THE STRENGTH OF TRUTH-THEORIES

RICHARD G HECK JR.

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 large 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.

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. It is tempting, therefore, to want to use this fact to interpret Tarski's famous claim in "The Semantic Conception of Truth" that the metalanguage must be 'essentially richer' than the object langauge (Tarski, 1944, p. 354). As we shall see, however, that would be to confuse a question about *expressive power* with a question about *logical strength*. It is possible to formalize a materially adequate theory of truth for the language of set-theory in a meta-theory that is as weak as it is *a priori* possible for it to be: one interpretable in Robinson arithmetic. If so, then

---

This manuscript dates from about 2009, with some significant updates having been made around 2011. Around then, however, I decided that the paper was becoming unmanageable and that I was trying to do too many things in it. I have therefore exploded the paper into several pieces, which will be published separately.

I am putting this version on the web simply because it has been cited in a few different places and so should be publicly available. I should have done this a long time ago. I should have finished the paper a long time ago. But since 2010, my time has largely been devoted to finishing my two books on Frege, and even this draft remains a mess. Terminology and notation are inconsistent, and some of the proofs aren't quite right. So, *caveat lector*.

Tarski's claim about essential richness cannot concern logical strength. Not, at least, if it is to have any hope of being true.

One might yet wonder, though, if there is not some way of understanding what a truth-theory buys us in terms of logical strength. And here is where we meet the central motivation for the present paper. It seems to be widely believed that a truth-theory *by itself* has no logical power at all. The proof of *PA*'s consistency mentioned above 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 a conservative extension of *PA*. Of course, there are other ways of comparing the strength of theories. In particular, it is compatible with a theory $\mathcal{T}$'s being a conservative extension of another theory $\mathcal{U}$ that $\mathcal{T}$ should not be interpretable in $\mathcal{U}$. But if we take *PA* and add a truth-theory, then the result *is* interpretable in *PA* so long as we do *not* extend induction. So that might seem to seal it: Truth-theories have no logical strength on their own.

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, 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 want to know how strong truth-theories are on their own, 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 $\mathcal{T}$?

Once we have reframed the investigation in these terms, then several sorts of questions become natural:

(1) What is the weakest theory $\mathcal{T}$ such that the result of adding a truth-theory to $\mathcal{T}$ yields a materially adequate theory of truth for the language of $\mathcal{T}$?
(2) What, in general, is the strength of such a theory, as compared to that of $\mathcal{T}$, if we do not extend whatever induction axioms are present in $\mathcal{T}$ to permit semantic vocabulary?
(3) What happens if we do extend $\mathcal{T}$'s induction axioms? In particular, for which theories $\mathcal{T}$ does the result of adding the truth-theoretic axioms and extending $\mathcal{T}$'s induction scheme allow us to prove the consistency of $\mathcal{T}$?
(4) What is the strength of the theory mentioned in (3), as compared to that of the 'base' theory $\mathcal{T}$?

Much turns on precisely how we formulate the truth-theory and on what sorts of base theories are in question. In particular, we shall see that the usual way of 'adding a truth-theory', though it allows a nice answer to (2), gives us only dissatisfying answers to (3) and (4). But there is a different, and older, way to proceed—Tarski's original way—that allows

answers to these questions that are about as elegant as one could hope they would be.

The plan for the paper is as follows. In an effort to make the discussion as accessible as possible I will quickly introduce in Section 2 some of the central concepts from logic that we shall be using. In Section 3, we'll discuss the usual way of 'adding a truth-theory' and see that there is a materially adequate and fully compositional theory of truth for the language of arithmetic that is about as weak as it could be. Section 4 introduces some machinery from the study of sub-systems of *PA* which may be less familiar. This is applied in Section 5, where we will get our first characterization of the strength of truth-theories and see a first respect in which *PA* is a special case, as well as discover some annoying limitations of the approach we will have been pursuing to that point. Section 6 explores a different way of 'adding a truth-theory' and gives nice answers to the questions above. We'll also see another, more impressive respect in which *PA* is a special case. Finally, Section 7 briefly considers how our results bear upon some philosophical questions about the role truth-theories play in semantic consistency proofs.

2. Logical Preliminaries

2.1. **Interpretatability.** The *languages* in which we'll be interested here are first-order languages, constructed from atomic 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 atomic expressions. It is convenient to identify a language with the set of its atomic expressions, together with some indication of their logical type, that is, with what is sometimes called the 'signature' of the language.

A *theory* here is always a recursively axiomatized theory, unless otherwise stated, and, officially, we understand the notion in an intensional sense: A theory is not a set of axioms but a 'presentation' of a set of axioms. Formally, a theory can be understood as given by a formula in one free variable, where the axioms of the theory are the sentences of whose Gödel numbers that formula is true. When a theory has only finitely many axioms, the distinction between intensional and extensional conceptions more or less lapses. But it does matter, in general, as Feferman (1960) made abundantly clear.

A theory is 'stated in' a language.

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 the other, then $\ulcorner A \vee \neg A \urcorner$ will be a

theorem of the one but not the other; this is true even if the (non-logical) axioms of the two theories are the same.

If the language of one theory contains that of the other, then one way to compare them is to ask if 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, which was first explored in a systematic way by Tarski, Mostowski, and Robinson (1953), although the basic idea is much older.

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*[1] 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) atomic expressions of $\mathcal{L}_\mathcal{T}$ in terms of those of $\mathcal{L}_\mathcal{B}$[2] 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: $\forall x(\delta(x) \to \phi^*(x))$, 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 atomic term t of $\mathcal{L}_\mathcal{T}$, and of the closure condition

$$\forall x_1 \cdots x_n(\delta(x_1) \wedge \cdots \wedge \delta(x_n) \to \delta(f^*(x_1, \ldots, x_n)))$$

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 $\mathcal{T}$, 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,[3] then all of this will be provable in $\mathcal{B}$ and $\mathcal{T}$ themselves. So, in particular, $\mathcal{T}$ will prove $\mathrm{Con}(\mathcal{B}) \to \mathrm{Con}(\mathcal{T})$ and so cannot prove $\mathrm{Con}(\mathcal{B})$, though $\mathcal{B}$ might well prove $\mathrm{Con}(\mathcal{T})$.

Note that interpretability is transitive and reflexive.

---

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

[2] 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.

[3] Facts concerning interpretability can generally be verified in the theory known as $I\Delta_0 + \omega_1$, for which see below.

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 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 atomic vocabulary does not have to be the identity function. But of course it can be, and in that case the interpretation can 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 form.

Now, a couple definitions that apply (sensibly) only to non-finitely axiomatized theories.

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

Local interpretability obviously follows from interpretability, which is also known as 'global' interpretability. The converse is not true. 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.

As said above, Peano arithmetic is going to turn out to be something of a special case. This is because *PA* is *reflexive* (Mostowski, 1952), in the following sense.

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

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 Theorem)**.** *Suppose that $\mathcal{T}$ is locally intepretable in $\mathcal{B}$ and that $\mathcal{B}$ is reflexive. Then $\mathcal{T}$ is (globally) interpretable in $\mathcal{B}$.*

The proof of this result was first published in Feferman's classic paper "The Arithmetization of Metamathematics in a General Setting", which was also, of course, where the 'unexpected phenomena' just mentioned first appeared.

