma, k) \equiv \mathsf{val}(\tau, k)]$$

[32] Nothing in the defintion requires the theory of sequences to be extensional.

[33] Visser (2008) gives lots of details about sequential theories, including the facts mentioned here. Note that we shall also need to use such facts as that the code of a sequence is always greater than its length. We can always arrange for this sort of thing to be true.

[34] Of course, the other constants are definable in terms of the ones already mentioned, but, in the present context, this is not a particularly interesting or important fact.

In the case of the language of arithmetic, we'll also have these axioms for the non-logical constants:[35]

(0) $\mathsf{Den}_\sigma(\ulcorner 0 \urcorner, x) \equiv x = 0$

(S) $\mathsf{Den}_\sigma(\ulcorner \mathsf{S}t \urcorner, x) \equiv \exists y(\mathsf{Den}_\sigma(t, y) \wedge y = \mathsf{S}x)$

(+) $\mathsf{Den}_\sigma(\ulcorner t + u \urcorner, x) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge x = y + z]$

(×) $\mathsf{Den}_\sigma(\ulcorner t \times u \urcorner, x) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge x = y \times z]$

(<) $\mathsf{Sat}_\sigma(\ulcorner t < u \urcorner) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge y < z]$

The pattern should be clear.[36]

Finally, then, we need to define the notion of truth itself:

(T) $\mathrm{T}(A) \equiv A$ is a sentence $\wedge \forall \sigma(\mathsf{Sat}_\sigma(A))$

That is Tarski's definition: Truth is satisfaction by every sequence.[37]

So, that's what a theory of truth is. Here is some notation.[38]

Definition. Let $\mathcal{T}$ be sequential. Then $\mathsf{CT}^-[\mathcal{T}]$ is the theory that extends $\mathcal{T}$ by adding truth-theoretic axioms of the sort just discussed for the logical and non-logical vocabulary of the language of $\mathcal{T}$.

Note that $\mathsf{CT}^-[\mathcal{T}]$ does not extend any induction scheme that might be present in $\mathcal{T}$. There is no real chance, then, that $\mathsf{CT}^-[\mathcal{T}]$ is going to prove the consistency of $\mathcal{T}$. So one might suspect that $\mathsf{CT}^-[\mathcal{T}]$ would logically be no stronger than $\mathcal{T}$. If so, then, as we shall see in Section 3.2, one would suspect wrongly, at least in general.

[35] As is well known, denotation is actually definable in the language of arithmetic in such a way that the clauses involving it can be proven in PA and, in fact, in much weaker theories, so those clauses are often regarded as not really necessary. One can also forego the use of sequences and instead treat quantification substitutionally: $\forall v_i(A(v_i))$ is true iff, for each n, $A(\overline{n})$—the result of substituting the numeral for n for v_i—is true. But both these manoeuvers are specific to the language of arithmetic and are not available in general. Since we want our results to extend smoothly and naturally to other cases, such as the language of set theory, we will not use these shortcuts.

[36] It appears to have been Wang (1952) who first worked out the details of this sort of construction.

[37] Where we are discussing theories of truth over weak arithmetics, there is a worry about Tarski's definition, namely, that it 'hides' a quantifier in the definition of truth, so that elimination of that definition can make a formula in which $\mathrm{T}(x)$ occurs logically more complex after the elimination than it appeared to be. If we use Tarski's definition, for example, then $\mathsf{CT}^-[\mathcal{T}]$ (which will be defined shortly) will in many cases not prove: $\mathrm{T}(\ulcorner \neg A \urcorner) \equiv \neg \mathrm{T}(\ulcorner A \urcorner)$. The usual proof of this rests upon the fact that, if A is a sentence, then $\forall \sigma(\mathsf{Sat}_\sigma(A))$ iff $\exists \sigma(\mathsf{Sat}_\sigma(A))$, and that in turn is normally proven by an induction that is not formalizable in $\mathsf{CT}^-[\mathcal{T}]$. For this reason, it is sometimes preferable to use an alternate definition:

T: $\mathrm{T}(A) \equiv A$ is a sentence $\wedge \mathsf{Sat}_{<>}(A)$

on which truth is satisfaction by the null sequence.

Thanks to Cezary Cieśliński for bringing this issue to my attention. As it happens, however, it is actually better, for our purposes, to use Tarski's original definition, so we shall stick with it.

[38] Here, CT stands for: compositional truth.

First, however, let us note that $\mathsf{CT}^-[\mathcal{T}]$ is by no means a trivial extension of $\mathcal{T}$.

Lemma 3.1. $\mathsf{CT}^-[\mathcal{T}]$ *is a materially adequate, fully compositional theory of truth for the language of* $\mathcal{T}$*. In particular: For each formula* $A(v_1, \ldots, v_n)$ *in the language of* $\mathcal{T}$*,* $\mathsf{CT}^-[\mathcal{T}]$ *proves:*

$$\mathsf{Sat}_\sigma(\ulcorner A(v_1, \ldots, v_n)\urcorner) \equiv A(\mathsf{val}(\sigma, 1), \ldots \mathsf{val}(\sigma, n))$$

A fortiori, for each sentence A*,* $\mathsf{CT}^-[\mathcal{T}]$ *proves:* $\mathrm{T}(\ulcorner A\urcorner) \equiv A$*.*

Proof. A rigorous proof would be by induction on the complexity of sentences of $\mathcal{L}$, the induction here being 'external', not induction in $\mathcal{T}$ itself. But this should be fairly obvious.[39] A little experimentation will reveal that proofs of 'T-sentences' need no more than is available in $\mathsf{Q}_{\mathsf{seq}}$: We're not proving any general laws, just a bunch of particular facts, and Q is very good at proving particular facts, no matter how bad it may be at proving general laws. □

To put this differently: $\mathsf{CT}^-[\mathcal{T}]$ defines truth for sentences in the language of $\mathcal{T}$. Since $\mathcal{T}$ is sequential, it interprets Q, so we know from Tarski's indefinability theorem that $\mathcal{T}$ itself cannot define truth for all sentences in the language of $\mathcal{T}$ (assuming it is consistent). So $\mathsf{CT}^-[\mathcal{T}]$ is always expressively more powerful than $\mathcal{T}$.

Before we continue to explore $\mathsf{CT}^-[\mathcal{T}]$, let me state a couple of obvious corollaries of Lemma 3.1 that we shall need below.

Corollary 3.2. $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$ *is a materially adequate, fully compositional theory of truth for the language of arithmetic. So a fortiori is* $\mathsf{CT}^-[\mathcal{T}]$*, so long as* $\mathcal{T} \supseteq \mathsf{Q}_{\mathsf{seq}}$*.*

Since any theory in which a compositional theory of truth might be formulated has to be sequential, and every sequential theory inteprets Q, Corollary 3.2 is best possible: A compositional theory of truth for the language of arithmetic can be built upon the weakest possible basis.

Corollary 3.3. $\mathsf{CT}^-[\mathcal{T}]$ *proves, of each axiom of* $\mathcal{T}$*, that it is true.*

Proof. Let A be an axiom of $\mathcal{T}$. By Lemma 3.1, $\mathsf{CT}^-[\mathcal{T}]$ proves $\mathrm{T}(\ulcorner A\urcorner) \equiv A$. Since $\mathsf{CT}^-[\mathcal{T}]$ obviously proves A, it proves $\mathrm{T}(\ulcorner A\urcorner)$, too. □

The same, of course, goes for the theorems of $\mathcal{T}$,[40] but we shall not need that fact.

Corollary 3.4. *Let* $\mathcal{T}$ *be a finitely axiomatized sequential theory. Then* $\mathsf{CT}^-[\mathcal{T}]$ *proves the obvious, disjunctive formalization of "all axioms of* $\mathcal{T}$ *are true".*

[39] Leigh and Nicolai (2013, §3.1) give a detailed proof, though for the disentangled case, which is Lemma 4.1 below.

[40] I.e., $\mathsf{CT}^-[\mathcal{T}]$ proves, of *each* theorem of $\mathcal{T}$, that it is true.

Note the contrast with Corollary 3.3: If $\mathcal{T}$ is infinitely axiomatized, there is no reason whatsoever to suspect that $\mathsf{CT}^-[\mathcal{T}]$ will prove that *all* axioms of $\mathcal{T}$ are true, although it does prove that *each* axiom of $\mathcal{T}$ is true.

3.2. $\mathsf{CT}^-[\mathcal{T}]$ is Stronger than $\mathcal{T}$.

We are now ready to prove our first main result.[41]

Theorem 3.5. *Let $\mathcal{T} \supseteq \mathsf{I}\Delta_0 + \Omega_1$, and suppose that $\mathsf{CT}^-[\mathcal{T}]$ proves that all axioms of $\mathcal{T}$ are true.*[42] *Then $\mathsf{CT}^-[\mathcal{T}]$ proves the consistency of $\mathcal{T}$ on a cut and so is not interpretable in $\mathcal{T}$. Moreover, $\mathsf{CT}^-[\mathcal{T}]$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and even $\mathsf{I}\Delta_0 + \Omega_1 + \mathsf{Con}(\mathcal{T})$.*

The natural proof of this needs to use $\mathsf{I}\Delta_0 + \Omega_1$ because, as I said earlier, it is only here that we can do syntax naturally. We'll see later that this assumption can be weakened. I'll also limit discussion to arithmetical theories. The results will transfer naturally to theories in other sorts of languages. In that case, the various conditions will need to be stated in terms of interpretability.

The key to the proof of Theorem 3.5 is the realization that we can *almost* mimic the 'trivial' proof of the consistency of $\mathcal{T}$ that we learned from Tarski. That proof proceeds as follows: First, we show that all $\mathcal{T}$'s axioms are true; then we show that the rules of inference preserve truth; then we conclude, by induction, that all $\mathcal{T}$'s theorems are true. Since '$0 = \mathsf{S}0$' is not true, it isn't a theorem of $\mathcal{T}$, and so $\mathcal{T}$ is consistent.

This won't work in the present case, of course, because we do not have 'semantic induction', that is, induction for formulae containing semantic vocabulary. But we could overcome that lack by the method of cuts if we could show that the formalization of "n line proofs have true conclusions" is inductive. Then we would have that, although $\mathsf{CT}^-[\mathcal{T}]$ does not prove $\mathsf{Con}(\mathcal{T})$, it does prove it on a a subcut of the inductive formula just mentioned.

If that were the only obstacle, the proof would be easy. But there is another. It is an hypothesis of Theorem 3.5 that $\mathsf{CT}^-[\mathcal{T}]$ can prove that all of $\mathcal{T}$'s *non*-logical axioms are true. But, to mimic Tarski's proof,

[41] This result, and the others reported in this section, were taught to me by Albert Visser, who tells me he regards them as folklore. The proofs are my own, and the complications we shall meet arose as I tried to work out the details. Fischer (2009) also explores the interpretability of truth-theories in their respective base theories, but his investigations are limited to theories extending PA. In a more recent paper (Fischer, 2014), he uses such considerations to investigate whether having a truth-theory allows proofs to be 'sped-up'.

[42] Note that, if $\mathcal{T}$ is infinitely axiomatized, we will have to choose some specification of its axioms, *both* in order to formalize "all axioms of $\mathcal{T}$ are true" *and* to formalize $\mathsf{Con}(\mathcal{T})$. In Theorem 3.5, then, we are using the same specification both times. So if, as mentioned in footnote 53, we are using a 'funny' specification of the axioms of some theory $\mathcal{T}$ that allows $\mathcal{T}$ to prove that all its axioms, so specified, are true, then $\mathcal{T}$ will also prove the consistency on a cut of the theory whose axioms are so specified. That is why we need to think of theories intensionally.

we also need to prove that all of the *logical* axioms are true and that the rules of inference are truth-preserving. This turns out to be more difficult than one might have expected. It helps to assume that the logic in which we're working is formulated as a Hilbert-style proof system rather than a natural deduction system, with just two rules of inference: *modus ponens* and universal generalization. This allows us to speak simply in terms of the truth of the various lines of a proof, rather than in terms of whether the formula on a given line follows from the premises on which that line depends.[43]

The propositional axioms are easy enough.[44] Consider, for example, $p \to (q \to p)$ and reason in $\mathsf{CT}^-[\mathcal{T}]$. Let A and B be formulae. Using the clause for $\to$ twice, $\mathsf{Sat}_\sigma(\ulcorner A \to (B \to A)\urcorner)$ iff $\mathsf{Sat}_\sigma(A) \to (\mathsf{Sat}_\sigma(B) \to \mathsf{Sat}_\sigma(A))$. But the latter is of course a logical truth. So, generalizing, for any A and B, and for all σ, $\mathsf{Sat}_\sigma(\ulcorner A \to (B \to A)\urcorner)$, which is to say that all instances of $p \to (q \to p)$ are true.

The propositional rule, *modus ponens*, is also easy. What we need to show is that, if both A and $A \to B$ are satisfied by all sequences, then so is B. If $\forall\sigma(\mathsf{Sat}_\sigma(\ulcorner A \to B\urcorner))$, then, by the clause for the conditional: $\forall\sigma(\mathsf{Sat}_\sigma(A) \to \mathsf{Sat}_\sigma(B))$. But then, by logic: $\forall\sigma(\mathsf{Sat}_\sigma(A)) \to \forall\sigma(\mathsf{Sat}_\sigma(B))$. So, if $\forall\sigma(\mathsf{Sat}_\sigma(A))$, then $\forall\sigma(\mathsf{Sat}_\sigma(B))$.

Unfortunately, we run into problems with quantification. (Don't we always.) Consider universal instantiation, the simplest formulation of which is:

$$\forall v_i(\phi v_i) \to \phi v_j$$

subject to the usual restrictions. The argument for its truth proceeds as follows. Suppose some sequence σ does not satisfy some instance. Then, by the clause for $\to$, we have $\mathsf{Sat}_\sigma(\ulcorner \forall v_i(\phi v_i)\urcorner)$ and $\neg\mathsf{Sat}_\sigma(\ulcorner \phi v_j\urcorner)$. Now consider a sequence τ that is just like σ, except that what it assigns to v_i is whatever σ assigns to v_j. So $\tau \overset{i}{\sim} \sigma$, and hence $\mathsf{Sat}_\tau(\ulcorner \phi v_i\urcorner)$. But:

(i) v_i stands in ϕv_i only where v_j stands in ϕv_j
(ii) τ assigns v_i the same value that σ assigns v_j

So we must have $\neg\mathsf{Sat}_\tau(\ulcorner \phi v_i\urcorner)$, since $\neg\mathsf{Sat}_\sigma(\ulcorner \phi v_j\urcorner)$. Contradiction.

In making the last move in that argument, however, we were appealing to a general principle concerning 'variable-switching':

[43] The difficulty presented by a natural deduction system is that the correctness of a line then involves the consequent's being satisfied by a sequence if all the premises are, and this introduces more logical complexity than we have with the axiomatic treatment.