2.2. **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. Let's meet them.

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

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

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 of **Q**, $\{0, S, +, \times, <\}$, is what we shall call 'the language of arithmetic' and denote: $\mathcal{A}$.

Q is in many ways extremely weak. It fails to prove such obvious facts as that $x \neq Sx$. But it is in other ways strong. For our purposes, the crucial fact is that **Q** is strong enough to allow us to do Gödel numbering and therefore some very basic syntax. For example, **Q** will allow us to say and prove such things as that '$0 = S0 \wedge 0 = SS0$' is the conjunction of '$0 = S0$' and '$0 = SS0$'. Here's the short version: **Q** is terrible at proving generalizations, but it's very good at proving particular facts.

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 $\exists x\phi$ (resp., $\forall x\phi$), where ϕ is Δ_0. A formula is Σ_n (resp., Π_n) if it is $\exists x\phi$ (resp., $\forall x\phi$), where ϕ is Π_{n-1} (resp., Σ_{n-1}).

We can now say precisely how good **Q** is at proving particular facts: **Q** proves all true Σ_1 sentences of the language of arithmetic.

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

$$A(0) \wedge \forall x(A(x) \rightarrow A(Sx) \rightarrow \forall x(A(x)),$$

where $A(x)$ is any formula at all. The theory $I\Theta$ is **Q** plus induction for formulae in the set Θ: So $A(x)$ has to be in Θ. Thus, $I\Delta_0$ is **Q** plus induction for Δ_0 formulae, and $I\Sigma_1$ is **Q** plus induction for Σ_1 formulae. $I\Delta_0$ is in one sense clearly stronger than **Q**: It proves lots of important generalizations about the natural numbers. But in another sense it is

still a very weak theory: It is interpretable in **Q**.[4] Another respect in which $I\Delta_0$ is weak is that, although one can define the relation $y = 2^x$ by means of a Δ_0 formula $\exp(x, y)$, we cannot prove in $I\Delta_0$ that exponentiation is total; that is, we cannot prove: $\forall x \exists y(\exp(x, y))$. The obvious proof uses by induction on $\exists y(\exp(x, y))$, which is Σ_1. But for that very reason, the totality of exponentiation is provable in $I\Sigma_1$, as is the totality of every other primitive recursive function.[5] So $I\Sigma_1$ is much stronger than $I\Delta_0$: Indeed, $I\Sigma_1$ proves $\mathrm{Con}(I\Delta_0)$.

The final theory we shall need is known as $I\Delta_0 + \omega_1$. Here, $\omega_1(x)$ is a certain function that, like 2^x, is Δ_0-definable but not $I\Delta_0$-provably total. The precise definition varies between authors, but one definition (Visser, 1991, p. 83) is:

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

where $|x|$ is the least y such that $2^{Sy} > Sx$. As said, the relation $y = \omega_1(x)$ can be defined by a Δ_0 formula $\Omega_1(x, y)$, and $I\Delta_0 + \omega_1$ is then $I\Delta_0$ 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".[6] And, like $I\Delta_0$, it is interpretable in **Q** (Hájek and Pudlák, 1993, p. 367).

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, P, A, and M in place of 0, S, $+$, and $\times$. We shall therefore want axioms asserting that there is a unique zero, and that P, A, and M are function-like:

Z	$\exists x(Zx \wedge \forall y(Zy \rightarrow x = y))$
P	$\forall x \exists y(Pxy \wedge \forall z(Pxz \rightarrow y = z))$
A	$\forall x \forall y \forall z \exists z(Axyz \wedge \forall w(Axyw \rightarrow z = w))$
M	$\forall x \forall y \forall z \exists z(Axyz \wedge \forall w(Axyw \rightarrow z = w))$

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	$\neg Px0$
QR2	$Pxz \wedge Pyz \rightarrow x = y$

[4] That $I\Delta_0$ is *locally* interpretable in **Q** was first proven by Edward Nelson (1986). That it is globally interpretable was proven by Alex Wilkie (Wilkie and Paris, 1987). The proof is discussed both by Hájek and Pudlák (1993, pp. 366–70) and by Burgess (2005, §2.2).

[5] Indeed, $I\Sigma_1$ is proof-theoretically equivalent to primitive recursive arithmetic.

[6] Wilkie and Paris (1987) were the first to recognize the importance of $I\Delta_0 + \omega_1$. 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.

QR3 $\quad Ax0x$

QR4 $\quad (Pyz \wedge Axzu) \wedge (Axyw \wedge Pwv) \rightarrow u = v$

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

It should be clear that $\boldsymbol{Q}$ and Q_R are interpretable in one another, in a very straightforward way. Similar things can be said about the other theories mentioned.

3. Theories of Truth

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 $\boldsymbol{Q}$, 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 $\mathrm{lh}(s, h)$ and $\mathrm{val}(s, n, x)$ for which $\mathcal{T}$ proves:[7]

$$\begin{aligned}
&\exists s(\mathrm{lh}(s, 0)) \\
\forall s \forall n\{\mathrm{lh}(s, n) \rightarrow &\forall m < n \exists x(\mathrm{val}(s, n, x))\} \\
\forall s \forall n\{\mathrm{lh}(s, n) \rightarrow &\forall y \exists t[\mathrm{lh}(t, Sn) \wedge \\
&\qquad \forall z \forall k < n(\mathrm{val}(s, k, z) \equiv \mathrm{val}(t, k, z)) \wedge \\
&\qquad \mathrm{val}(t, n, z)]\}
\end{aligned}$$

Here, $\mathrm{lh}(s, n)$ means: s is a sequence of length n; $\mathrm{val}(s, n, x)$ means: the $(n + 1)$-st element of s is x. So the second of the principles says that every sequence of length n has an element at each position below n; the third says that each sequence can be extended by appending an arbitrary element of the domain; the first assures us that there is a 'null' sequence with which we can begin. We shall use '<>' as a term denoting one of the null sequences whose existence is so guaranteed.[8]

$\boldsymbol{Q}$ is not sequential, but there are lots of sequential theories that are interpretable in $\boldsymbol{Q}$. For example, $I\Delta_0$ is sequential, and it is interpretable in $\boldsymbol{Q}$. More importantly, for our purposes, we can simply add a theory of sequences to $\boldsymbol{Q}$, by adding new predicates $\mathrm{lh}(s, n)$ and $\mathrm{val}(s, n, x)$, subject to the principles that characterize sequential theories. This new theory, which we might call Q_{seq}, is interpretable in $\boldsymbol{Q}$, since it is obviously intepretable in any sequential theory. This fact will allow us to extend

[7] We can take 's is a sequence' to be defined as: $\exists n(\mathrm{lh}(s, n))$.

[8] In the cases in which we are interested, there generally will be such a term in the language. If not, then we can conservatively extend whatever theory we are employing by adding such a term, subject to the axiom: $\mathrm{lh}(<>, 0)$.

our main results to Q, even though they do not apply to Q directly. Note, morever, that every sequential theory interprets Q.[9]

It should be obvious that we can easily allow $\text{val}(s, n, x)$ 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 our sequences are infinite.