[44] Assuming we define truth as Tarski did, in terms of satisfaction by all sequences. If we use the alternate definition mentioned in note 37, and say that a line is true iff its universal closure is satisfied by <>, then we find ourselves needing to prove: $\forall\sigma(\mathsf{Sat}_\sigma(A)) \equiv \mathsf{Sat}_{<>}(\mathrm{ucl}(A))$. That only adds to our problems. This could probably be avoided, though, if we truncated sequences in the way suggested in note 45.

> If ϕv_j results from replacing all free occurrences of v_i in ϕv_i by v_j, and if τ is just like σ but sets $\tau_i = \sigma_j$, then $\mathsf{Sat}_\sigma(\ulcorner\phi v_j\urcorner)$ iff $\mathsf{Sat}_\tau(\ulcorner\phi v_i\urcorner)$.

There is clearly no hope of proving this without 'semantic' induction.

The problem is all the more serious if we allow instantiation not just by variables but by arbitrary terms and so formulate UI in the form:

$$\forall v_i(\phi v_i) \to \phi t$$

In that case, the proof also requires the claim that all terms denote. And we will have to face some form of this problem so long as our language does indeed contain terms that are not variables.

There are similar problems concerning universal generalization:

$$A \to \phi(v_i) \vdash A \to \forall v_i(\phi(v_i))$$

where of course A must not contain v_i free. Suppose that $A \to \forall v_i(\phi(v_i))$ is not satisfied by all sequences. Then there is a sequence σ such that $\mathsf{Sat}_\sigma(A)$ and $\neg\mathsf{Sat}_\sigma(\ulcorner\forall v_i\phi(v_i)\urcorner)$. By the clause for $\forall$, we have a sequence $\tau \overset{i}{\sim} \sigma$ such that $\neg\mathsf{Sat}_\tau(\ulcorner\phi(v_i)\urcorner)$. Since v_i is not free in A, then, we have $\mathsf{Sat}_\tau(A)$, as well. But how do we know that? Because whether a formula is satisfied by a sequence depends only upon what is assigned to variables that occur free in that formula, viz.:

$$\forall i[\text{free-in}(A, v_i) \to \mathsf{val}(\sigma, i) = \mathsf{val}(\tau, i)] \to \mathsf{Sat}_\sigma(A) \equiv \mathsf{Sat}_\tau(A)$$

But we will not be able to prove this without semantic induction.[45]

Careful examination of the proofs that the logical axioms are true, and that the rules are truth-preserving, shows that those proofs need the following semantic claims.

(1) If ϕt is the result of replacing all free occurrences of v_i in ϕv_i with t, and if $\mathsf{Den}_\sigma(t, a)$, $\tau \overset{i}{\sim} \sigma$, and $\mathsf{val}(\tau, i) = a$, then $\mathsf{Sat}_\sigma(\phi t)$ iff $\mathsf{Sat}_\tau(\phi v_i)$.

(2) If σ and τ agree on all free variables contained in A, then $\mathsf{Sat}_\sigma(A)$ iff $\mathsf{Sat}_\tau(A)$.

The proofs of these depend upon the corresponding claims for terms:

(3) If $u(t)$ is the result of replacing all occurrences of v_i in $u(v_i)$ with t, and if $\mathsf{Den}_\sigma(t, a)$, $\tau \overset{i}{\sim} \sigma$, and $\mathsf{val}(\tau, i) = a$, then $\mathsf{Den}_\sigma(u(t), a)$ iff $\mathsf{Den}_\tau(u(v_i), a)$.

(4) If σ and τ agree on all free variables contained in t, then $\mathsf{Den}_\sigma(t, a)$ iff $\mathsf{Den}_\tau(t, a)$.

We also need:

(5) For every term t, $\exists x(\mathsf{Den}_\sigma(t, x))$.

[45] As Visser pointed out to me, this particular issue can be avoided if we reformulate our truth-theory so that a sequence satisfies a formula only if it assigns values to all *and only* the variables free in that formula. This complicates the statement of the theory, however, and it does not help with our other problems.

This last will be trivial, however, if there are no terms in the language other than variables. The other claims are greatly simplified in that case, too, since t will always be a variable.

We thus have no hope whatsoever of showing that $\mathsf{CT}^{-}[\mathcal{T}]$ proves that 'logic is true', i.e., that the logical axioms are all true and that the rules of inference are truth-preserving. All is not lost, however, because we can use the method of cuts. The idea is to show that, though $\mathsf{CT}^{-}[\mathcal{T}]$ does not prove the listed semantic principles, it does prove their relativizations to some cut. Then it will follow that any formula that is on the cut and is an instance of a logical axiom is true, and that any rule of inference involving only formulae on the cut will be truth-preserving. And that will allow us to show that there can be no $\mathcal{T}$-proof of a contradiction on that cut.

Consider, for example:

(1*) For all σ and τ, if ϕt is of complexity $< n$ and is the result of replacing all occurrences of v_i in ϕv_i with t, and if $\mathsf{Den}_\sigma(t, a)$, $\tau \overset{i}{\sim} \sigma$, and $\mathsf{val}(\tau, i) = a$, then $\mathsf{Sat}_\sigma(\ulcorner \phi t \urcorner)$ iff $\mathsf{Sat}_\tau(\ulcorner \phi v_i \urcorner)$.

The usual proof of (1) can be adapted to show that (1*) is inductive with respect to n. The usual proofs of (2)–(4) can similarly be adapted to show that their 'starred versions' are inductive. The case of (5) is more complicated, however. The corresponding inductive formula is:

(5*) If t is of complexity $< n$, then $\exists x(\mathsf{Den}_\sigma(t, x))$.

In the case of the language of arithmetic, this will certainly be inductive. But if we were to add expressions to the language for fast growing functions, then we might have difficulty keeping the value of the term in the cut, so to speak. The problem can be side-stepped, however, by considering, in the first instance, only purely relational languages, such as the langauge of relational arithmetic. Then, as mentioned earlier, (5) is trivial.

We first prove Theorem 3.5, then, for the special case of relational languages.

Theorem 3.6. *Let $\mathcal{T} \supseteq (\mathsf{I}\Delta_0 + \Omega_1)_R$, the relational version of $\mathsf{I}\Delta_0 + \Omega_1$, where $\mathcal{L}_\mathcal{T}$ is relational, and suppose that $\mathsf{CT}^{-}[\mathcal{T}]$ proves that all axioms of $\mathcal{T}$ are true. Then $\mathsf{CT}^{-}[\mathcal{T}]$ proves the consistency of $\mathcal{T}$ on a cut and so is not interpretable in $\mathcal{T}$. Moreover, $\mathsf{CT}^{-}[\mathcal{T}]$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and even $\mathsf{I}\Delta_0 + \Omega_1 + \mathsf{Con}(\mathcal{T})$.*

Proof. As noted, the usual proofs of (1)–(4) can be adapted to show that their starred versions are inductive, so, by Lemma 2.4, $\mathsf{CT}^{-}[\mathcal{T}]$ proves their relativizations to some cut and therefore proves that 'logic is true' on this cut; we can also assume that $\mathsf{CT}^{-}[\mathcal{T}]$ proves that the axioms of $\mathsf{I}\Delta_0 + \Omega_1$ hold on the cut. We now 'work on this cut', as it is said: Relativizing everything to the cut, we can prove that "n line proofs have true conclusions" is inductive and will therefore be able to construct a subcut on which the relativization of "for all n, n line proofs have true conclusions" is true. The relativization of "all theorems of $\mathcal{T}$ are true" to that cut will then be provable, and so we will be able to prove the consistency of $\mathcal{T}$ on that cut.

To fill in a little detail, consider a formula $\phi(n)$ that says: if n is the Gödel number of a proof such that (i) n lies in our cut and (ii) every formula in the proof also lies in this cut, then (iii) every formula occurring in the proof is true. I.e., if we let $\lambda(x)$ be a formula describing the cut on which logic is true, then $\phi(n)$ is:[46]

$$\lambda(n) \wedge \mathsf{Bew}_{\mathcal{T}}(n) \wedge \forall m < \mathsf{lh}(n)[\lambda(\mathsf{val}(n,m))] \rightarrow \\ \forall m < \mathsf{lh}(n)[\mathrm{T}(\mathsf{val}(n,m))]$$

Now consider:

$$\forall k \leq n(\phi(k))$$

The usual argument that establishes that all proofs have true conclusions can be used to show that this is inductive, since all the formulas involved here lie in our cut, and logic is true on that cut. By Lemma 2.2, there is a cut $\kappa(x)$ on which $\forall k \leq n(\phi(k))$ holds, and, by Lemma 2.3, we can assume that the axioms of $\mathsf{I}\Delta_0 + \Omega_1$ are available on this cut, as well.

So we have:[47]

$$\forall n\{\kappa(n) \rightarrow \\ \forall k \leq n[\lambda(k) \wedge \mathsf{Bew}_{\mathcal{T}}(k) \wedge \forall m < \mathsf{lh}(k)[\lambda(\mathsf{val}(k,m))] \rightarrow \\ \forall m < \mathsf{lh}(k)[\mathrm{T}(\mathsf{val}(k,m))]\}$$

Taking k to be n, we thus have:

$$\forall n\{\kappa(n) \wedge \lambda(n) \wedge \mathsf{Bew}_{\mathcal{T}}(n) \wedge \forall m < \mathsf{lh}(n)[\lambda(\mathsf{val}(n,m))] \rightarrow \\ \forall m < \mathsf{lh}(n)[\mathrm{T}(\mathsf{val}(n,m))]\}$$

What we want is:

$$(*) \qquad \forall n\{\kappa(n) \wedge \mathsf{Bew}_{\mathcal{T}}(n) \rightarrow \forall m < \mathsf{lh}(n)[\mathrm{T}(\mathsf{val}(n,m))]\}$$

We thus need to eliminate the other conjuncts of the antecedent:

$$\lambda(n) \qquad \text{and} \qquad \forall m < \mathsf{lh}(n)[\lambda(\mathsf{val}(n,m))]$$

by showing that they follow from the other two: $\kappa(n)$ and $\mathsf{Bew}_{\mathcal{T}}(n)$. But $\kappa(x)$ is a sub-cut of $\lambda(x)$, and, for the other, we need only make sure that, if $\lambda(n)$, then:

$$\mathsf{seq}(n) \rightarrow \forall m < \mathsf{lh}(n)(\mathsf{val}(n,m) < n)$$

[46] The third conjunct will often be redundant, given the usual sorts of Gödel numberings: If n lies in the cut, then the Gödel numbers of the formulae occurring in the proof it codes will be $\leq n$. But of course it cannot hurt to include it.

[47] The bounded quantifiers here are relativized to $\kappa(x)$, as well, but the relativization is redundant, since cuts are closed downwards and $\leq$ is transitive on this cut.

But if we have a reasonable coding of sequences, we will be able to prove that this is true on our cut, since $\mathsf{I}\Delta_0 + \Omega_1$ is available there.

From (*), then, we easily derive

$$\forall n \forall m \{\kappa(n) \wedge \kappa(m) \wedge \mathsf{Bew}_{\mathcal{T}}(n, m) \rightarrow \mathrm{T}(m)\}$$

in which case also:

$$\forall n \{\kappa(n) \rightarrow \neg\mathsf{Bew}_{\mathcal{T}}(n, \ulcorner 0 = 1 \urcorner)\}$$

since $\neg\mathrm{T}(\ulcorner 0 = 1 \urcorner)$. So $\mathcal{T}$ is consistent on $\kappa(x)$.

The rest now follows from Proposition 2.8. □

With Theorem 3.6 in hand, we can extend the result to non-relational languages and so establish Theorem 3.5.

Proof of Theorem 3.5. Let $\mathcal{T}_R$ be the relational version of $\mathcal{T}$. What we are going to see is that $\mathsf{CT}^-[\mathcal{T}_R]$ is interpretable in $\mathsf{CT}^-[\mathcal{T}]$. It is easy enough to interpret $\mathcal{T}_R$ in $\mathcal{T}$, of course, via such translations as:

$$r(\ulcorner \mathsf{A}xyz \urcorner) = \ulcorner x + y = z \urcorner$$

That is not all we need to do, however. We need to interpret the *semantics* of the relational language in that of the non-relational language, as well. But we can just translate $\mathsf{Sat}_\sigma(A)$ as $\mathsf{Sat}_\sigma(r(A))$, where $r(x)$ is a formula of the language of $\mathcal{T}$ that expresses the translation from $\mathcal{L}_{\mathcal{T}_R}$ to $\mathcal{L}_{\mathcal{T}}$.[48] Since $r(x)$ commutes with the logical connectives, proving the translations of the semantic axioms for the connectives will be easy. For example, the translation of

$$\mathsf{Sat}_\sigma(\ulcorner A \wedge B \urcorner) \equiv \mathsf{Sat}_\sigma(A) \wedge \mathsf{Sat}_\sigma(B)$$

is

$$\mathsf{Sat}_\sigma(r(\ulcorner A \wedge B \urcorner)) \equiv \mathsf{Sat}_\sigma(r(A)) \wedge \mathsf{Sat}_\sigma(r(B))$$

But $r(\ulcorner A \wedge B \urcorner)$ just is $\ulcorner r(A) \wedge r(B) \urcorner$. And since we did not relativize the interpretation, the case of quantification is no harder.

The clauses for the non-logical constants are also easy. Consider, for example, that for $\mathsf{A}xyz$, which is essentially:

$$\mathsf{Sat}_\sigma(\ulcorner \mathsf{A}v_i v_j v_k \urcorner) \equiv \mathsf{A}(\mathsf{val}(\sigma, i), \mathsf{val}(\sigma, j), \mathsf{val}(\sigma, k))$$

Its translation is:

$$\mathsf{Sat}_\sigma(r(\ulcorner \mathsf{A}v_i v_j v_k \urcorner)) \equiv \mathsf{val}(\sigma, i) + \mathsf{val}(\sigma, j) = \mathsf{val}(\sigma, k)$$

But $r(\ulcorner \mathsf{A}v_i v_j v_k \urcorner)$ is $v_i + v_j = v_k$, so this becomes:

$$\mathsf{Sat}_\sigma(\ulcorner v_i + v_j = v_k \urcorner) \equiv \mathsf{val}(\sigma, i) + \mathsf{val}(\sigma, j) = \mathsf{val}(\sigma, k)$$

[48] Since the translation is recursive, it will of course be representable in $\mathcal{T}$. (In fact, as Visser (1992, §7.3) notes, careful formulation makes it p-time.) In general, of course, it will be represented by some formula $R(x, y)$, not by a term like $r(x)$. But this point affects nothing that follows and only complicates the exposition. (We certainly need to know that every formula has a translation, and we may need to know that every formula has exactly one translation. But $\mathsf{I}\Delta_0 + \Omega_1$ will prove such facts.)

which is immediate from the clauses for $+$ and for variables.

So $\mathsf{CT}^-[\mathcal{T}_R]$ is interpretable in $\mathsf{CT}^-[\mathcal{T}]$.

Now, $\mathsf{CT}^-[\mathcal{T}_R]$ proves the consistency of $\mathcal{T}_R$ on a cut. That is, $\mathsf{CT}^-[\mathcal{T}_R]$ proves

$$(*) \qquad \forall n\{\kappa(n) \to \neg\mathsf{Bew}_{\mathcal{T}_R}(n, \ulcorner 0 = 1 \urcorner)\}$$

for some cut formula $\kappa(x)$. So $\mathsf{CT}^-[\mathcal{T}]$ proves the translation of (*), since it proves the translation of any theorem of $\mathsf{CT}^-[\mathcal{T}_R]$.

What we want to see now is that $\mathsf{CT}^-[\mathcal{T}]$ actually proves (*) itself. This is true, but we need to be careful. The version of (*) that $\mathsf{CT}^-[\mathcal{T}_R]$ proves is formulated in the language of *relational* arithmetic; any version of (*) that $\mathsf{CT}^-[\mathcal{T}]$ might prove would be formulated in the language of arithmetic itself. So what we want to see is that $\mathsf{CT}^-[\mathcal{T}]$ proves its own version of (*). But, of course, the translation of (*) that $\mathsf{CT}^-[\mathcal{T}]$ proves is already formulated in the language of arithmetic, and the simplistic, and unrelativized, nature of that translation guarantees that $\mathsf{CT}^-[\mathcal{T}]$ does indeed prove its own version of (*).

Now, suppose $\mathsf{Bew}_{\mathcal{T}}(n, \ulcorner 0 = 1 \urcorner)$. Then, since $\mathcal{T}$ is interpretable in $\mathcal{T}_R$, also $\mathsf{Bew}_{\mathcal{T}_R}(m, \ulcorner 0 = 1 \urcorner)$, for some m. We would therefore have that

$$(**) \qquad \forall n\{\kappa(n) \to \neg\mathsf{Bew}_{\mathcal{T}}(n, \ulcorner 0 = 1 \urcorner)\}$$

if we knew that $\mathsf{CT}^-[\mathcal{T}]$ could prove:

$$\kappa(n) \wedge \mathsf{Bew}_{\mathcal{T}}(n, \ulcorner 0 = 1 \urcorner) \to \exists m[\kappa(m) \wedge \mathsf{Bew}_{\mathcal{T}_R}(m, \ulcorner 0 = 1 \urcorner)]$$

I.e, we need to know that the translation of proofs does not cause them to grow so fast that they take us out of our cut. But we can suppose that the axioms of $\mathsf{I}\Delta_0 + \Omega_1$ are available on $\kappa(x)$, and that will suffice. □

Theorem 3.7. *$\mathsf{CT}^-[\mathsf{I}\Delta_0 + \Omega_1]$ is interpretable in $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$.*

Proof. The technique involved in this proof is similar to that used in the proof of Theorem 3.5, but it is applied more subtly. It will be clear that it is not special to this particular case. What the proof shows, in effect, is that we can always relativize the *semantic part* of a theory like $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$ to a cut, assuming the cut itself is defined purely arithmetically, and that means that the method of cuts can be used with semantic theories like $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$.

We know, of course, that we can interpret $\mathsf{I}\Delta_0 + \Omega_1$ in Q by relativizing to a cut $\kappa(x)$. The problem is to do so while preserving the semantic part of $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$. We cannot actually expect $\mathsf{CT}^-[\mathsf{I}\Delta_0 + \Omega_1]$ to prove the relativizations of the semantic axioms of $\mathsf{CT}^-[\mathsf{Q}_{\mathsf{seq}}]$. That would mean, in particular, proving the relativization of the clause for $\exists$, which would be:

$$\kappa(\sigma) \to \mathsf{Sat}_\sigma(\ulcorner \exists v_i(\phi v_i) \urcorner) \equiv \exists\tau[\kappa(\tau) \wedge \tau \overset{i}{\sim} \sigma \wedge \mathsf{Sat}_\tau(\ulcorner \phi v_i \urcorner)]$$

This says, in effect, that $\exists v_i(\phi v_i)$ is true iff there is a number *in the cut* that satisfies ϕv_i, and, in general, that is false. But what we can do is re-interpret satisfaction itself so that $\mathsf{Sat}_\sigma(A)$ means: the *relativization* of A is satisfied by σ. That is, we translate $\mathsf{Sat}_\sigma(A)$ as: $\mathsf{Sat}_\sigma(t^\kappa(A))$, where $t^\kappa(x)$ is a syntactic function meaning: the relativization of A to $\kappa(x)$.[49] So what we need to prove is:

$$\kappa(\sigma) \rightarrow \mathsf{Sat}_\sigma(t^\kappa(\ulcorner \exists v_i(\phi v_i)\urcorner)) \equiv \exists\tau[\kappa(\tau) \land \tau \overset{i}{\sim} \sigma \land \mathsf{Sat}_\tau(t^\kappa(\ulcorner \phi v_i \urcorner))]$$

Now, $t^\kappa(\ulcorner \exists v_i(\phi v_i)\urcorner)$ is $\exists v_i(\kappa(v_i) \land t^\kappa(\phi v_i))$, so this becomes:

$$\kappa(\sigma) \rightarrow \mathsf{Sat}_\sigma(\ulcorner \exists v_i(\kappa(v_i) \land t^\kappa(\phi v_i))\urcorner) \equiv \\ \exists\tau[\kappa(\tau) \land \tau \overset{i}{\sim} \sigma \land \mathsf{Sat}_\tau(t^\kappa(\ulcorner \phi v_i \urcorner))]$$

And this is easily proven.

Left to right: By the clauses for $\exists$ and $\land$, $\mathsf{Sat}_\sigma(\ulcorner \exists v_i(\kappa(v_i) \land t^\kappa(\phi v_i))\urcorner)$ iff $\exists\tau[\tau \overset{i}{\sim} \sigma \land \mathsf{Sat}_\tau(\ulcorner \kappa(v_i)\urcorner) \land \mathsf{Sat}_\tau(t^\kappa(\ulcorner \phi v_i \urcorner))]$. But $\kappa(v_i)$ is a concrete formula of the language of arithmetic that defines the cut—a long one, but one we could actually write down—so we can prove a Sat-sentence for it. In particular, we have:

$$\mathsf{CT}^-[\mathcal{T}] \vdash \mathsf{Sat}_\tau(\ulcorner \kappa(v_i)\urcorner) \equiv \kappa(\mathsf{val}(\tau, i))$$

But if $\kappa(\mathsf{val}(\tau, i))$, then, since $\kappa(\sigma)$, also $\kappa(\tau)$. That is: If a sequence is in the cut, and some number is in the cut, then the sequence we get by replacing some member of the original sequence by the new number is also in the cut. Although this is not provable in Q, it is provable in $\mathsf{I}\Delta_0 + \Omega_1$, so it will be true on the cut given by $\kappa(x)$, and we are done.

The converse is similar, and similar arguments work for the other semantic clauses. □

Theorem 3.8. *Let $\mathcal{T} \supseteq \mathsf{Q}$ and suppose that $\mathsf{CT}^-[\mathcal{T}]$ proves that all axioms of $\mathcal{T}$ are true. Then $\mathsf{CT}^-[\mathcal{T}]$ proves the consistency of $\mathcal{T}$ on a cut and interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and even $\mathsf{I}\Delta_0 + \Omega_1 + \mathsf{Con}(\mathcal{T})$.*

Proof. From Theorem 3.5 and Theorem 3.7. □

Corollary 3.9. *Suppose $\mathcal{T} \supseteq \mathsf{Q}$ is finitely axiomatized and consistent. Then $\mathsf{CT}^-[\mathcal{T}]$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and hence is not interpretable in $\mathcal{T}$.*

Proof. By Corollary 2.9, Corollary 3.4, and Theorem 3.8. □

So, in the case of finitely axiomatized theories, $\mathsf{CT}^-[\mathcal{T}]$ is always stronger than $\mathcal{T}$ in the precise sense that it is not interpretable in $\mathcal{T}$.

Can we say, however, just how much stronger $\mathsf{CT}^-[\mathcal{T}]$ is than $\mathcal{T}$? What would be really nice is if we could prove a converse of Corollary 3.9.

Hoped For Result 3.10. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\mathsf{CT}^-[\mathcal{T}]$ is interpretable in $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$, and so $\mathsf{CT}^-[\mathcal{T}]$ and $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ are mutually interpretable.*

[49] Being primitive recursive, $\tau^\kappa(x)$ is of course repesentable in Q. As above, it will actually be represented by a formula, but this will make no difference to what follows.

If we could prove 3.10, then we would have shown that, at least in the finitely axiomatized case, $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ was, in one good sense, of the same strength as $\mathsf{CT}^-[\mathcal{T}]$. Unfortunately, however, while I do not know of any definite reason to think that 3.10 is false, the sort of proof given below of the closely related Theorem 4.7 does not work in the present setting.[50] But it is known that one can get fairly close to 3.10.

Theorem 3.11. *Let $\mathcal{T}$ be a finitely axiomatized, sequential theory. Then $\mathsf{CT}^-[\mathcal{T}]$ is interpretable in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$.*

This follows from results proven by Enayat and Visser (2014, esp. Theorem 4.5) in their recent explorations of full satisfaction classes. Note, however, that Theorem 3.11 does not lead to mutual interpretability, since we have no reason to suppose that $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$ is interpretable in $\mathsf{CT}^-[\mathcal{T}]$, even if $\mathcal{T}$ contains $\mathsf{I}\Sigma_1$.

3.3. **Peano Arithmetic Is a Special Case (I).** The proof of Corollary 3.9 depends essentially upon the assumption that $\mathcal{T}$ is finitely axiomatized. This is because, as mentioned previously, if $\mathcal{T}$ is infinitely axiomatized, then there is no reason, in general, to suppose that $\mathsf{CT}^-[\mathcal{T}]$ will prove that *all* of $\mathcal{T}$'s axioms are true, although it will prove that *each* of them is. But then Theorem 3.5 will not apply. We do, however, have the following obvious corollaries.[51]

Corollary 3.12. *Let $\mathcal{T}$ be sequential. Then $\mathsf{CT}^-[\mathcal{T}]$ is locally interpretable in $\mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U}) : \mathcal{U}$ a finite, sequential fragment of $\mathcal{T}\}$.*

Proof. Every finite fragment of $\mathsf{CT}^-[\mathcal{T}]$ is contained in $\mathsf{CT}^-[\mathcal{U}]$, for some finite fragment $\mathcal{U} \supseteq \mathcal{T}$; we may assume that $\mathcal{U}$ is sequential. Then $\mathsf{CT}^-[\mathcal{U}]$ is interpretable in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{U}) \subseteq \mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U})\}$. □

Corollary 3.13. *If $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$ is reflexive, then $\mathsf{CT}^-[\mathcal{T}]$ is interpretable in $\mathcal{T}$.*

Proof. A reflexive theory, by definition, is one that proves the consistency of each of its finite sub-theories. So $\mathcal{T}$ contains $\mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U})\}$ and so itself locally interprets $\mathsf{CT}^-[\mathcal{T}]$. It then follows from Orey's Compactness Theorem that $\mathcal{T}$ *globally* interprets $\mathsf{CT}^-[\mathcal{T}]$. □

So, in particular, we have:

Corollary 3.14 (Enayat and Visser, 2012, Theorem 5.1)**.** *$\mathsf{CT}^-[\mathsf{PA}]$ is intepretable in PA.*

[50] See footnote 73 for an explanation of why.

[51] We need the finite fragments $\mathcal{U}$ in question to be sequential because the proof talks about $\mathsf{CT}^-[\mathcal{U}]$, and that only makes sense if $\mathcal{U}$ is sequential. One might therefore wonder how we know that there are any finite, sequential fragments of $\mathcal{T}$. But $\mathcal{T}$ is sequential, by hypothesis, and its being so involves there being a certain finite set of formulae that $\mathcal{T}$ can prove. So there is some finite fragment of $\mathcal{T}$ that is sequential, and so then is every fragment containing it.

That gives us one sense in which what happens when one adds a truth-theory to PA can be very different from what happens when one adds a truth-theory to some other theory, in particular, to a finitely axiomatized theory.

It's worth emphasizing that the reason we get Corollary 3.14 is *not* that PA has full induction, or anything of that sort.[52] There are really two reasons for PA's anomalous behavior. The first is that, as I have mentioned several times already, if $\mathcal{T}$ is not finitely axiomatizable, then there is no reason to expect that $\mathsf{CT}^-[\mathcal{T}]$ will prove that *all* axioms of $\mathcal{T}$ are true, although it will prove that *each* of them is. And, indeed, we have:[53]

Corollary 3.15. $\mathsf{CT}^-[\mathsf{PA}]$ *does* not *prove that all axioms of* PA *are true.*

Proof. From Theorem 3.5 and Corollary 3.14. □

The second reason is that PA is not just not finitely axiomatizable but is reflexive. As Corollary 3.13 shows, we have the analogue of Corollary 3.14 for any reflexive theory. That means, for example, that it holds for arbitrary theories constructed in the following sort of way:

$$\begin{aligned} T_0 &= \mathsf{I}\Sigma_1 & C_0 &= \mathsf{Con}(\mathsf{I}\Sigma_1) \\ T_1 &= \mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1) & C_1 &= \mathsf{Con}(\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1)) \\ T_{n+1} &= T_n + C_n & C_{n+1} &= \mathsf{Con}(T_{n+1}) \\ \mathsf{I}\Sigma_1^{cl} &= \cup T_n \end{aligned}$$

It is obvious that $\mathsf{I}\Sigma_1^{cl}$ is reflexive, since every finite sub-theory of $\mathsf{I}\Sigma_1^{cl}$ is contained in one of the T_n, and T_{n+1} proves $\mathsf{Con}(T_n)$, trivially.[54]

3.4. **Semantic Consistency Proofs.** As I have emphasized, what was shown in Section 3.2 is *not* that $\mathsf{CT}^-[\mathcal{T}]$ proves that $\mathcal{T}$ is consistent. If $\mathcal{T}$ is a finitely axiomatized (sequential) theory, then $\mathsf{CT}^-[\mathcal{T}]$ will prove that $\mathcal{T}$'s axioms are true, but $\mathsf{CT}^-[\mathcal{T}]$ does not have the induction axioms needed to prove that all of $\mathcal{T}$'s theorems are true (or even that all the theorems of pure logic are true). The natural question to ask, then, is: What exactly do we need to get a proof of $\mathcal{T}$'s consistency? We need $\mathcal{T}$ to contain some induction axioms in the first place, and then we need

[52] Although, as Enayat and Visser mention, we do get the analogue of Corollary 3.14 for "theories. . . that have access to the full scheme of induction over their ambient 'numbers'" (Enayat and Visser, 2012, p. 13, fn. 12).