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	$\text{Den}_\sigma(v_i, x) \equiv \text{val}(\sigma, i, x)$, where v_i is the i^{th} variable
$=$	$\text{Sat}_\sigma(\ulcorner t = u \urcorner) \equiv \exists x \exists y [\text{Den}_\sigma(t, x) \wedge \text{Den}_\sigma(u, y) \wedge x = y]$
$\neg$	$\text{Sat}_\sigma(\ulcorner \neg A \urcorner) \equiv \neg \text{Sat}_\sigma(A)$
$\wedge$	$\text{Sat}_\sigma(\ulcorner A \wedge B \urcorner) \equiv \text{Sat}_\sigma(A) \wedge \text{Sat}_\sigma(B)$
$\forall$	$\text{Sat}_\sigma(\ulcorner \forall v_i A(v_i) \urcorner) \equiv \forall \tau [\tau \overset{i}{\sim} \sigma \rightarrow \text{Sat}_\sigma(\ulcorner A(v_i) \urcorner)]$

And similarly for the other logical constants.[10] Here, '$\text{Den}_\sigma(t, x)$' means: t denotes x with respect to the sequence σ; '$\text{Sat}_\sigma(A)$' 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.:[11]

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

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

0	$\text{Den}_\sigma(\text{'0'}, x) \equiv x = 0$
S	$\text{Den}_\sigma(\ulcorner St \urcorner, x) \equiv \exists y (\text{Den}_\sigma(t, y) \wedge y = Sx)$
$+$	$\text{Den}_\sigma(\ulcorner t + u \urcorner, x) \equiv \exists y \exists z [\text{Den}_\sigma(t, y) \wedge \text{Den}_\sigma(u, z) \wedge x = y + z]$
$\times$	$\text{Den}_\sigma(\ulcorner t \times u \urcorner, x) \equiv \exists y \exists z [\text{Den}_\sigma(t, y) \wedge \text{Den}_\sigma(u, z) \wedge x = y \times z]$
$<$	$\text{Sat}_\sigma(\ulcorner t < u \urcorner) \equiv \exists y \exists z [\text{Den}_\sigma(t, y) \wedge \text{Den}_\sigma(u, z) \wedge y < z]$

The pattern should be clear.[12]

In the case of the language of arithmetic, there are at least two simplifications one often meets in practice. First, denotation is actually definable in arithmetic in such a way that the clauses involving it can be proven, so those clauses are often regarded as not really necessary.

---

[9] For lots of details on sequential theories, including the facts mentioned here, see Visser's "Pairs, Sets, and Sequences in First-order Theories" (Visser, 2008). 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.

[10] 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.

[11] It will be important below that this can be made to be Δ_1. One way to see that it can be is to note that $\text{lh}(\sigma,)$ and $\text{val}(\sigma, i)$ are what Boolos (1993, pp. 24–7) calls 'Σ pterms', and Δ_1 formulas are closed under substitution of Σ pterms.

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

Second, one can 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. We shall avoid these simplifications here, however. Both simplificiations are specific to the language of arithmetic and are not available in general. We want our results to extend smoothly and naturally to other cases, such as the language of set theory.

Finally, then, we need to define the notion of truth itself. Tarski, as is familiar, defines truth in terms of satisfaction by every sequence, thus:

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

Where we are discussing theories of truth over weak arithmetics, however, 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 was before it.[13] For this reason, it is sometimes preferable to use an alternate definition:

T: $\quad \mathrm{T}(A) \equiv A$ is a sentence $\wedge \mathrm{Sat}_{\langle\rangle}(A)$

So, on this definition, truth is satisfaction by the null sequence. As it happens, however, it is actually better, for our purposes, to use the original definition, so we shall stick with it.

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

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

Note that $\mathcal{T}^T$ does not extend any induction scheme that might be present in $\mathcal{T}$. There is no real chance, then, that $\mathcal{T}^T$ is going to prove the consistency of $\mathcal{T}$. So one might suspect that $\mathcal{T}^T$ would logically be no stronger than $\mathcal{T}$. If so, then, as we shall see, one would suspect wrongly, at least in general. But we shall not be ready to prove that until Section 5.

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

Lemma 3.1. *$\mathcal{T}^T$ is a materially adequate, fully compositional theory of truth for the language of $\mathcal{T}$. In particular: For each sentence A in the language of $\mathcal{T}$, $\mathcal{T}^T$ proves: $\boldsymbol{T}(\ulcorner A \urcorner) \equiv A$.*

Proof. A rigorous proof would be by induction on the complexity of sentences of $\mathcal{L}$. But this should be fairly obvious. A little experimentation

[13] Thanks to Cezary Cieśliński for bringing this issue to my attention. As one example of the problem, if we use Tarski's definition, then $\mathcal{T}^T$ (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 $\forall \sigma \mathrm{Sat}_\sigma(A)$ iff $\exists \sigma \mathrm{Sat}_\sigma(A)$, and that in turn is normally proven by an induction that may not be formalizable in $\mathcal{T}^T$.

will reveal that proofs of 'T-sentences' need no more than is available in Q_{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: $\mathcal{T}^T$ 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 $\mathcal{T}$. So $\mathcal{T}^T$ is always expressively more powerful than $\mathcal{T}$.

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

Corollary 3.2. *$\mathcal{T}^T$ proves, of each axiom of $\mathcal{T}$, that it is true.*

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

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

Corollary 3.3. *Let $\mathcal{T}$ be a finitely axiomatized sequential theory. Then $\mathcal{T}^T$ proves the formalization of "all axioms of $\mathcal{T}$ are true".*[14]

Note the contrast with Corollary 3.2: If $\mathcal{T}$ is infintely axiomatized, there is no reason whatsoever to suspect that $\mathcal{T}^T$ will prove that *all* axioms of $\mathcal{T}$ are true, although it does prove that *each* axiom of $\mathcal{T}$ is true. Indeed, we have the following.

Proposition 3.4. *PA^T does* not *prove that all axioms of PA are true.*

This follows from Corollary 5.10, to be proven below, and it begins to illustrate one of the senses in which *PA* is a special case.

3.2. **Object-language *versus* Object-theory.** I suspect that Lemma 3.1 will have surprised at least some people.[15] If it didn't, then maybe this will.[16]

[14] Since $\mathcal{T}$ is sequential, it interprets **Q**, which means that we can develop enough syntax in $\mathcal{T}$ to allow us to formalize "all axioms of $\mathcal{T}$ are true".

[15] I'm cheating. I've already seen it surprise a fair number of people.

[16] Something close to this result is present in Wang's paper "Truth Definitions and Consistency" (Wang, 1952), though Wang is interested in definitions of truth, whereas we are interested in theories of truth. What Wang shows (see his theorem 11) is that a theory he calls S_2, which is basically adjunctive set theory (see below) plus predicative second-order quantification plus separation, defines truth for the language of set theory. As he notes, the result could be improved "if we could develop number theory in S_2 without using" separation. At the time, it was not known that this could be done, but the interpretability of **Q** in adjunctive set theory would soon be proven by Tarski, Mostowski, and Robinson (1953).