[53] Of course, this depends upon how the axioms of PA are specified. I am assuming them to be given by some purely syntactic, Σ_1 formula. If we specify the axioms as "those axioms of PA, syntactically specified, that are true", then $\mathsf{CT}^-[\mathsf{PA}]$ will trivially prove that all the axioms of *that* theory are true. As said, however, we treat theories intensionally.

[54] Another well-known reflexive theory is Primitive Recursive Arithmetic, or PRA, but as usually formulated, the language of PRA is not finite, so I am not sure whether or how the results proven here apply to it.

to replace $\mathsf{CT}^{-}[\mathcal{T}]$ with a version that extends the induction axioms to permit semantic predicates—in particular, the satisfaction predicate—to occur therein. But how much 'semantic induction' do we require?

It is not at all obvious, in general, what it means to 'extend a theory's induction scheme'. The scheme might itself be stated in such a way as to exclude formulae containing semantic vocabulary. To take a trivial example, the scheme might require that its instances contain no predicates other than identity. In the cases in which we shall be interested, however, the right way to proceed is both clear and well established. Intuitively, the point is that we may simply regard such formulae as $\mathsf{Sat}_\sigma(x)$ as being among the atomic formulae from which the construction of more complex formulae begins. More precisely, we may make use of the so-called relativized arithmetical hierarchy (Hájek and Pudlák, 1993, pp. 81ff).

Definition. Let X be any set of formulas. A formula is said to be $\Delta_0(X)$ if it belongs to the smallest class of formulae that (i) contains all atomic (arithmetical) formulae and all formulae in X and (ii) is closed under Boolean operations and bounded quantification.

A formula is $\Sigma_1(X)$ if it is of the form $\exists y(\phi(y))$, where $\phi(y)$ is $\Delta_0(X)$, etc.

In our case, if we take Sem to be the set of atomic semantic formulae—$\mathsf{Den}_\sigma(t,x)$, $\mathsf{Sat}_\sigma(x)$, and so forth—then what it means to 'extend induction' in the case of $\mathsf{I}\Delta_0$, say, is that we permit induction on $\Delta_0(\mathsf{Sem})$ formulae. The resulting theory is thus $\mathsf{I}\Delta_0(\mathsf{Sem})$. Similarly for $\mathsf{I}\Sigma_1$, etc.

Definition. Suppose that $\mathcal{T}$ is among $\mathsf{I}\Delta_0$, $\mathsf{I}\Sigma_n$, and so forth. Then: $\mathsf{CT}[\mathcal{T}]$ is the result of (i) adding a fully compositional truth-theory, in the same sense as with $\mathsf{CT}^{-}[\mathcal{T}]$, and (ii) extending the induction scheme to permit semantic vocabulary, i.e., extending it to formulae in $\Delta_0(\mathsf{Sem})$ or $\Sigma_n(\mathsf{Sem})$.

Since $\mathsf{CT}[\mathcal{T}]$ does extend whatever induction scheme might be present in $\mathcal{T}$, it does at least have a chance of formalizing the usual sort of semantic proof of $\mathsf{Con}(\mathcal{T})$. To do so, $\mathsf{CT}[\mathcal{T}]$ will need to be able to do two things: (i) Carry out the induction at the core of that proof; and (ii) Prove that all of the logical and non-logical axioms of $\mathcal{T}$ are true.

The inductive core of the proof can be carried out using the same formula used in the proof of Theorem 3.6, namely:

$$\mathsf{Bew}_{\mathcal{T}}(n) \wedge \lambda(n) \wedge \forall m < \mathsf{lh}(n)[\lambda(\mathsf{val}(n,m))] \rightarrow \\ \forall m < \mathsf{lh}(n)\forall\sigma[\mathsf{Sat}_\sigma(\mathsf{val}(n,m))]$$

This is $\Pi_1(\mathsf{Sem})$, since $\mathsf{Bew}_{\mathcal{T}}(n)$ is Σ_1.[55] Moreover, as a look back at (1)–(5) will show, the formulae involved in the various inductions needed

[55] It's also important that $\mathsf{val}(x,y,z)$ and $\mathsf{lh}(x,y)$ are Δ_1. But they certainly will be in $\mathsf{I}\Sigma_1$.

to prove that logic is true are all $\Pi_1(\mathsf{Sem})$, as well—except for the one concerning denotation, which is $\Sigma_1(\mathsf{Sem})$. But $\mathsf{I}\Sigma_1$ has induction for Π_1 formulae (Hájek and Pudlák, 1993, p. 63, theorem 2.4). So we have:[56]

Theorem 3.16. *Suppose $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$ and suppose further that $\mathsf{CT}[\mathcal{T}]$ proves that all axioms of $\mathcal{S}$ (which may or may not be $\mathcal{T}$) are true. Then $\mathsf{CT}[\mathcal{T}]$ proves $\mathsf{Con}(\mathcal{S})$.*

Corollary 3.17. *Suppose $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$ is finitely axiomatized. Then $\mathsf{CT}[\mathcal{T}]$ proves $\mathsf{Con}(\mathcal{T})$.*

This might seem like a nice, neat result. Since $\mathsf{I}\Sigma_n$ is finitely axiomated, we'll get that $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves $\mathsf{Con}(\mathsf{I}\Sigma_1)$, that $\mathsf{CT}[\mathsf{I}\Sigma_2]$ proves $\mathsf{Con}(\mathsf{I}\Sigma_2)$, and so forth. Unfortunately, however, things are not nearly so tidy.

Everyone knows that $\mathsf{CT}[\mathsf{PA}]$ proves $\mathsf{Con}(\mathsf{PA})$. But it's a good deal less obvious how it does so than people often seem to suppose. What you usually hear people say—and what I myself usually say—is that the proof goes like this: First, you prove that all of the axioms are true; then you prove.... But wait! How are we supposed to prove that *all* of the axioms of PA are true?[57] We can easily enough prove, of *each* axiom, that it is true, since we can prove its T-sentence and we can prove it. But that is an entirely different matter. There are truckloads of very important cases where PA can prove that *each* number blurgs without being able to prove that *every* number blurgs, not least of which is when 'x blurgs' means: x does not code a proof of a contradiction. So again: How do we prove that *all* of PA's axioms are true?

The answer is that the truth of all the axioms falls out of a single instance of the extended induction scheme. Consider the formula:

$$\phi(a, z, \sigma) \overset{df}{\equiv} \exists\tau \left[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau, 0, z) \wedge \mathsf{Sat}_\tau(a)\right]$$

Here, a is meant to code a formula with v_0 free, e.g., $A(v_0, \vec{y})$, where $\vec{y}$ indicates additional free variables that might occur. So what $\phi(\ulcorner A(v_0, \vec{y})\urcorner, \sigma, z)$ says is that $A(v_0, \vec{y})$ is satisfied by the sequence that is just like σ except that it assigns z to v_0—roughly speaking, that $A(v_0, \vec{y})$ is true of z.

The formula $A(v_0, \vec{y})$ is what we really care about, so I will henceforth use $A(v_0, \vec{y})$ as if it were a variable with which we are reasoning in $\mathsf{CT}[\mathsf{PA}]$, in order to make what follows intelligible.

[56] Kotlarski (1986) seems to imply that this result can be strengthened to $\mathcal{T} \supseteq \mathsf{I}\Delta_0$. But Kotlarski is simply not careful enough about the case of the logical axioms. Enayat and Visser (2012) have shown that Kotlarski's result can be salvaged in the model-theoretic setting in which he is working by strengthening the conditions on satisfaction classes. In the present axiomatic setting, one could similarly add an axiom to the truth-theory stipulating that 'variable switching' works as it should. But that does not seem very interesting. It thus remains an open question whether the theory I am calling $\mathsf{CT}[\mathsf{I}\Delta_0]$ proves $\mathsf{Con}(\mathsf{I}\Delta_0)$, let alone whether it proves $\mathsf{Con}(\mathsf{PA})$.

[57] Wang (1952, p. 260) credits Rosser with the observation that this question needs to be addressed.

We want to show that all instances of the induction scheme are true. So, what we want to show is that

(*) $$A(0,\vec{y}) \wedge \forall v_0\, [A(v_0,\vec{y}) \rightarrow A(Sv_0,\vec{y})] \rightarrow \forall v_0(A(v_0,\vec{y}))$$

is true, for all $A(v_0,\vec{y})$. And (*) will be true just in case it is satisfied by every sequence σ. But then, by the clauses for the connectives, that holds just in case, for every sequence σ:[58]

$$\begin{gathered}\mathsf{Sat}_\sigma(\ulcorner A(0,\vec{y})\urcorner) \wedge \\ \mathsf{Sat}_\sigma(\ulcorner \forall v_0[A(v_0,\vec{y}) \rightarrow A(Sv_0,\vec{y})]\urcorner) \rightarrow \\ \mathsf{Sat}_\sigma(\ulcorner \forall v_0(A(v_0,\vec{y}))\urcorner)\end{gathered}$$

This is what we now need to prove. We have the induction axiom:

$$\begin{gathered}\phi(\ulcorner A(v_0,\vec{y})\urcorner, 0, \sigma) \wedge \\ \forall v_0[\phi(\ulcorner A(v_0,\vec{y})\urcorner, v_0, \sigma) \rightarrow \phi(\ulcorner A(v_0,\vec{y})\urcorner, \mathsf{S}v_0, \sigma)] \rightarrow \\ \forall v_0[\phi(\ulcorner A(v_0,\vec{y})\urcorner, v_0, \sigma)]\end{gathered}$$

So it will be enough to show that:

(a) $\mathsf{Sat}_\sigma(\ulcorner A(0,\vec{y})\urcorner)$ implies $\phi(\ulcorner A(v_0,\vec{y})\urcorner, 0, \sigma)$

(b) $\mathsf{Sat}_\sigma(\ulcorner \forall v_0[A(v_0,\vec{y}) \rightarrow A(\mathsf{S}v_0,\vec{y})]\urcorner)$ implies $\forall v_0[\phi(\ulcorner A(v_0,\vec{y})\urcorner, v_0, \sigma) \rightarrow \phi(\ulcorner A(v_0,\vec{y})\urcorner, \mathsf{S}v, \sigma)]$

(c) $\forall v_0[\phi(\ulcorner A(v_0,\vec{y})\urcorner, v_0, \sigma)]$ implies $\mathsf{Sat}_\sigma(\ulcorner \forall v_0 A(v_0,\vec{y})\urcorner)$

None of these are terribly difficult, given three important facts:

(i) If σ and τ agree on the free variables present in some formula ψ, then $\mathsf{Sat}_\sigma(\psi)$ iff $\mathsf{Sat}_\tau(\psi)$.

(ii) If $\mathsf{Sat}_\sigma(\ulcorner\psi(0)\urcorner)$ and $0 = \mathsf{val}(\sigma, 0)$, then $\mathsf{Sat}_\sigma(\ulcorner\psi(v_0)\urcorner)$.

(iii) If $\mathsf{Sat}_\sigma(\ulcorner\psi(\mathsf{S}v_0)\urcorner)$, $\tau \overset{0}{\sim} \sigma$, and $\mathsf{val}(\tau,0) = S(\mathsf{val}(\sigma,0))$, then $\mathsf{Sat}_\tau(\ulcorner\psi(v_0)\urcorner)$.[59]

All of these are provable in $\mathsf{CT[PA]}$ by the usual sorts of arguments.

We get (a) immediately from (ii).

For (b), we need to derive:

$$\begin{gathered}\forall v_0\{\exists\tau[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau, 0, v_0) \wedge \mathsf{Sat}_\tau(\ulcorner A(v_0,\vec{y})\urcorner)] \rightarrow \\ \exists\tau[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau, 0, \mathsf{S}v_0) \wedge \mathsf{Sat}_\tau(\ulcorner A(v_0,\vec{y})\urcorner)]\}\end{gathered}$$

from $\mathsf{Sat}_\sigma(\ulcorner \forall v_0[A(v_0,\vec{y}) \rightarrow A(\mathsf{S}v_0,\vec{y})]\urcorner)$. Applying the semantic clauses to the latter, we have:

(*) $$\forall \chi \overset{0}{\sim} \sigma\, [\mathsf{Sat}_\chi(\ulcorner A(v_0,\vec{y})\urcorner) \rightarrow \mathsf{Sat}_\chi(\ulcorner A(\mathsf{S}v_0,\vec{y})\urcorner)]$$

Now fix v_0 and suppose that for some τ such that $\tau \overset{0}{\sim} \sigma$ and $\mathsf{val}(\tau, 0, v_0)$, $\mathsf{Sat}_\tau(\ulcorner A(v_0,\vec{y})\urcorner)$. So by (*), $\mathsf{Sat}_\tau(\ulcorner A(\mathsf{S}v_0,\vec{y})\urcorner)$. Now let χ be just like τ

[58] So here, $\ulcorner A(0,\vec{y})\urcorner$ really means: the Gödel number of the result of substituting the Gödel number for 0 for v_0 in the formula whose Gödel number is a. Similarly elsewhere.

[59] We're assuming, of course, that *all* free occurences of v_0 have been replaced by occurrences of $\mathsf{S}v_0$.

except that it assigns the variable 'v_0' the successor of what τ assigns it. Then, by (iii), $\mathsf{Sat}_\chi(\ulcorner A(v_0, \vec{y})\urcorner)$. So we have

$$\chi \overset{0}{\sim} \sigma \wedge \mathsf{val}(\chi, 0, \mathsf{S}v_0) \wedge \mathsf{Sat}_\chi(\ulcorner A(v_0, \vec{y})\urcorner)$$

as wanted.

The argument for (c) is similar and is left to the reader.

What makes all of this go, then, is the fact that PA is schematically axiomatized: An extended instance of the induction scheme for PA can be made to yield all of the unextended instances. But, by the same token, the argument works *only* because PA is schematically axiomatized. If $\mathcal{T}$ is an infinitely axiomatized theory that is *not* schematically axiomatized, then there is no reason whatsoever to expect that $\mathsf{CT}[\mathcal{T}]$ should prove that all of $\mathcal{T}$'s axioms are true.

So, as Visser once put it, the fact that CT[PA] proves Con(PA) is something of a happy accident. Too happy, for our purposes, as we are about to see.

Theorem 3.18. $\mathsf{CT}[\mathsf{I}\Sigma_1]$ *proves that all axioms of* PA *are true.*

Proof. The argument just given needed only a handful of instances of extended induction. One was for the formula $\phi(a, \sigma, z)$. This is $\Sigma_1(\mathsf{Sem})$. The other thing we need to check is that the general principles (i)–(iii) on which we relied can be proven in $\mathsf{CT}[\mathsf{I}\Sigma_1]$. The proofs of these are all by induction, but, other than the semantic notions, there is nothing in these that isn't primtive recursive and so Δ_1 in $\mathsf{I}\Sigma_1$;[60] the universal quantifier over sequences makes the relevant claims Π_1(Sem). So the proof that all axioms of PA are true can be carried out in $\mathsf{CT}[\mathsf{I}\Sigma_1]$. □

Corollary 3.19. $\mathsf{CT}[\mathsf{I}\Sigma_1]$ *proves* Con(PA).

Proof. By Theorem 3.16 and Theorem 3.18. □

If Corollary 3.19 seems surprising, it is because one might have thought we needed to assume the axioms of PA in order to able to prove that all of the axioms of PA are true. Well, we don't.[61] As Tarski himself put it, we need not assume "axioms which have the same meaning as the axioms of the science under investigation", but only ones that "suffice... for the establishment of all sentences having the same meaning as the theorems of the science being investigated" (Tarski, 1958, p. 211). And it turns out that assuming extended Σ_1 induction *is* assuming axioms that suffice to establish all axioms of PA.

Indeed, since Con(PA) is a single theorem of CT[PA], the full power of CT[PA] can't be needed for the proof, anyway: Only finitely many axioms of CT[PA] will be needed, so Con(PA) has to be provable in $\mathsf{CT}[\mathsf{I}\Sigma_n]$, for

[60] The quantifier over variables in the antecedent of (i) can be bounded by ψ.

[61] And, for the reason to be given in the next paragraph, it wouldn't help us if we did: There's no way we can use all the axioms of PA in a single proof.

some n. Nonetheless, I take Corollary 3.19 to be a bad result in the context of the present investigation, in so far as it suggests that we do not yet have things properly formulated. It's a perfectly natural question what sort of truth-theory you need to formalize the obvious semantic consistency proof of $\mathsf{I}\Sigma_1$. It's disappointing if the answer turns out to be, "One that proves $\mathsf{Con}(\mathsf{PA})$".[62]

More importantly, the thought guiding this investigation is that truth is, somehow, closely related to consistency. There is a clear sense in which adding a theory of truth to a theory gives us consistency: Corollary 3.9 and Corollary 3.17 gives us that. But we would also like to find some sense in which adding a consistency statement to some theory gives us a theory of truth for that theory's language. But we are unable to prove Hoped For Result 3.10, and it follows from Corollary 3.19 that its analogue in this case—something like: $\mathcal{T} + \mathsf{Con}(\mathcal{T})$ is interpretable in $\mathsf{CT}[\mathcal{T}]$—is certainly false.[63]

Fortunately, it turns out that we can do better.

4. Disentangling Syntax from the Object Theory

4.1. **Reviving an Old Approach to Truth-theories.** What's responsible for Corollary 3.19?

Semantic consistency proofs make use of two different sorts of theories, for two very different sorts of reasons. On the one hand, we have a 'base theory' that gives us the syntactic machinery we need to formulate our truth-theory and then to reason within it. Among other things, for example, the induction axioms allow us to formalize arguments by induction on the complexity of expressions, or the length of proofs, or what have you. On the other hand, there is the object theory, which is the theory we mean to be reasoning *about*: the theory whose consistency we mean to be proving, for example. We need to know that all the axioms of the object theory are true, and the idea is to get the truth of the axioms from the axioms themselves, deriving their truth via their T-sentences.

As we have seen, however, that is not at all how things work in the case of PA. The fact that *all* of the axioms of PA are true is not derived from the axioms of PA via the T-sentences, and, on reflection, it's easy to see that it can't be: The truth of *each* axiom of PA can be derived from that very axiom, but that's it; the truth of *all* the axioms of PA is a consequence, not of those axioms, but of a handful of instances of extended Σ_1 induction. So what leads to Corollary 3.19 is the fact that a single theory is playing both of the roles I just distinguished: In

[62] Of course, PA itself proves $\mathsf{Con}(\mathsf{I}\Sigma_1)$, and the argument is semantic in character—it uses a partial truth-theory for the language of arithmetic—but it is very much not the sort of argument that we are discussing.

[63] I have discussed some other reasons to be dissatisfied with the usual way of handling theories of truth elsewhere (Heck, 2014b, §3.3).

$\mathsf{CT}[\mathsf{I}\Sigma_1]$, $\mathsf{I}\Sigma_1$ is *both* the underlying syntax *and* what provides us with the axioms of the theory we had meant to be reasoning about. We must extend induction to Σ_1 formulae containing semantic vocabulary in order to formalize certain sorts of arguments, but there are instances of that same form that entail the truth of principles stated in the object language that go well beyond what we'd meant to be assuming in our object theory. The problem, to put it as directly as possible, is that we can't strengthen the syntax without thereby strengthening the object theory.

The solution to the problem is therefore obvious: We need to disentangle the syntactic theory from the object theory. And, interestingly enough, this is how Tarski himself proceeds in "The Concept of Truth in Formalized Languages":

> A meta-language which meets our requirements must contain three groups of expressions: (1) expressions of a general logical kind; (2) expressions having the same meaning as all the constants of the language to be discussed. . . ; (3) expressions of the structural-descriptive type which denote single signs and expressions of the language considered, whole classes and sequences of such expressions or, finally, the relations existing between them. (Tarski, 1958, pp. 210–1)

The expressions mentioned under (3) belong, of course, to syntax. Tarski does not actually say that these expressions will be disjoint from those mentioned under (2), but it is natural to read him that way. That is plainly how he conceives the matter in his discussion of the calculus of classes (Tarski, 1958, pp. 172ff), which is far too weak to interpret syntax. Tarski was of course aware—at least by the time his paper was published—that syntax can be interpreted in arithmetic: His famous theorem on the indefinability of truth depends upon that fact. But the *positive* part of Tarski's project—showing how it is possible to define truth in a consistent manner, suitable for the purposes of meta-mathematics—in no way depends upon this now common manoeuver. The basic idea of separating the syntax from the object theory is thus an old one.

So let $\mathcal{L}$ be the (finite) language for which we want to give a truth-theory. We let $\mathcal{S}$ be a disjoint (and fixed) language in which we will formalize syntax. The most natural choice for $\mathcal{S}$, and the one that would be closest to Tarksi's original intentions, would be the language whose sole primitive is the binary function-symbol: $a \frown b$, intended to mean: the concatenation of a and b (Quine, 1946; Corcoran et al., 1974; Grzegorczyk, 2005). To keep things familiar, however, we shall take $\mathcal{S}$ to be isomorphic to the language of arithmetic and assume that syntax is coded in the usual way.[64] Think of $\mathcal{S}$ as the language of arithmetic

[64] The fact that $\mathcal{L}$ is disjoint from $\mathcal{S}$ is of course no obstacle to our coding facts about $\mathcal{L}$ in $\mathcal{S}$.

written in boldface, or something of the sort. Our theory of syntax can then be taken be Q, or $\mathsf{I}\Sigma_1$, or whatever we wish.

Now, if we're going to do the semantics of $\mathcal{L}$, then we're going to need to be able to talk about the things $\mathcal{L}$ talks about. In particular, if we're going to have the usual Tarski-style clauses for the primitive expressions of the object language, then we are going to need to have the expressive resources of $\mathcal{L}$ available to us, as Tarski notes at (2). So the obvious choice for the language of our semantic theory would be $\mathcal{S} \cup \mathcal{L}$. There are, however, complications. Suppose that $\mathcal{L}$ is the language of set theory. Then the quantifiers in sentences of $\mathcal{L}$ would normally be understood as ranging over all and only the sets. The quantifiers in sentences of $\mathcal{S}$, however, do not range over all sets. So we need to keep the domains of $\mathcal{S}$ and $\mathcal{L}$ separate. The simplest way to do so is to let the semantic theory be many-sorted, so that's what we'll do. Variables ranging over the domain of $\mathcal{S}$ will be italic; those ranging over the domain of $\mathcal{L}$ will be upright.[65]

If we do go this way, then we're also going to need a separate theory of sequences or, better, of assignments of objects to variables: There will be no hope at all of coding sequences of objects from the domain of $\mathcal{L}$ as objects in $\mathcal{S}$, at least not in general. So we shall takes ourselves to have the following theory of assignments available:[66]

$$\forall v[\mathsf{var}(v) \rightarrow \forall\alpha\forall\mathbf{x}\exists\beta(\beta \overset{v}{\sim} \alpha \wedge \mathsf{val}(\beta, v) = \mathbf{x})]$$

What this says is that, given any assignment, the value it assigns to a given variable can always be changed as one pleases. Assignments live in yet a third sort. Variables ranging over them will be Greek letters. That there is at least one assignment, and that every assignment assigns a unique object to each variable, are truths of logic, in this formulation.[67]

Given this theory of assignments, we can then state a truth-theory for $\mathcal{L}$. The theory will be the familiar one, though with some adjustments to take account of the present framework. For example, the universal closures of the following will be axioms common to all semantic theories, independent of $\mathcal{L}$:

(v) $\mathsf{var}(v) \rightarrow \mathsf{Den}_\alpha(v, \mathbf{x}) \equiv \mathbf{x} = \mathsf{val}(\alpha, v)$
($\neg$) $\mathsf{Sat}_\alpha(\ulcorner\neg A\urcorner) \equiv \neg\mathsf{Sat}_\alpha(A)$
($\wedge$) $\mathsf{Sat}_\alpha(\ulcorner A \wedge B\urcorner) \equiv \mathsf{Sat}_\alpha(A) \wedge \mathsf{Sat}_\alpha(B)$
($\forall$) $\mathsf{Sat}_\alpha(\ulcorner\forall\mathrm{v}_i(A(\mathrm{v}_i))\urcorner) \equiv \forall\beta[\beta \overset{i}{\sim} \alpha \rightarrow \mathsf{Sat}_\beta(\ulcorner A(\mathrm{v}_i)\urcorner)]$

The other axioms of the theory will depend upon $\mathcal{L}$. If $\mathcal{L}$ is the language of set theory, then the only other axiom will be:

[65] The two-sorted theory can of course be interpreted in a single sorted theory via the usual relativization to a pair of domains. This is more or less what Craig and Vaught (1958) do. We'll discuss their work further below.

[66] So $\beta \overset{v}{\sim} \alpha$ now abbreviates: $\forall w(w \neq v \rightarrow \mathsf{val}(\beta, w) = \mathsf{val}(\beta, v))$.

[67] This sort of idea is borrowed from Craig and Vaught (1958).

($\in$) $\mathsf{Sat}_\alpha(\ulcorner t \in u \urcorner) \equiv \exists \mathbf{x} \exists \mathbf{y}[\mathsf{Den}_\alpha(t, \mathbf{x}) \wedge \mathsf{Den}_\alpha(u, \mathbf{y}) \wedge \mathbf{x} \in \mathbf{y}]$

In the case of the language of arithmetic, we'll have axioms like:

(0) $\mathsf{Den}_\alpha(\ulcorner 0 \urcorner, \mathbf{x}) \equiv \mathbf{x} = 0$

(+) $\mathsf{Den}_\alpha(\ulcorner t + u \urcorner, \mathbf{x}) \equiv \exists \mathbf{y} \exists \mathbf{z}[\mathsf{Den}_\sigma(t, \mathbf{y}) \wedge \mathsf{Den}_\sigma(u, \mathbf{z}) \wedge \mathbf{x} = \mathbf{y} + \mathbf{z}]$

Note that, in both these cases, the *used* expressions '0' and '+' are expressions of $\mathcal{L}$, not of $\mathcal{S}$.

So that is the theory in which I propose henceforth to work. As for notation:[68]

Definition. Let $\mathcal{T}$ be an arithmetical theory. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ is the semantics for $\mathcal{L}$ just described.

We can think of $\mathsf{TT}^-_{\eta}[\xi]$ as a two-place functor: Given a theory $\mathcal{T}$ in $\mathcal{S}$ and a language $\mathcal{L}$, it returns a new theory that constitutes a semantics for $\mathcal{L}$ based upon $\mathcal{T}$ as syntax. Our interest is in the properties of this functor.[69]

Note that we are not (yet) extending any induction scheme that might be present in $\mathcal{T}$, so $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ is not going to be formalizing semantic consistency proofs of the sort discussed in Section 3.4. More generally, induction in $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ does not apply to statements involving assignments, or semantics, or the object language. The induction axioms must be 'purely syntactical'.

4.2. **The Weakness of Tarskian Truth-theories.** Our goal now is to show that the various results proven in Section 3.2 hold also in the present setting. As we shall see, however, these results are usually available in an improved form.

Lemma 4.1. $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *is a materially adequate theory of truth for* $\mathcal{L}$*. In particular: For each formula* $A(v_1, \dots, v_n)$ *of* $\mathcal{L}$*,* $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *proves:*

$$\mathsf{Sat}_\sigma(\ulcorner A(v_1, \dots, v_n) \urcorner) \equiv A(\mathsf{val}(\sigma, 1), \dots \mathsf{val}(\sigma, n))$$

A fortiori, for each sentence A*,* $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *proves:* $\mathrm{T}(\ulcorner A \urcorner) \equiv A$.

Proof. Essentially the same as that of Lemma 3.1. □

Lemma 4.2. $\mathsf{TT}^-_{\mathcal{A}}[\mathsf{Q}]$ *is interpretable in* Q.

Proof. The basic idea here is very simple: Since no theory stated in $\mathcal{L}$ is so far in evidence, we can give $\mathcal{L}$ the completely trivial interpretation whose domain is $\{0\}$, that takes each term to denote 0, and that takes every predicate to have an empty extension. The theory of assignments

[68] Here, TT stands for: Tarskian truth.

[69] There are some other choices hidden here about how exactly syntax is being formalized. That, as Visser emphasized to me, is true even when we are working in a theory of concatentation. I am assuming for the moment that these choices are made in some sensible, uniform fashion. I.e., I am bracketing any intensionality due to decisions about how exactly to code syntax.

is then completely trivial: $\mathrm{val}(v, \mathrm{x})$ will always be true, for each v and x. A semantic theory for $\mathcal{L}$, so interpreted, is then easily constructed. □

A similar proof works for languages other than the language of arithmetic, though there are complications if $\mathcal{L}$ contains no terms other than variables.[70]

Lemma 4.2 gives us a first indication of why it is worth disentangling syntax from the object theory. If we develop our truth-theory in the usual way, where syntax and the object-theory are intertwined, then the weakest materially adequate truth-theory is $\mathsf{CT}^{-}[\mathsf{Q}_{\mathrm{seq}}]$. And it follows from Theorem 3.8 that $\mathsf{CT}^{-}[\mathsf{Q}_{\mathrm{seq}}]$ is *not* interpretable in Q. In the disentangled setting, however, there is a materially adequate truth-theory for the language of arithmetic that is as weak as it could possibly be: It is interpretable in Q.[71]

It is worth emphasizing, as well, that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ is not just a curiosity but is of real mathematical utility. Lemma 4.1 plays an important role in the proof, due to Craig and Vaught (1958), that every axiomatizable theory that has no finite models has a finitely axiomatizable conservative extension. Their argument is an extension of one due to Kleene (1952).