In any event, I do not claim originality for this result. As we shall see, it was to become fairly well known but a few years after Wang's work. But it has not been widely appreciated among philosophers.

Corollary 3.5. *${Q_{\text{seq}}}^T$ is a materially adequate, fully compositional theory of truth for the language of arithmetic.*

As mentioned earlier, it is at least tempting to try to interpret Tarski's claim that the metalanguage must be 'essentially richer' than the object langauge (Tarski, 1944, p. 354) in terms of logical strength.[17] Some people seem to think that having a theory of truth for the language of arithmetic means being able to prove that *PA* is consistent. Better informed people know that you can have a theory of truth without extending induction. But even some of them seem to think that a theory of truth for the language of arithmetic must be at least as strong as *PA*, and that this is something we learned from Tarski. What Corollary 3.5 shows is that this is just false. If that's surprising, it's probably because one is thinking of theories of truth as if their subject-matter were not *languages* but *theories*. But there is no such thing as 'a truth-theory for *PA*'. There is only a theory of truth for the *language* of *PA*, that is, for the language of arithmetic. The question whether such a theory is materially adequate is the question whether it allows us to prove the T-sentence for each sentence of the *language* it concerns. And ${Q_{\text{seq}}}^T$ is perfectly capable of proving all the T-sentences for the language of *PA*, that is, for the language of arithmetic.

The point is not specific to arithmetic. Consider, for example, the theory known as 'adjunctive set theory' or, following Visser, '*WS*' (for 'weak' set theory). Its axioms are:

$$\begin{aligned} \text{Null set:} \quad & \exists x \forall y (y \notin x) \\ \text{Adjunction:} \quad & \forall x \forall y \exists z \forall w (w \in z \equiv w \in x \vee w = y) \end{aligned}$$

WS is interpretable in *Q*, and it is, in a sense, the weakest sequential theory (Visser, 2008). But now consider WS^T. It is easy to see that WS^T proves a T-sentence for each sentence of the language of set theory. That is: WS^T is a materially adequate theory of truth for the language of set theory. So, if you were tempted to say that a theory of truth for *ZF* cannot be developed in any theory weaker than *ZF* itself, you might want to reconsider.

One might think I am being uncharitable. There is a perfectly natural interpretation of what people mean when they speak of 'a theory of truth for *PA*'. What they mean is: A materially adequate theory of truth for the language of arithmetic in which it is possible to prove that each of the axioms of *PA* is true.[18] And if *that's* what a theory of truth for *PA* is, then it's true that a materially adequate theory of truth for *PA* can only be developed in a theory at least as strong as *PA*. But this fact is completely trivial and isn't anything we needed Tarski to teach us. Each

[17] See the exchange between DeVidi and Solomon (1999) and Ray (2005).

[18] This is what Wang (1952, p. 244) calls a 'normal' truth-theory.

axiom of *PA* will follow from (i) its T-sentence and (ii) the statement that it is true.

$$\begin{aligned} \mathrm{T}(A) \equiv A \quad & \text{T-sentence} \\ \mathrm{T}(A) \quad & \text{Each axiom is true} \\ A \quad & \text{Logic} \end{aligned}$$

The lesson I propose we should learn from Corollary 3.5 is thus this: Tarski's theorem on the indefinability of truth *has nothing to do with logical strength*. It has all and only to do with *expressive power*.[19] This should have been obvious already. The reason you cannot develop a truth-theory for the language of *PA* inside *PA* itself is not that *PA* isn't strong enough. Tarski's indefinability theorem applies even to 'true arithmetic'—to the theory whose 'axioms' are all the arithmetical truths—and true arithmetic is not only stronger than *ZFC* but is stronger than *any* consistent formal theory.[20] But Corollary 3.5 reinforces this point, since it tells us that you can develop a theory of truth for the language of *PA* in ${Q_{\text{seq}}}^T$, a theory whose consistency is provable not just in *PA* but already in $I\Sigma_1$, and in weaker theories still.[21] What distinguishes ${Q_{\text{seq}}}^T$ from *PA* is the fact that the the language of the former is more expressive than the language of the latter, in the precise sense that there are sets that are definable in ${Q_{\text{seq}}}^T$ that are not definable in *PA*. The set of (Gödel numbers of) true sentences of the language of arithmetic is the salient example.

If this point has not been widely appreciated, Tarski is partly to blame. The central task of section 4 of "The Concept of Truth in Formalized Languages" is to explain how to generalize the definition of truth that Tarski had given (in the previous section of that paper) for the language of the calculus of classes. And so Tarski first "undertake[s]. . . the construction of a corresponding meta-language and the establishment of a meta-theory which forms the proper field of investigation" (Tarski, 1958, p. 210). After explaining what the meta-language must contain (we'll discuss that below), he writes:

> . . . [T]he full axiom system of the meta-theory includes three groups of sentences: (1) axioms of a general logical kind; (2) axioms which have the same meaning as the axioms of the science under investigation or are logically

[19] A similar point is made by Ray (2005), but the results here make it clear just how great the gap is between these two ways of understanding Tarski's claim.

[20] Proof: Let $\mathcal{T}$ be a formal theory. Then the statement that $\mathcal{T}$ is consistent can be formalized in arithmetic; if it is true, then true arithmetic of course proves it. So true arithmetic proves $\mathrm{Con}(\mathcal{T})$, for every consistent formal theory $\mathcal{T}$. Moreover, by the arithmetized completeness theorem, true arithmetic interprets every consistent formal theory.

[21] We'll see shortly that ${Q_{\text{seq}}}^T$ is mutually interpretable with $Q + \mathrm{Con}(Q_{\text{seq}})$, which is itself mutually interpretable with $Q + \mathrm{Con}(Q)$.

> stronger than them, but which in any case suffice... for the establishment of all sentences having the same meaning as the theorems of the science being investigated; finally, (3) axioms which determine the fundamental properties of the primitive concepts of a structural-descriptive type [that is, of syntax]. (Tarski, 1958, p. 211)

Note how Tarski speaks, at (2), of "the *science* under investigation". In a way, this is fine: If Tarski wants to investigate certain sciences—that is, theories—he's welcome to do so, and of course there are plenty of interesting results to be proved about theories using the techniques Tarski developed. (For example, using those techniques, we can prove that *PA* is consistent.) It is because Tarski thinks of 'sciences', rather than languages, as the object of investigation that, at (2), he includes (translations of) the axioms of the *object*-theory among those of the *meta*-theory. As applied to the sort of case we have been discussing, what this means is that the meta-theory in which Tarski proposes to develop a definition of truth suitable for use in investigations of *PA* will have to include either the axioms of *PA* themselves or else something sufficient for proving them.

But what has not been clearly appreciated, it seems to me, is that the axioms of *PA* need to be included among the axioms of the meta-theory *only* in so far as we want to prove certain results about *PA* specifically. If our goal is simply a materially adequate theory of truth for the *language* of *PA*, that is, for the language of arithmetic, then we have no need of (most of) those axioms. Whether Tarski himself appreciated this point in the 1930s, I do not know. But it was known to logicians no later than 1952, when it was used by Kleene (1952). We'll look at Kleene's argument in Section 6.1 or rather, at a later reformulation of it.