Consider some recursively axiomatizable theory $\mathcal{T}$. We take a weak, finitely axiomatizable theory of syntax—Q, basically—a weak theory of assignments, and the Tarski clauses for the language of $\mathcal{T}$. I.e., we work in $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$. Then we can prove the T-sentence for each sentence of the language of $\mathcal{T}$ (Craig and Vaught, 1958, p. 296, Lemma 2.4). So now, since the set of $\mathcal{T}$'s axioms is recursive, it is representable in Q, and we need only add one more axiom: All of $\mathcal{T}$'s axioms are true. This theory clearly contains $\mathcal{T}$, and the fact that it is a conservative extension of $\mathcal{T}$ can be proven by the usual sort of model-theoretic argument (Craig and Vaught, 1958, p. 298, Lemma 2.7).

[70] The problem is that, in this case, we will not be able to specify the domain via a formula $\delta(x)$ with just x free. What we can do, however, is use a parameter. This gives us an interpretation with a parameter, so, in the general case, $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ will only be parametrically interpretable in Q.

[71] Moreover, we see that a materially adequate theory of truth for $\mathcal{L}$ need make use of *no information whatsoever* about whatever it is that $\mathcal{L}$ talks about. As said, any theory of truth that is going to be materially adequate, in the sense that it proves all 'disquotational' T-sentences, is of course going to have to have the expressive resources of the object language available to it. But that is all. We haven't even mentioned any theory formulated in $\mathcal{L}$ to this point, let alone made use of one.

These results have another sort of significance. If, as I am inclined to believe (Heck, 2005, 2007), a speaker's semantic competence consists in her tacitly knowing a truth-theory for her language, one might worry that this would credit ordinary speakers with far too much tacit knowledge. But knowing such a theory need involve no more than knowing $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$, and the logical strength of that theory derives entirely from its syntactic component.

4.3. **The Strength of Tarskian Truth-theories.** Our next set of results concern the logical strength of $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$.

Lemma 4.2 tells us that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ is very weak. But this does not really help us to characterize the strength of truth-theories. For one thing, the interpretation of $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ in Q wreaks havoc on the meanings of the primitives of $\mathcal{L}$: It all but treats $\mathcal{L}$ as uninterpreted. How, then, might we force the truth-theory to respect the meanings of $\mathcal{L}$'s primitives? One plausible answer is to require the interpretation to preserve some theory stated in $\mathcal{L}$. Indeed, we might well understand Tarski as taking the object theory to play something like this role. (Though we do not need to suppose, as Tarski may have, that this theory should in any sense consist of 'meaning postulates'.)[72] Moreover, the question how strong truth-theories are is best understood as the question: What does 'adding a truth-theory' give us, in terms of logical strength? That is, if we have some theory $\mathcal{T}$ and we 'add a truth-theory' to it, how strong is the resulting theory, compared to $\mathcal{T}$ itself? In our terminology, the question is thus how $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ compares, in logical strength, to $\mathcal{T}$. Lemma 4.2 just concerns the special case where $\mathcal{T}$ is the null theory.

As was explained in Section 2.2, there are different ways of comparing theories, so we can ask various sorts of questions about the relationship between $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ and $\mathcal{T}$. One question is whether $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is a conservative extension of $\mathcal{T}$. And we have, in fact, already seen that it is: That is just the result due to Craig and Vaught (1958) that was mentioned above.

But there is a different, and ultimately more interesting, question we can ask, namely, whether $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is interpretable in $\mathcal{T}$. And to this question, the answer is "no", at least if $\mathcal{T}$ is finitely axiomatized, for essentially the reasons we saw earlier.

Corollary 4.3. *$\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ proves, of each axiom of $\mathcal{T}$, that it is true.*

Proof. From Lemma 4.1, as Corollary 3.3 was derived from Lemma 3.1. □

Corollary 4.4. *Let $\mathcal{T}$ be a finitely axiomatized sequential theory. Then $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{L}$ proves the obvious, disjunctive formalization of "all axioms of $\mathcal{T}$ are true".*

Theorem 4.5. *Let $\mathcal{T}$ be a consistent theory in $\mathcal{L}$. Then $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ plus "all axioms of $\mathcal{T}$ are true" proves the consistency of $\mathcal{T}$ on a cut and interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

Proof. Essentially the same as that of Theorem 3.8. □

It's worth emphasizing that the hypothesis that all axioms of $\mathcal{T}$ are true does no significant work in this proof beyond giving us the 'base case' of

[72] These remarks are largely based upon observations due to John Burgess.

the argument that allows us to prove $\mathsf{Con}(\mathcal{T})$ on a cut. The actual work is all done in $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$.

Corollary 4.6. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ proves the consistency of $\mathcal{T}$ on a cut and interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$, hence is not interpretable in $\mathcal{T}$.*

Proof. From Corollary 2.9, Corollary 4.4, and Theorem 4.5. □

We thus see again that compositional truth-theories have at least some logical power: If we start with a finitely axiomatized theory $\mathcal{T}$ and add an absolutely minimal but still compositional theory of truth for the language of $\mathcal{T}$, the result is a theory that is logically stronger than $\mathcal{T}$ in the sense that it is not interpretable in $\mathcal{T}$.

So we can still prove the central result of Section 3.2 in the 'disentangled' setting. More interestingly, we can now prove a sort of converse of Corollary 4.6, as well, one that tells us exactly how strong $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is, at least when $\mathcal{T}$ is finitely axiomatized.[73]

Theorem 4.7. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is interpretable in $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

The proof of Theorem 4.7 is similar to the proof of Theorem 4.10. It is in one sense easier, since Theorem 4.10 applies to theories in which induction has been extended to semantic vocabulary and that gives us more work to do. But it is, in a different sense, more complicated, since Q is so weak. Nicolai (2014, §4) gives all the details.

Corollary 4.8. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is mutually interpretable with $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

[73] That some such result should be provable was first suggested to me by Visser, with reference to his paper "The Predicative Frege Hierarchy" (Visser, 2009c), and Theorem 4.7 should probably be credited to Visser. My initial thought was that arguments like those he uses to prove his Theorem 5.2 could be used to prove Hoped For Result 3.10. Indeed, in an earlier version of this material (Heck, 2009, Theorem 5.10), I even claimed as much. As Carlo Nicolai eventually pointed out to me, however, no such proof can work.

The obstacle to proving 3.10 is roughly this. We can use the Henkin–Feferman construction to build a model of $\mathcal{T}$ inside $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$, and this construction in effect delivers a semantic theory for the language of $\mathcal{T}$, just as in the proof of Theorem 4.10 below. But the expressions to which that semantic theory applies live in the *original* model of $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$, *not* in the model of $\mathcal{T}$ we have built. That is no problem if we are trying to interpret $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$, since the syntactic objects do not live in the domain of the object theory, and that is precisely because of how the syntax and the object theory have been disentangled. If we are trying to interpret $\mathsf{CT}^-[\mathcal{T}]$, however, then the syntax itself does have to be interpreted inside the model of $\mathcal{T}$ we have built. And there is no reason to think that can be done. If we knew that the domain of the model $\mathcal{T}$ was isomorphic to, or at least embeddable in, the domain of our original model, then we would stand a chance. But that clearly need not be true, since the original model could perfectly well be standard and yet $\mathcal{T}$ could be, say, $\mathsf{Q} + \neg\mathsf{Con}(\mathsf{Q})$, which has only non-standard models.

Nicolai (2014, §4.1) discusses some related issues.

So we may think of $\mathsf{TT}_{\mathcal{L}}[\mathsf{Q}] + (\cdot)$ as a functor that, given a finitely axiomatized theory, always hands you back a stronger theory. In fact, what it does is 'upGödel' that theory.[74] We can think of Pudlák's form of the second incompleteness theorem, too, as defining a functor: Given a consistent, sequential, finitely axiomatized theory $\mathcal{T}$ containing Q, it hands us $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$, which is guaranteed to be logically stronger than $\mathcal{T}$ itself in the sense that it is not interpretable in $\mathcal{T}$. What we have found, then, is that, for finitely axiomatized theories, $\mathsf{TT}_{\mathcal{L}}[\mathsf{Q}] + (\cdot)$ is the same functor, *modulo* mutual interpretability.

4.4. **Semantic Consistency Proofs, Again.** We have seen that, if $\mathcal{T}$ is finitely axiomatized, then $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is not interpretable in $\mathcal{T}$, because it proves the consistency of $\mathcal{T}$ on a cut. It does so because it proves the basis case and the induction step of a semantic proof of $\mathcal{T}$'s consistency. That leaves us where we were at the end of Section 3.2. The next question to ask, then, is what we need to add if we are to get a proof of the consistency of $\mathcal{T}$. As we saw in Section 3.4, the answer is going to be something along the lines of 'induction for Σ_1 formulae'. In the framework in which we were then working, however, this answer turned out to be disappointing, even if correct. It's true that $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves the consistency of $\mathsf{I}\Sigma_1$, but it also proves the consistency of PA.

The disentangled setting in which we are now working allows us to resolve this problem.[75] What we need to add is, indeed, something along the lines of 'induction for Σ_1 formulae'. But we can now strengthen our theory of syntax *without* thereby strengthening the object theory whose consistency we are trying to prove. Let me emphasize what this says about the role induction plays in semantic consistency proofs: The induction we need for the proof is a *syntactic* principle, not an arithmetical one. It's a principle that has to do, at least in the application we need to make of it, with inductions on proofs; it has *nothing* to do with whatever the object language happens to be about, which could be numbers, sets, graphs, or what have you. This is obvious once stated, but the usual way of formulating truth-theories obscures the point.

So now we need a definition paralleling that of $\mathsf{CT}[\mathcal{T}]$.

Definition. $\mathsf{TT}_{\mathcal{L}}[\mathcal{T}]$ is $\mathsf{TT}^{-}_{\mathcal{L}}[\mathcal{T}]$ with the induction axioms in $\mathcal{T}$ extended to permit semantic vocabulary, reference to assignments, and vocabulary from the object language.

[74] Thanks to Visser for the wonderful neologism.

[75] This sort of idea is taken up as well by Leigh and Nicolai (2013), who discuss 'disentangled' theories of the sort I am about to describe at some length. (As they note, such theories were first presented in an earlier version of this material (Heck, 2009).) The main difference between their discussion and mine is that they fix the syntactic theory to be PA, whereas I allow it to vary. So their $CDT[O]$ is what I would call $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + O$.

As before, this definition isn't perfectly general. But we know how to apply it to the cases that matter here: We simply allow all the new vocabulary to occur in the induction axioms, treating the new primitives as, well, primitives. A complication is that there are now three different sorts of quantifiers—over syntactic objects, assignments, and whatever is in the domain of the object language—that can occur in a formula. But we can just ignore this difference when determining the complexity of a formula. Thus, e.g., '$\exists \mathrm{x}(\mathsf{Den}_\sigma(t, \mathrm{x}))$' counts as Σ_1 for our purposes, and '$\forall \sigma \exists t \exists \mathrm{x}(\mathsf{Den}_\sigma(t, x))$' counts as Π_2.

It is clear that we can now adapt the arguments given in Section 3.4 to our new framework. In particular, we will be able to formalize a semantic proof of $\mathsf{Con}(\mathcal{T})$ in $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1] + \mathcal{T}$, where $\mathcal{L}$ is the language of $\mathcal{T}$. So we have:[76]

Theorem 4.9. *Let $\mathcal{T}$ be a theory in a finite language $\mathcal{L}$. Then $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1]$ plus "all axioms of $\mathcal{T}$ are true" proves $\mathsf{Con}(\mathcal{T})$.*

Hence, if $\mathcal{T}$ is finitely axiomatized, then $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1] + \mathcal{T}$ proves $\mathsf{Con}(\mathcal{T})$.

This, I suggest, is the right way to think about the formalization of semantic consistency proofs.

I should emphasize, before we continue, that the sentence $\mathsf{Con}(\mathcal{T})$ that Theorem 4.9 asserts can be proved in $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is a sentence of the *syntactic* language $\mathcal{S}$, *not* of the object language $\mathcal{L}$, that is, the language of the theory $\mathcal{T}$.[77] Of course, so long as the object language is not ridiculously lacking in expressive power—if it is the language of arithmetic, say, or the language of set theory—and so long as $\mathcal{T}$ is not too terribly weak, then there will also be a sentence of $\mathcal{L}$ that 'expresses' the statement that $\mathcal{T}$ is consistent via coding. So we need to distinguish the sentence $\mathsf{Con}_\mathcal{S}(\mathcal{T})$ of the syntactic language that Theorem 4.9 says can be proven in $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1] + \mathcal{T}$ from the sentence $\mathsf{Con}_\mathcal{L}(\mathcal{T})$ of the object language about which nothing has yet been said. And it can be shown that the object language sentence $\mathsf{Con}_\mathcal{L}(\mathcal{T})$ *cannot* be proven in $\mathsf{TT}_\mathcal{L}[\mathsf{I}\Sigma_1] + \mathcal{T}$. This follows from a much more general observation, due to Halbach, that even $\mathsf{TT}_\mathcal{L}[\mathsf{PA}] + \mathcal{T}$ is a conservative extension of $\mathcal{T}$ (Leigh and Nicolai, 2013, §3.2). That might well seem peculiar, but, as I have argued elsewhere (Heck, 2014b, §5), it simply highlights in a different way how the usual approach to theories of truth conflates the syntactic theory with the object theory.

Henceforth, then, $\mathsf{Con}(\mathcal{T})$ always means: $\mathsf{Con}_\mathcal{S}(\mathcal{T})$.

4.5. **Limitative Results.** We can now establish an analogue of Theorem 4.7 for the case in which induction has been extended, thus showing

[76] The remarks in footnote 42, about how the axioms of $\mathcal{T}$ are specified, apply here, too, of course.

[77] Thanks to Volker Halbach for making me take account of this point.

that the disentangled setting really does allow us to resolve the problem revealed by Corollary 3.19.

Theorem 4.10. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is interpretable in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$.*

Indeed, this same pattern extends through the arithmetical hierarchy.

Theorem 4.11. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then, for all $n \geq 1$, $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ is interpretable in $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$.*

I'll present the proofs of these two results in the next section. First, let me note a few corollaries.

Corollary 4.12. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ is mutually interpretable with $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$.*

Proof. If $\mathcal{T}$ is finitely axiomatized, then $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ proves $\mathsf{Con}(\mathcal{T})$, by Theorem 4.9, and so contains $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$ as a sub-theory. □

Corollary 4.13. *If $n \geq m \geq 1$, then $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{I}\Sigma_n \nvdash \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$. In particular, $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_n \nvdash \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$.*

Proof. By Theorem 4.10, $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m]+\mathsf{I}\Sigma_n$ is interpretable in $\mathsf{I}\Sigma_m+\mathsf{Con}(\mathsf{I}\Sigma_n)$. But $\mathsf{Con}(\mathsf{I}\Sigma_n)$ is provable in $\mathsf{I}\Sigma_{n+1}$ (Hájek and Pudlák, 1993, p. 108), so if $n \geq m$, $\mathsf{I}\Sigma_m + \mathsf{Con}(\mathsf{I}\Sigma_n)$ is actually a sub-theory of $\mathsf{I}\Sigma_{n+1}$. Thus, $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{I}\Sigma_n$ is interpretable in $\mathsf{I}\Sigma_{n+1}$.

So, if $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{I}\Sigma_n \vdash \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$, then $\mathsf{Q} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ is a sub-theory of $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{I}\Sigma_n$ and hence is also interpretable in $\mathsf{I}\Sigma_{n+1}$. But that contradicts Pudlák's version of Gödel's second. □

So, while $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves $\mathsf{Con}(\mathsf{PA})$, $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ most certainly does not, since it does not even prove $\mathsf{Con}(\mathsf{I}\Sigma_2)$. In fact, $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ does not even interpret $\mathsf{I}\Sigma_2$.[78]

This sort of result even extends to $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$, in which induction is completely unrestricted and is available for every formula (subject to the usual restrictions on capturing of variables, etc), no matter what mix of vocabulary from syntax, semantics, and the object language it might contain.

Corollary 4.14. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is mutually locally interpretable with $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$.*

Proof. Each finite fragment of $\mathsf{PA}+\mathsf{Con}(\mathcal{T})$ is contained in one of the $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$, which is interpretable in $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n]+\mathcal{T}$ and so in $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}]+\mathcal{T}$. And each finite fragment of $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}]+\mathcal{T}$ is contained in one of the $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n]+\mathcal{T}$ and so is interpretable in $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$ and so in $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$. □

[78] This is because $\mathsf{I}\Sigma_2$ is the same theory as $\mathsf{I}\Sigma_1$ plus reflection for Σ_3 formulas (Beklemishev, 2005, p. 231, Theorem 7). So $\mathsf{I}\Sigma_2$ proves $\mathsf{Con}(\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1))$. Thanks to Volodya Shavrukov for confirming my suspicion and for the reference.

Since $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$ is, like PA, reflexive,[79] $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is globally interpretable in $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$, by Orey's Compactness Theorem. It is not at all obvious, however, whether $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is reflexive (when $\mathcal{T}$ is finitely axiomatized). It would be nice if it was, though, since then we could remove "locally" from Corollary 4.14.

4.6. **The Proof of Theorem 4.10.** We are going to need a version of the so-called arithmetized completeness theorem (Hájek and Pudlák, 1993, pp. 104–5) that is provable in $\mathsf{I}\Sigma_1$.[80] There are two different ways one often sees this theorem stated, and the proof of Theorem 4.10 rests upon the way these two statements of it relate to one another.

Theorem 4.15 (Arithmetized Completeness Theorem)**.** *Let $\mathcal{T}$ be a recursively axiomatized theory. Then:*

(1) $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$ *interprets $\mathcal{T}$.*
(2) $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$ *proves that $\mathcal{T}$ has a model, one whose complexity is what Hajék and Pudlák call* $low\ \Sigma_0^*(\Sigma_1)$*, or LL_1.*

By a 'model' here is meant precisely what one would think is meant: A certain sort of set, arithmetically coded, of course.[81] The model is understood to come with a corresponding compositional truth-theory, that is, with notions of denotation, satisfaction, and truth for which the usual Tarskian clauses can be proved, and of course sequences will serve to code the theory of assignments.[82] That the model *is* a model of $\mathcal{T}$ amounts to its being provable, in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$, that each axiom of $\mathcal{T}$ is, in the sense of truth associated with the model, true, that is, true in the model.

To say that the model is $low\ \Sigma_0^*(\Sigma_1)$, or LL_1, means that everything that occurs in it—the formulae that determine the domain, the interpretations of the primitives of $\mathcal{L}$, and the associated notion of satisfaction—are all LL_1. I am not going to attempt to explain what '$low\ \Sigma_0^*(\Sigma_1)$' means. It doesn't really matter for our purposes—and, frankly, I don't really

[79] This is because PA is not just reflexive but *essentially* reflexive: Every extension of PA in the same language is reflexive.

[80] The proof of Theorem 4.7 rests upon the availability of the same sort of result in $\mathsf{I}\Delta_0 + \Omega_1$ (Visser, 2009c, §5; Nicolai, 2014, §4).

[81] It does not seem to be widely appreciated among philosophers how much set theory can be coded even in very weak theories of arithmetic. Everyone knows that PA is capable of talking about finite sets of numbers, but PA can talk about lots of infinite sets, too. This is because, even though PA cannot define truth for the whole of the language of arithmetic, it *can* define truth for ever larger fragments. In particular, there is a Σ_n sentence $\mathrm{Sat}_{n,\sigma}(x)$ such that $\mathsf{I}\Sigma_1$ proves the Tarski clauses for Σ_n formulae and therefore proves, for each Σ_n formula $A(x)$ the Sat-sentence: $\mathrm{Sat}_{n,\sigma}(\ulcorner A(v_0)\urcorner) \equiv A(\mathsf{val}(\sigma, 0))$. One can therefore use Σ_n formulae as codes for Σ_n-definable sets when working in $\mathsf{I}\Sigma_1$ (Hájek and Pudlák, 1993, §I.1(d), esp. p. 60, Remark 1.80).

[82] Note that this works because the model we get is, obviously, one in the natural numbers (as $\mathsf{I}\Sigma_1$ understands them), and this is true even if $\mathcal{T}$ is, say, ZFC.

understand it very well.[83] I will explain why the complexity of the model matters and why its being LL_1 is enough for the proof of Theorem 4.10. The *really* important thing is that the complexity of the model is independent of $\mathcal{T}$.

Proof of Theorem 4.10. If we are going to intepret $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$, we need to deal with three things:

(i) $\mathcal{T}$
(ii) the semantic theory for $\mathcal{L}$, including the theory of assignments
(iii) the underlying syntax, $\mathsf{I}\Sigma_1$

A significant part of the last will be no problem, since we already have $\mathsf{I}\Sigma_1$ available. But we will need to make sure that we can prove the extended induction axioms. We'll deal with that last.

The arithmetized completeness theorem tells us that $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$ can give us (i) and (ii): It interprets $\mathcal{T}$, and it gives us a a model for $\mathcal{T}$, with which we get a semantics for $\mathcal{L}$. But these aren't enough by themselves: We need to make sure that they fit together the right way. To see why, suppose $\mathcal{T}$ is $\mathsf{I}\Sigma_2$. Then "0" is a term, and among the axioms of $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_2$ that we need to interpret are these two:

$$\forall x(0 \neq \mathsf{S}x)$$
$$\mathsf{Den}_\alpha(\ulcorner 0 \urcorner, 0)$$

The first comes from $\mathsf{I}\Sigma_2$ itself; the second, from the semantics. The point to note is that the term "0" occurs in both of these and so must be interpreted the same way both times, or at least in ways that are compatible. The mere fact that $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_2)$ both interprets $\mathsf{I}\Sigma_2$ and gives us a semantics for the language of $\mathsf{I}\Sigma_2$ doesn't guarantee that. For all we know so far, the former could interpret "0" as "S0" while the latter told us that "0" denotes SSS0.

This needn't happen, however, because the two versions of the arithmetized completeness theorem are closely related. It is really the second that is more fundamental. The way you get an interpretation of $\mathcal{T}$ once you have a model of $\mathcal{T}$ is the same way you can *always* get an interpretation of $\mathcal{T}$ once you have a model of $\mathcal{T}$: You just interpret it the way the model tells you to interpret it. So if the model tells you that some term t denotes u, you translate t as 'u'. If the model tells you that some predicate $R(x, y)$ has as its extension the set S, then you translate $R(x, y)$ as meaning: $<x, y> \in S$.[84] And, of course, you restrict the quantifiers to the domain of the model. The fact that the model is a model of $\mathcal{T}$

[83] The definition is on p. 85 of Hajék and Pudlák's book, for those who would like to explore it.

[84] Note that this is all intensional: In the theory in which we are working, we'll be *given* the extension of $R(x, y)$ in a certain way, that is, by means of a certain formula; we then use that very formula to construct the translation of $R(x, y)$.

will then imply that $\mathcal{T}$'s axioms, so translated, are provably true. Which means that we've successfully interpreted $\mathcal{T}$.

What this means in our case is that the interpretation and the model do 'fit together in the right way'. If the semantic theory says that "0" denotes $S0$, then the interpretation of "0" will be "$S0$".

So that takes care of the interpretation of $\mathcal{T}$ and the interpretation of the semantics for $\mathcal{L}$. What's left is (iii), the underlying syntax, $\mathsf{I}\Sigma_1$. As noted earlier, much of that is trivial, since we're working in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$ and so have $\mathsf{I}\Sigma_1$ readily available. So if we were just trying to interpret $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$, we'd be done. What we're actually trying to interpret, however, is $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$, and so what we lack at this point—all we lack—is a demonstration that the extended induction axioms can be proven in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$, given the interpretation of $\mathcal{T}$, and of the semantics for $\mathcal{L}$, that we've already got.

It is here, then, that we need to make use of what we know about the complexity of the model and, in particular, of its associated notions of denotation, satisfaction, and truth. If the formula we were using to interpret $\mathsf{Sat}_\alpha(x)$ were, say, Σ_2, then we'd have no hope whatsoever of proving the translations of induction axioms containing $\mathsf{Sat}_\alpha(x)$ in $\mathsf{I}\Sigma_1$. But we know that everything that appears in the model—including $\mathsf{Sat}_\alpha(x)$ and its friends—is LL_1.

Now, the induction axioms we're trying to prove are those for formulae of the form $\exists v_1 \cdots \exists v_n(\phi)$, where the initial quantifers may be of any of the three available types, and ϕ is built from atomic formulae of $\mathcal{S}$ and from the translations of the expressions of the object language and atomic semantic formulae ($\mathsf{Den}_\alpha(t, \mathrm{x})$, $\mathsf{val}(\sigma, x)$, etc). All of that is LL_1, so the induction axioms we're trying to prove are thus $\Sigma_1(LL_1)$. And it just so happens that $\mathsf{I}\Sigma_1$ proves induction for $\Sigma_1(LL_1)$ formulae (Hájek and Pudlák, 1993, p. 85, Lemma 2.78). □

Proof of Theorem 4.11. The proof is essentially the same as the one just given. In $\mathsf{I}\Sigma_2$, however, we get a better bound on the complexity of the model: It is $\mathit{low}\ \Delta_2$. So the question is whether $\mathsf{I}\Sigma_n$ proves induction for $\Sigma_n(\mathit{low}\ \Delta_2)$ formulae, when $n \geq 2$. It does.[85] □

4.7. **Peano Arithmetic Is a Special Case (II).** I've remarked several times now that PA is in certain respects unrepresentative. We're now in a position to see another way in which that is so.

Corollary 4.16. $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{PA}$ *is interpretable in* PA.

[85] It is provable in $\mathsf{I}\Sigma_2$ that the $\Sigma_2(\mathit{low}\ \Delta_2)$ sets just are the Σ_2 sets (Hájek and Pudlák, 1993, p. 83, Theorem 2.71). That is much stronger than, but implies, the claim that $\mathsf{I}\Sigma_2$ has induction for $\Sigma_2(\mathit{low}\ \Delta_2)$ formulae. It is easy to generalize this result to show that the $\Sigma_n(\mathit{low}\ \Delta_2)$ sets are just the Σ_n sets and so that $\mathsf{I}\Sigma_n$ has induction for $\Sigma_n(\mathit{low}\ \Delta_2)$ formulae.

Proof. Any finite fragment of $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{PA}$ is contained in one of the theories: $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{I}\Sigma_n$ and so by Theorem 4.11 is interpretable in $\mathsf{I}\Sigma_m + \mathsf{Con}(\mathsf{I}\Sigma_n)$. But PA, being reflexive, contains every such theory. So every finite fragment of $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m]+\mathsf{PA}$ is interpretable in PA, which shows that $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{PA}$ is locally interpretable in PA. Orey's Compactness Theorem then does the rest. □

Corollary 4.17. $\mathsf{TT}_{\mathcal{A}}[\mathsf{PA}] + \mathsf{PA}$ *is interpretable in* PA.

Proof. Any finite fragment of $\mathsf{TT}_{\mathcal{A}}[\mathsf{PA}] + \mathsf{PA}$ is contained in one of the $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_m] + \mathsf{PA}$. So $\mathsf{TT}_{\mathcal{A}}[\mathsf{PA}] + \mathsf{PA}$ is locally interpretable in PA and hence is globally interpretable in PA. □

Thus, the nice pattern we had with Theorem 4.10 and Theorem 4.11 breaks down when we take PA as *object* theory.[86] This is because, as previously, there is no reason to expect $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ to be able to prove that *all* axioms of $\mathcal{T}$ are true when $\mathcal{T}$ is infinitely axiomatized.

Indeed, we can easily prove that $\mathsf{TT}_{\mathcal{A}}[\mathsf{PA}] + \mathsf{PA}$ does not prove that all axioms of PA are true.

Corollary 4.18. $\mathsf{TT}_{\mathcal{A}}[\mathsf{PA}]$ *plus "all axioms of* PA *are true" proves* $\mathsf{Con}(\mathsf{PA})$. *Indeed,* $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1]$ *plus "all axioms of* PA *are true" proves* $\mathsf{Con}(\mathsf{PA})$.

Proof. From Theorem 4.9. □

Corollary 4.19. $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathsf{PA}$ *does not prove that all axioms of* PA *are true.*

Proof. From Corollary 4.17, Corollary 4.18, and Gödel's second. □

What this means is that, once we have disentangled the syntax from the object-language, the 'happy accident' that permits $\mathsf{CT}[\mathsf{PA}]$ to prove $\mathsf{Con}(\mathsf{PA})$ is revealed as something more like a dirty trick. It is *only* because of the interaction between the extended induction principle and the theory whose consistency we are trying to prove that $\mathsf{CT}[\mathsf{PA}]$ proves $\mathsf{Con}(\mathsf{PA})$.

5. Closing

We may summarize the central results of this paper as follows.