4. Cuts and Consistency

The question I now want to address is this: What does adding a truth-theory give us, as far as logical strength is concerned? It is obvious that $\mathcal{T}^T$ is at least as strong as $\mathcal{T}$, since it contains $\mathcal{T}$. But is it any stronger than $\mathcal{T}$?

There is no perfectly general answer to this question. What we shall see, however, is that, if $\mathcal{T}$ is finitely axiomatized, then $\mathcal{T}^T$ is stronger than $\mathcal{T}$, in the sense that $\mathcal{T}^T$ is not interpretable in $\mathcal{T}$. The proof uses a method called 'shortening of cuts' which is due to Robert Solovay and which plays a major role in the study of models of arithmetic. Since this method is not widely known among philosophers, I shall spend some time introducing it.

4.1. **The Method of Cuts.**

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 are

almost sure to be formulae $\phi(x)$ for which $\mathcal{T}$ proves the hypotheses of the relevant induction axiom—$\phi(0)$ and $\forall x(\phi(x) \rightarrow \phi(Sx))$—but for which $\mathcal{T}$ does *not* prove its conclusion: $\forall x\phi(x)$.[22] 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, $\boldsymbol{Q}$ does not prove that no number is its own successor. But $\boldsymbol{Q}$ does prove both $0 \neq S0$, which follows immediately from the first axiom of $\boldsymbol{Q}$, and $x \neq Sx \rightarrow Sx \neq SSx$, which follows just as immediately from the second. So $x \neq Sx$ 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) $\mathcal{T} \vdash \forall x(\iota(x) \rightarrow \iota(Sx))$.

Inductive formulas can be used to establish results about interpretability. The crucial result is this one.

Theorem 4.1. *Let $\iota(x)$ be a formula that is inductive in $\mathcal{T} \supseteq Q$ and that is no worse than Π_1. Then $\mathcal{T}$ interprets $Q + \forall x(\iota(x))$.*

It's not essential for what follows that the reader understand the proof of this theorem, so I shall not present it in any detail. But the method used in its proof—the shortening of cuts—is one we shall need below, so it is worth having some sense for how it works. I shall therefore explain the ideas behind the proof of Theorem 4.1 by continuing to discuss the example already mentioned: We'll see how to prove that $\boldsymbol{Q}$ interprets $Q + \forall x(x \neq Sx)$.

The basic idea is simply to restrict the domain to the numbers that satisfy $x \neq Sx$—which, one might say, might as well *be* the natural numbers, so far as $\boldsymbol{Q}$ 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 'real' natural 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

(1) $\iota(x)$ is inductive in $\mathcal{T}$
(2) $\mathcal{T} \vdash \forall x[\iota(x) \rightarrow \forall y < x(\iota(y))]$

[22] In the case of $I\Sigma_n$, one can actually exhibit such a formula (Hájek and Pudlák, 1993, p. 172). But note that, if there were no such cases, then the fact that the induction axiom was missing wouldn't cost us anything.

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 crucial result relating inductive formulas and cuts is this one.

Lemma 4.2 (Hájek and Pudlák 1993, p. 368)**.** *Let $\iota(x)$ be inductive in $\mathcal{T} \supseteq Q$. Then there is a cut $\kappa(x)$ in $\mathcal{T}$ for which $\mathcal{T} \vdash \forall x(\kappa(x) \rightarrow \iota(x))$. That is: Every inductive formula can be shortened to a 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 $\boldsymbol{Q}$ does not prove that $\leq$ is transitive.

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. In particular, we first consider the formula:

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

$\chi(x)$ says roughly that x satisfies $\iota(x)$ and that $\leq$ is transitive below x. It's easy to see that $\boldsymbol{Q}$ proves: $\chi(x)$ is inductive if $\iota(x)$ is.

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

$$\kappa(x) \stackrel{df}{\equiv} \forall w < x(\chi(w))$$

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

So, although $\boldsymbol{Q}$ can't prove that $x \neq Sx$ is a cut, there is a 'subcut' $\kappa(x)$ of $x \neq Sx$ in $\boldsymbol{Q}$. So we might now try simply restricting attention to $\kappa(x)$, the thought being that this will give us an interpretation in which $x \neq Sx$ holds and in which the axioms of $\boldsymbol{Q}$ 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 cuts to restrict attention to numbers that do have sums and products.[23] Doing so allows us to prove the following.

[23] There is an accessible treatment in Burgess's book *Fixing Frege* (Burgess, 2005, §2.2).

Lemma 4.3. *If $\mathcal{T} \supseteq Q$, then every formula $\iota(x)$ inductive in $\mathcal{T}$ can be shortened to a cut $\kappa(x)$ on which $\mathcal{T}$ proves the relativizations of the axioms of $\boldsymbol{Q}$.*

We can now see how Theorem 4.1 follows from Lemma 4.3.

Proof of Theorem 4.1. Start with a very simple case: $x \neq Sx$. We want to see that $\boldsymbol{Q}$ interprets $Q + \forall x(x \neq Sx)$. Since $x \neq Sx$ is inductive in $\boldsymbol{Q}$, by Lemma 4.3, there is a subcut $\kappa(x)$ of $x \neq Sx$ on which $\boldsymbol{Q}$ proves the relativizations of the axioms of $\boldsymbol{Q}$. Our interpretation is thus a 'pure relativization' to $\kappa(x)$. So we need only show that $\boldsymbol{Q}$ proves

$$\forall x(\kappa(x) \rightarrow x \neq Sx).$$

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

So now consider the case where $\iota(x)$ is Δ_0. By Lemma 4.3, there is a subcut $\delta(x)$ of $\iota(x)$ on which $\mathcal{T}$ proves the relativizations of the axioms of $\boldsymbol{Q}$. So now we just have to check that $\mathcal{T}$ proves the relativization of $\forall x(\iota(x))$. This is not as straightforward as in the previous case, because now there are quantifiers in $\iota(x)$ that themselves have to be relativized. But $\iota(x)$ is Δ_0, which means that the quantifiers in $\iota(x)$ are all *bounded*, and that means that the relativization is redundant, in the sense that, if $\iota(x)$ is Δ_0 and $\delta(x)$ is a cut in $\mathcal{T}$, then $\forall x(\delta(x) \rightarrow \iota(x))$ is going to be $\mathcal{T}$-provably equivalent to $\forall x(\delta(x) \rightarrow \iota^*(x))$, where $\iota^*(x)$ is the relativization of $\iota(x)$ to $\delta(x)$.[24] The proof is by induction on the complexity of expressions, of course, but the basic idea is simple enough. Consider, for example, $\forall y < t(\phi(y))$. This is relativized as: $\forall y < t(\delta(y) \rightarrow \phi(t))$. Since $\delta(x)$ is $\mathcal{T}$-provably closed under S, $+$, and $\times$, t, whatever it may be, is $\mathcal{T}$-provably going to satisfy $\delta(x)$, whence, since $\delta(x)$ is a cut, we have that $y < t \rightarrow \delta(y)$, and the new condition is redundant. Similarly for the existential case.