Theorem. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$.*

(i) *If $\mathcal{T} \supseteq$ Q, then $\mathsf{CT}^-[\mathcal{T}]$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and hence is not interpretable in $\mathcal{T}$.* (Corollary 3.9)

(ii) *If $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$, then $\mathsf{CT}[\mathcal{T}]$ proves $\mathsf{Con}(\mathcal{T})$.* (Corollary 3.17)

(iii) $\mathsf{CT}[\mathsf{I}\Sigma_1]$ *proves* $\mathsf{Con}(\mathsf{PA})$ (Corollary 3.19)

(iv) $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ *is mutually interpretable with* $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$*, and so is not interpretable in $\mathcal{T}$.* (Corollary 4.8)

[86] Similar results, as one can easily see, will hold for other reflexive theories.

(v) $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ *is mutually interpretable with* $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$*, for* $n > 0$. (Theorem 4.11)

(vi) $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ *is mutually locally interpretable with* $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$. (Corollary 4.14)

Note that nothing has been said specifically about such theories as $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_1]$, and I think it would be well worth investigating them. It's of course immediate that $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is not interpretable in $\mathcal{T}$, since even $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ isn't. But is there some nice characterization of exactly how strong $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is? In general, one would suppose it is stronger than $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$. On the other hand, one would suppose that $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is weaker than $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ and, in particular, that it does not interpret $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$. So where precisely does it sit? And what of intermediate theories, like $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Delta_0] + \mathcal{T}$?

The results proven here are obviously similar to one of the central results of Visser's paper "Can We Make the Second Incompleteness Theorem Coordinate Free?"

Theorem 5.1 (Visser, 2009a, Theorem 4.1)**.** *Suppose* $\mathcal{T}$ *is sequential. Then* $\mathsf{PC}(\mathcal{T})$ *is mutually interpretable with* $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.

Here, $\mathsf{PC}(\mathcal{T})$ is the result of adding predicative second-order logic to $\mathcal{T}$. And that result gives rise to a characterization of consistency statements.

Theorem 5.2 (Visser, 2009a, Theorem 4.4)**.** $\mathsf{Con}(\mathcal{T})$ *is the unique* Π_1 *sentence* P *(modulo* $\mathsf{I}\Delta_0 + \mathsf{exp}$*-provable equivalence) such that* $\mathsf{PC}(\mathcal{T})$ *is mutually interpretable with* $\mathsf{Q} + P$.

A similar characterization is forthcoming from our results.[87]

Theorem 5.3 (Nicolai, 2014, Proposition 3)**.** *If* $\mathcal{T}$ *is finitely axiomatizable, then* $\mathsf{Con}(\mathcal{T})$ *is the unique* Π_1 *sentence* P *(modulo* $\mathsf{I}\Delta_0 + \mathsf{exp}$*-provable equivalence) such that* $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ *is mutually interpretable with* $\mathsf{Q} + P$.

Unfortunately, this result, unlike Visser's, is limited to finitely axiomatized theories. But it is natural to wonder whether we might not be able to handle non-finitely axiomatized theories by proving something like:

Conjecture 5.4. *Let* $\mathcal{T}$ *be a consistent theory in* $\mathcal{L}$*. Then* $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *plus "all axioms of* $\mathcal{T}$ *are true" is mutually interpretable with* $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.

And if we could prove that, then perhaps we would also be able to prove:

Conjecture 5.5. $\mathsf{Con}(\mathcal{T})$ *is the unique* Π_1 *sentence* P *(modulo* $I\Delta_0 + \mathsf{exp}$*-provable equivalence) such that* $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ + *"all axioms of* $\mathcal{T}$ *are true" is mutually interpretable with* $\mathsf{Q} + P$.

This suspicion has now been borne out. Nicolai (2014, Corollaries 6–8) has recently proven both Conjecture 5.4 and Conjecture 5.5.

[87] This was conjectured in an earlier version of this paper.

It also seems worth asking whether we get similar results for the case in which induction has been extended:

Conjecture 5.6. *Let $\mathcal{T}$ be a consistent theory in $\mathcal{L}$. Then $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1]$ plus "all axioms of $\mathcal{T}$ are true" is mutually interpretable with $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$.*

Conjecture 5.7. $\mathsf{Con}(\mathcal{T})$ *is the unique Π_1 sentence P (modulo $I\Delta_0 + \mathsf{exp}$-provable equivalence) such that $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1]$ plus "all axioms of $\mathcal{T}$ are true" is mutually interpretable with $\mathsf{I}\Sigma_1 + P$.*

These remain open as does, of course, the question whether some restriction on the complexity of the formula specifying the axioms is needed.

Even when provable, however, such results do not serve the central purpose Visser wanted his to serve, which was to give a characterization of consistency statements that is independent of any issues involving coding. One can take some steps to minimize the extent to which the results proven here do depend upon coding. As mentioned earlier, the syntactic part of e.g. $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ could be (and maybe even should be) formulated using a theory of concatenation, say, the theory TC due to Grzegorczyk (2005), which is mutually interpretable with Q (Visser, 2009b). We can then add the obvious sorts of induction axioms to TC, thus arriving at theories we can use in place of $\mathsf{I}\Sigma_n$ and PA. The proofs of the results just summarized will transfer smoothly to such a framework. But, as Visser pointed out to me, there are still many choices to be made about, for example, whether we are using prefix, infix, or postfix notation, exactly what we take a variable to be, and so forth. So some seemingly inessential choices still seem to need making.

Still, the results proven here make it clear how close the connection is between truth and consistency and also, in light of Visser's results, between truth and predicative comprehension.[88] They should also make it clear that theories of truth in which the syntactic theory is disentangled from the object theory are of substantial technical utility, at least, since they allow us to formulate and prove a number of nice results that are otherwise unavailable.

My own view is that these results also have significant philosophical implications, but that matter will have to be left to another occasion (Heck, 2014b).[89]

[88] Mostowski (1950) seems to have been the first to realize that there is some such connection. Philosophers' appreciation of it is largely due to Parsons (1974). Van Wesep (2013) has returned to the issue recently.

[89] This paper is one of several to emerge from an earlier manuscript, "The Strength of Truth Theories" (Heck, 2009). That paper not only contained the technical material presented here but a discussion of its philosophical significance and its bearing upon questions about the nature of what Tarski called "essential richness". For reasons that ought to have been obvious to me much sooner, that paper became unmanageable and

References

Beklemishev, L. D. (2005). 'Reflection principles and provability algebras in formal arithmetic', *Russian Mathematical Surveys* 60: 197–268. 41

Burgess, J. P. (2005). *Fixing Frege*. Princeton NJ, Princeton University Press. 7, 11, 12

Buss, S. R. (1986). *Bounded Arithmetic*. Napoli, Bibliopolis. 8

Corcoran, J., Frank, W., and Maloney, M. (1974). 'String theory', *Journal of Symbolic Logic* 39: 625–37. 33

Craig, W. and Vaught, R. L. (1958). 'Finite axiomatizability using additional predicates', *Journal of Symbolic Logic* 23: 289–308. 34, 36, 37

Enayat, A. and Visser, A. (2012). New constructions of full satisfaction classes. Manuscript available at http://dspace.library.uu.nl/bitstream/handle/1874/266885/preprint303.pdf. 26, 27, 29

—— (2014). Full satisfaction classes in a general setting (Part I). Unpublished manuscript. 26

Feferman, S. (1960). 'Arithmetization of metamathematics in a general setting', *Fundamenta Mathematicae* 49: 35–92. 3, 5, 6, 13

Fischer, M. (2009). 'Minimal truth and interpretability', *Review of Symbolic Logic* 2: 799–815. 18

—— (2014). 'Truth and speed-up', *Review of Symbolic Logic* 7: 319–40. 18

had to be split into several pieces. That, and the pressing need to finish two books on Frege, is what explains why this paper has taken so long to appear.

Thanks to Volker Halbach and Jeffrey Ketland for helpful discussions early in the history of my work on this topic, and to Josh Schechter for equally helpful whiteboard sessions later on. Comments on the earlier manscript from Cezary Cieśliński and Ali Enayat were also very useful, as were comments from two anonymous referees on this version.

Talks incorporating some of these ideas were given at a conference on philosophical logic, organized by Delia Graff Fara, that was held at Princeton University in April 2009; at the New England Logic and Language Colloquium and at the Philosophy of Mathematics Seminar at Oxford University, both in May 2011; and at a meeting of the Logic Group at the University of Connecticut, in April 2012. Thanks to everyone present for their questions and comments, especially J. C. Beall, John P. Burgess, Carlo Nicolai, Hartry Field, Graham Leigh, Daniel Isaacson, Charles Parsons, Agustín Rayo, and Lionel Shapiro, as well as Volker and Josh, again. Special thanks to my commentator at Princeton, Josh Dever, whose comments were insightful and lucid, as well as helpful.

I owe the greatest debt, however, to Albert Visser. Just as my ideas were starting to come together, Albert taught me Theorem 3.5 and its attendant corollaries. That changed the direction of my project and helped make it clear to me what strategy I should try to use to prove the sorts of limitative results presented in Section 4.5. Since then, Albert has read many drafts of this paper, provided extensive feedback, and cheerfully answered a nearly unending stream of questions. (Ali also answered a lot of questions toward the end of the process.) This paper could never have been written without Albert's generous assistance, for which I am extremely grateful.

Grzegorczyk, A. (2005). ‘Undecidability without arithmetization’, *Studia Logica* 79: 163–230. 33, 47

Hájek, P. (2007). ‘Mathematical fuzzy logic and natural numbers’, *Fundamenta Informaticae* 81: 155–63. 8

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First-order Arithmetic*. New York, Springer-Verlag. 7, 8, 9, 10, 12, 13, 28, 29, 41, 42, 44

Halbach, V. (2011). *Axiomatic Theories of Truth*. Oxford, Oxford University Press. 2

Heck, R. G. (2005). ‘Reason and language’, in C. MacDonald and G. MacDonald (eds.), *McDowell and His Critics*. Oxford, Blackwells, 22–45. 36

—— (2007). ‘Meaning and truth-conditions’, in D. Greimann and G. Siegwart (eds.), *Truth and Speech Acts: Studies in the Philosophy of Language*. New York, Routledge, 349–76. 36

—— (2009). The strength of truth-theories. Unpublished manuscript available at `http://rgheck.frege.org/pdf/unpublished/StrengthOfTruthTheories.pdf`. 38, 39, 47

—— (2014a). ‘Frege arithmetic and “everyday mathematics”’, *Philosophia Mathematica* 22: 279–307. 8

—— (2014b). The logical strength of compositional principles. Manuscript. 32, 40, 47

—— (2014c). What is essential richness? Manuscript. 1

Kleene, S. (1952). ‘Finite axiomatizability of theories in the predicate calculus using additional predicate symbols’, *Memoirs of the American Mathematical Society* 10: 27–68. 36

Kotlarski, H., Krajewski, S., and Lachlan, A. H. (1981). ‘Construction of satisfaction classes for nonstandard models’, *Canadian Mathematical Bulletin* 24: 283–93. 2

Kotlarski, H. (1986). ‘Bounded induction and satisfaction classes’, *Zeitschrift für Mathematische Logik* 32: 531–544. 29

Leigh, G. E. (2013). Conservativity for theories of compositional truth via cut elimination. Forthcoming in the *Journal of Symbolic Logic*. Manuscript available at `http://arxiv.org/abs/1308.0168`. 2

Leigh, G. E. and Nicolai, C. (2013). ‘Axiomatic truth, syntax and metatheoretic reasoning’, *The Review of Symbolic Logic* 6: 613–36. 17, 39, 40

Mostowski, A. (1950). ‘Some impredicative definitions in the axiomatic set-theory’, *Fundamenta Mathematicae* 37: 111–24. 47

—— (1952). ‘On models of axiomatic systems’, *Fundamenta Mathematicae* 39: 133–58. 6

Nelson, E. (1986). *Predicative Arithmetic*. Mathematical Notes 32. Princeton NJ, Princeton University Press. 7

Nicolai, C. (2014). A note on typed truth and consistency assertions. Forthcoming in the *Journal of Philosophical Logic*. 8, 38, 42, 46

Orey, S. (1961). 'Relative interpretations', *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* 7: 146–53. 6

Parsons, C. (1974). 'Sets and classes', *Noûs* 8: 1–12. 47

Pudlák, P. (1985). 'Cuts, consistency statements and interpretations', *Journal of Symbolic Logic* 50: 423–41. 13

Quine, W. V. O. (1946). 'Concatenation as a basis for arithmetic', *Journal of Symbolic Logic* 11: 105–14. 33

Švejdar, V. (2007). 'An interpretation of Robinson arithmetic in Grzegorczyk's weaker variant', *Fundamenta Informaticae* 81: 347–54. 8

Tarski, A. (1944). 'The semantic conception of truth and the foundations of semantics', *Philosophy and Phenomenological Research* 4: 341–75. 1

—— (1953). 'A general method in proofs of undecidability', in Tarski et al. 1953, 1–35. 4

—— (1958). 'The concept of truth in formalized languages', in J. Corcoran (ed.), *Logic, Semantics, and Metamathematics*. Indianapolis, Hackett, 152–278. 31, 33

Tarski, A., Mostowski, A., and Robinson, A. (1953). *Undecidable Theories*. Amsterdam, North-Holland Publishing. 5, 50

Van Wesep, R. A. (2013). 'Satisfaction relations for proper classes: Applications in logic and set theory', *Journal of Symbolic Logic* 78: 245–68. 47

Visser, A. (2006). 'Categories of theories and interpretations', in A. Enayat and I. Kalantari (eds.), *Logic in Tehran: Proceedings of the Workshop and Conference on Logic, Algebra and Arithmetic, Held October 18–22, 2003*. Wellesley MA, A. K. Peters, 284–341. 5

—— (1991). 'The formalization of interpretability', *Studia Logica* 50: 81–105. 6, 7

—— (1992). 'An inside view of EXP', *Journal of Symbolic Logic* 57: 131–65. 23

—— (2008). 'Pairs, sets and sequences in first-order theories', *Archive for Mathematical Logic* 47: 299–326. 15

—— (2009a). 'Can we make the second incompleteness theorem coordinate free?', *Journal of Logic and Computation* 21: 543–60. 13, 46

—— (2009b). 'Growing commas: A study of sequentiality and concatenation', *Notre Dame Journal of Formal Logic* 50: 61–85. 47

—— (2009c). 'The predicative Frege hierarchy', *Annals of Pure and Applied Logic* 160: 129–53. 38, 42

Wang, H. (1952). 'Truth definitions and consistency proofs', *Transactions of the American Mathematical Society* 73: 243–275. 16, 29

Wilkie, A. J. and Paris, J. B. (1987). 'On the scheme of induction for bounded arithmetic formulas', *Annals of Pure and Applied Logic* 35: 261–302. 7, 8, 13

Department of Philosophy, Brown University, Providence RI 02912