So suppose $\iota(x)$ is Π_1; say it is $\forall y \phi(x, y)$, where $\phi(x, y)$ is Δ_0. Then what we need to show is that $\mathcal{T}$ proves

$$\forall x[\delta(x) \rightarrow \forall y(\delta(y) \rightarrow \phi^*(x, y))]$$

As we just saw, $\mathcal{T}$ will prove

$$\forall y(\delta(y) \rightarrow \phi(x, y)) \equiv \forall y(\delta(y) \rightarrow \phi^*(x, y)),$$

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

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

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

[24] Burgess (2005, pp. 101–4) again has a nice discussion of this sort of point.

It is *not* in general true that, if $\iota(x)$ is a Σ_1 cut in $\mathcal{T} \supseteq Q$, then $\mathcal{T}$ interprets $Q + \forall x(\iota(x))$. The standard counterexample is $\exists y(\exp(x, y))$. This is inductive even in $\boldsymbol{Q}$, but $\boldsymbol{Q}$ does not interpret $Q + \forall x \exists y(\exp(x, y))$.

It does *not* follow, however, that the method of shortening cuts only works with Π_1 formulae. Sometimes one can show, by other means, that $\mathcal{T}$ proves the relativization of some Σ_1 formula.

As it happens, shortening of cuts can be used to prove stronger forms of Lemma 4.3 and so of Theorem 4.1.

Lemma 4.4. *If $\mathcal{T} \supseteq Q$, then every formula $\iota(x)$ inductive in $\mathcal{T}$ can be shortened to a cut $\kappa(x)$ on which $\mathcal{T}$ proves the relativizations of the axioms of $I\Delta_0$, and even of $I\Delta_0 + \omega_1$.*

We'll need this stronger result below.

4.2. **The Unprovability of 'Small' Consistency.** We know from Gödel's second incompleteness theorem that no 'sufficiently strong' theory proves its own consistency. In the mid-1980s, Pavel Pudlák proved a beautiful version of Gödel's result, one that really ought to be better known. If we think of the numbers satisfying a cut as 'small' numbers,[25] then what the theorem says is that no theory containing $\boldsymbol{Q}$ can prove that there are no 'small' proofs of contradictions from its axioms. More formally, then, what Pudlák's theorem says is that no theory containing $\boldsymbol{Q}$ proves its own consistency 'on a cut'.

Theorem 4.5 (Pudlák 1985, Theorem 2.1)**.** *Suppose $\mathcal{T} \supseteq Q$ is consistent, and let $\kappa(x)$ be a cut in $\mathcal{T}$. Then $\mathcal{T}$ does not prove:*[26]

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

Moreover, this continues to hold even if $\kappa(x)$ is merely inductive, since it can always be shortened to a cut.

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 theorem, by contrast, applies to any theory containing $\boldsymbol{Q}$. 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 resepct emerges from the following consequence of Theorem 4.5.

Theorem 4.6 (Pudlák 1985, Corollary 3.5)**.** *Suppose $\mathcal{T}$ is finitely axiomatized, sequential, and consistent. Then $\mathcal{T}$ does not interpret $Q + \mathrm{Con}(\mathcal{T})$.*

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

[26] Here, $\mathrm{Bew}_{\mathcal{T}}(x, y)$ is an appropriate formalization of '*x is a $\mathcal{T}$-proof of y*'.

Whereas Gödel tells us that $\mathcal{T}$ cannot prove $\mathrm{Con}(\mathcal{T})$, Pudlák tells us that, if $\mathcal{T}$ is finitely axiomatized, it cannot even *interpret* $Q + \mathrm{Con}(\mathcal{T})$, let alone $\mathcal{T} + \mathrm{Con}(\mathcal{T})$.[27]

The proofs of these two results are (well) beyond the scope of the present discussion.[28]

Putting these together, we have:[29]

Corollary 4.7. *Let $\mathcal{S} \supseteq Q$ be a consistent, finitely axiomatizable sequential theory that proves $\mathrm{Con}(\mathcal{T})$ 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 4.1 it will interpret $Q + \mathrm{Con}(\mathcal{T})$. But if $\mathcal{S}$ were interpretable in $\mathcal{T}$, then $Q + \mathrm{Con}(\mathcal{T})$ would be interpretable in $\mathcal{T}$. □

It is Corollary 4.7 that will do much of the work below.

5. The Strength of Truth-theories

5.1. $\mathcal{T}^T$ is Stronger than $\mathcal{T}$.

We are now ready to prove our first main result.[30]

Theorem 5.1. *Let $\mathcal{T} \supseteq I\Delta_0 + \omega_1$ and suppose that $\mathcal{T}^T$ proves that all axioms of $\mathcal{T}$ are true. Then $\mathcal{T}^T$ proves the consistency of $\mathcal{T}$ on a cut and so is not interpretable in $\mathcal{T}$.*

The natural proof of this needs to use $I\Delta_0 + \omega_1$ 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.

The key to the proof 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 the axioms are true; then we show that the rules of inference preserve truth; then we conclude, by induction, that all theorems are true. Since '$0 = S0$' is not true, it isn't a theorem, so $\mathcal{T}$ is consistent.

---

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

[28] As well as the paper of Pudlák's already cited, the interested reader may consult Hajék and Pudlák (1993, pp. 173ff); see also Visser's recent paper on the second incompleteness theorem (Visser, 2009a).

[29] Exercise: Show that we do not have to assume that $\mathcal{T}$ is consistent.

[30] The results reported in this section were taught to me by Albert Visser, though the proofs are my own, and the complications we shall meet arose as I tried to work out the details. There is a result very similar in feel to Corollary 5.6 in a recent paper of his (Visser, 2009a, theorem 4.1). Note that Corollary 5.6 leads to an alterative statement of Theorem 4.4 of that paper, which is the characterization of consistency statements that is its central purpose. This version relies upon coding, however, which is part of what Visser is trying to avoid. We will prove related results below, Theorem 6.5 and Corollary 6.13, that do not have this flaw.

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 'n line proofs have true conclusions' is inductive. Then we would have that, although $\mathcal{T}^T$ does not prove $\mathrm{Con}(\mathcal{T})$, it does prove it on a cut.

If that were the only obstacle, the proof would be easy. Unfortunately, there is another. We're just assuming, at present, that $\mathcal{T}^T$ can prove that all of $\mathcal{T}$'s *non*-logical axioms are true. But, to mimic Tarski's proof, we also need to prove that all the logical axioms are true and that the rules of inference are truth-preserving. This turns out to be more difficult than one might suppose. It helps to assume that the logic in which we're working is formulated as an axiomatic 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.[31]

The propositional axioms are easy enough.[32] Consider, for example, $p \to (q \to p)$. Let A and B be formulae. Using the clause for $\to$ twice, $\mathrm{Sat}_\sigma(\ulcorner A \to (B \to A) \urcorner)$ iff $\mathrm{Sat}_\sigma(A) \to (\mathrm{Sat}_\sigma(B) \to \mathrm{Sat}_\sigma(A))$. But the latter is of course a logical truth. So, generalizing, for any A and B, and for all σ, $\mathrm{Sat}_\sigma(\ulcorner A \to (B \to A) \urcorner)$, 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 $\forall\sigma\mathrm{Sat}_\sigma(\ulcorner A \to B \urcorner)$, then, by the clause for the conditional: $\forall\sigma(\mathrm{Sat}_\sigma(A) \to \mathrm{Sat}_\sigma(B))$. But then, by logic: $\forall\sigma\mathrm{Sat}_\sigma(A) \to \forall\sigma\mathrm{Sat}_\sigma(B)$. So, if $\forall\sigma\mathrm{Sat}_\sigma(A)$, then $\forall\sigma\mathrm{Sat}_\sigma(B)$.

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 $\mathrm{Sat}_\sigma(\forall v_i(\phi v_i))$ and $\neg\mathrm{Sat}_\sigma(\phi v_j)$. Now consider a sequence 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 $\mathrm{Sat}_\tau(\phi v_i)$. But

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

[32] Assuming we define truth as Tarski did, in terms of satisfaction by all sequences. If we use the alternate definition, and say that a line is true iff its universal closure is satisifed by $<>$, then we find ourselves needing to prove: $\forall\sigma(\mathrm{Sat}_\sigma(A)) \equiv \mathrm{Sat}_{<>}(\mathrm{ucl}(A))$. That only adds to our problems.

since (i) v_i stands in ϕv_i only where v_j stands in ϕv_j and (ii) τ assigns v_i the same value that σ assigns v_j, then we must have $\neg\text{Sat}_\tau(\phi v_i)$, since $\neg\text{Sat}_\sigma(\phi v_j)$. Contradiction. In making this last move, however, we are appealing to a general principle concerning 'variable-switching', one we might formulate as: 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 $\text{Sat}_\sigma(\phi v_j)$ iff $\text{Sat}_\sigma(\phi v_i)$.[33] 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:

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

(And we will have to face this problem, one way or another, if our language does indeed contain terms that are not variables.) In that case, the proof also requires the claim that all terms denote.

There are similar problems concerning universal generalization:

$$A \rightarrow \phi(v_i) \vdash A \rightarrow \forall v_i \phi(v_i),$$

where of course A must not contain 'x' free. Suppose that $A \rightarrow \forall v_i \phi(v_i)$ is not satisfied by all sequences. Then there is a sequence σ such that $\text{Sat}_\sigma(A)$ and $\neg\text{Sat}_\sigma(\forall v_i \phi(v_i))$. By the clause for $\forall$, then, we have a sequence $\tau \overset{i}{\sim} \sigma$ such that $\neg\text{Sat}_\tau(\phi(v_i))$. Since v_i is not free in A, then, $\text{Sat}_\tau(A)$, 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.:

$$\forall i(\text{free-in}(A, v_i) \rightarrow \forall x(\text{val}(\sigma, i, x) \equiv \text{val}(\tau, i, x)) \rightarrow \text{Sat}_\sigma(A) \equiv \text{Sat}_\tau(A)$$

But not will we be able to prove this without semantic induction.[34]

Careful examination of the proofs that the logical axioms are true and 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 $\text{Den}_\sigma(t, a)$ and $\forall k \neq i(\text{val}(\tau, k) = \text{val}(\sigma, k) \wedge \text{val}(\tau, i) = a$, then, $\text{Sat}_\sigma(\phi t)$ iff $\text{Sat}_\tau(\phi v_i)$.

(2) Suppose that σ and τ agree on all free variables contained in A. Then $\text{Sat}_\sigma(A)$ iff $\text{Sat}_\tau(A)$.

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 $\forall k \neq i(\tau_k = \sigma_k) \wedge \tau_i = \sigma_j$, then $\text{Den}_\sigma(u(t), a)$ iff $\text{Den}_\tau(u(v_i), a)$.

[33] Note that ϕv_j does not contain v_i free.

[34] 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.

(4) Suppose that σ and τ agree on all free variables contained in t. Then $\text{Den}_\sigma(t, a)$ iff $\text{Den}_\tau(t, a)$.

We also need:

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

though this will be trivial if there are no terms in the language other than variables.

We thus have no hope whatsoever of showing that $\mathcal{T}^T$ 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 $\mathcal{T}^T$ 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 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 $\text{Den}_\sigma(t, a)$ and $\forall k \neq i(\tau_k = \sigma_k) \wedge \tau_i = a$, then $\text{Sat}_\sigma(\phi t)$ iff $\text{Sat}_\tau(\phi v_i)$.

The usual proof of (1) can be adapted to show that (1*) is inductive. There are similar formulae corresponding to (2)–(4), and the usual proofs of them can also 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 $\exists x(\text{Den}_\sigma(t, x))$.

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 5.1, then, for the special case of relational languages.

Theorem 5.2. *Let $\mathcal{T} \supseteq I\Delta_0 + \omega_1$, where $\mathcal{L}_T$ is relational, and suppose that $\mathcal{T}^T$ proves that all axioms of $\mathcal{T}$ are true. Then $\mathcal{T}^T$ proves the consistency of $\mathcal{T}$ on a cut and so is not interpretable in $\mathcal{T}$.*

Proof. As noted, the usual proofs of (1)–(4) can be adapted to show that their starred versions are inductive, so, by Lemma 4.4, $\mathcal{T}^T$ proves their relativizations to some cut on which it also proves the axioms of $I\Delta_0 + \omega_1$. What we now do is 'work on this cut', as it is said: Relativizing everything to the cut, we can prove the relativization of 'logic is true' on the cut. On that cut, we will be able to prove that 'n line proofs have true conclusions'

is inductive and will therefore be able to construct a cut 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:

$$\begin{gathered}\mathrm{Bew}_{\mathcal{T}}(n) \wedge \lambda(n) \wedge \forall m < \mathrm{len}(n)(\lambda(n_m)) \rightarrow \\ \forall m < \mathrm{len}(n)(\mathrm{T}(n_m))\end{gathered}$$

The second 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.

Now consider:

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

The usual argument 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. We therefore have a cut $\kappa(x)$ on which $\forall k \leq n(\phi(k))$ holds. I.e., we can prove:

$$\begin{gathered}\forall n\{\kappa(n) \rightarrow \\ \forall k \leq n[\lambda(k) \wedge \mathrm{Bew}_{\mathcal{T}}(k) \wedge \forall m < \mathrm{len}(k)(\lambda(k_m)) \rightarrow \\ \forall m < \mathrm{len}(k)(\mathrm{T}(k_m))]\}\end{gathered}$$

Taking k to be n, we have:

$$\begin{gathered}\forall n\{\kappa(n) \wedge \lambda(n) \wedge \mathrm{Bew}_{\mathcal{T}}(n) \wedge \forall m < \mathrm{len}(n)(\lambda(n_m)) \rightarrow \\ \forall m < \mathrm{len}(n)(\mathrm{T}(n_m))\}\end{gathered}$$

What we want is:

(*) $$\forall n\{\kappa(n) \wedge \mathrm{Bew}_{\mathcal{T}}(n) \rightarrow \forall m < \mathrm{len}(n)(\mathrm{T}(n_m))\}$$

We thus need to eliminate the other conjuncts of the antecedent by showing that those two conjuncts:

$$\text{(i) } \lambda(n) \quad \text{and} \quad \text{(ii) } \forall m < \mathrm{len}(n)(\lambda(n_m))$$

follow from the other two: $\kappa(n)$ and $\mathrm{Bew}_{\mathcal{T}}(n)$. For the first, by construction, $\kappa(x)$ is a sub-cut of $\lambda(x)$. For (2), we need only make sure that $\lambda(x)$ satisfies:

$$\mathrm{seq}(n) \rightarrow \forall m < \mathrm{len}(n)(n_m < n)$$

We can do this simply by making sure that the axioms of, say, $I\Delta_0$ are true on $\lambda(x)$, or we could build this condition directly into $\lambda(x)$.

From (*), then, we easily derive

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

and so:

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

So $\mathcal{T}$ is consistent on $\kappa(x)$. □

With Theorem 5.2 in hand, we can extend the result to non-relational languages and so establish Theorem 5.1.

Proof of Theorem 5.1. Let $\mathcal{T}_R$ be the relational version of $\mathcal{T}$. What we are going to see is that ${\mathcal{T}_R}^T$ is interpretable in $\mathcal{T}^T$. It is easy enough to interpret $\mathcal{T}$ in $\mathcal{T}_R$, of course, via such translations as:

$$r(\ulcorner Axyz \urcorner) = \ulcorner x + y = z \urcorner$$

But, of course, that is not all we need to do. We need to interpret the *semantics* of the relational language in that of the non-relational language. But this is fairly easy to do. The idea is just to translate $\mathrm{Sat}_\sigma(A)$ as $\mathrm{Sat}_\sigma(r(A))$, where $r(x)$ is a formula of the language of $\mathcal{T}$ that expresses the translation from $\mathcal{L}_{\mathcal{T}_R}$ to $\mathcal{L}_{\mathcal{T}}$.[35] And 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

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

is

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

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 $Axyz$, which is essentially:

$$\mathrm{Sat}_\sigma(\ulcorner Av_i v_j v_k \urcorner) \equiv A\sigma_i\sigma_j\sigma_k$$

Its translation is:

$$\mathrm{Sat}_\sigma(r(\ulcorner Av_i v_j v_k \urcorner)) \equiv (\sigma_i + \sigma_j = \sigma_k)$$

But $r(\ulcorner Av_i v_j v_k \urcorner)$ is $v_i + v_j = v_k$, so this becomes:

$$\mathrm{Sat}_\sigma(\ulcorner v_i + v_j = v_k \urcorner) \equiv (\sigma_i + \sigma_j = \sigma_k)$$

which is easily provable.

So ${\mathcal{T}_R}^T$ is interpretable in $\mathcal{T}^T$. But ${\mathcal{T}_R}^T$ proves $\mathrm{Con}(\mathcal{T}_R)$ on a cut and so interprets $Q + \mathrm{Con}(\mathcal{T}_R)$ and, in fact, interprets $I\Delta_0 + \omega_1 + \mathrm{Con}(\mathcal{T}_R)$. But $I\Delta_0 + \omega_1$ is strong enough to verify the fact that $\mathcal{T}$ is interpretable in

[35] Since the translation is recursive, it will of course be representable in $\mathcal{T}$. In general, of course, it will be represented by some formula $R(x, y)$, not by a function like $r(x)$. But this point affects nothing that follows and only complicates the exposition. (We probably do need to know that every formula has exactly one translation. But $I\Delta_0 + \omega_1$ will prove such facts.)

$\mathcal{T}_R$ and so to prove that, if $\mathrm{Con}(\mathcal{T}_R)$, then $\mathrm{Con}(\mathcal{T})$. So $I\Delta_0 + \omega_1 + \mathrm{Con}(\mathcal{T}_R)$ actually proves $\mathrm{Con}(\mathcal{T})$, whence $I\Delta_0 + \omega_1 + \mathrm{Con}(\mathcal{T})$, and so of course $Q + \mathrm{Con}(\mathcal{T})$, is interpretable in ${\mathcal{T}_R}^T$ and so in $\mathcal{T}^T$. □

Theorem 5.3. *Let $\mathcal{T} \supseteq Q$ and suppose that $\mathcal{T}^T$ proves that all axioms of $\mathcal{T}$ are true. Then $\mathcal{T}^T$ proves the consistency of $\mathcal{T}$ on a cut.*

Proof. Start the proofs of the preceding theorems by restricting everything to a cut on which the axioms of $I\Delta_0 + \omega_1$ are available. □

Corollary 5.4. *Suppose $\mathcal{T} \supseteq Q$ is finitely axiomatized. Then $\mathcal{T}^T$ is not interpretable in $\mathcal{T}$.*

Proof. From Theorem 5.3 and Corollary 3.3. □

Theorem 5.5. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $Q + \mathrm{Con}(\mathcal{T})$ interprets $\mathcal{T}^T$.*

Proof. The proof of this theorem is similar to that of Theorem 6.12 below, but harder, since Q is so weak. See Visser's paper "The Predicative Frege Hierarchy" for the details (Visser, 2009b). □

Corollary 5.6. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $Q + \mathrm{Con}(\mathcal{T})$ is mutually interpretable with $\mathcal{T}^T$.*

So there is a straightforward sense in which a 'full truth-theory' is a sort of functor that strengthens any finitely axiomatized theory you feed it. We'll see some similar, but much more general, results below.

There is one more result I want to mention before we continue.

Theorem 5.7. *$(I\Delta_0 + \omega_1)^T$ is interpretable in Q^T.*

We will not actually need this result for anything that follows, but the proof seems to me to be of substantial interest. The technique involved will be familiar from the proof of Theorem 5.1, but it is applied more subtly. It will be clear, too, that it is not special to this particular case. What it shows, in effect, is that we can always relativize the semantic part of a theory like Q^T to a cut.

Proof. We know, of course, that we can interpret $I\Delta_0 + \omega_1$ in Q. The problem is to do so while preserving the semantic part of Q^T. We cannot actually expect $(I\Delta_0 + \omega_1)^T$ to prove the relativizations of the semantic axioms of Q^T. That would mean, in particular, proving the relativization of the clause for $\exists$, which would be:

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

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, which is, in general, false. But what we can do is reinterpret satisfaction itself so that $\mathrm{Sat}_\sigma(A)$ means: the *relativization* of A is satisfied by σ. That is, we translate $\mathrm{Sat}_\sigma(A)$ as: $\mathrm{Sat}_\sigma(t^\kappa(A))$, where

$t^\kappa(x)$ is a syntactic function meaning: the relativization of A to $\kappa(x)$.[36] So what we need to prove is:

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

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

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

This, however, is easily proven.

Left to right: By the clauses for $\exists$ and $\wedge$, $\mathrm{Sat}_\sigma(\ulcorner \exists v_i(\kappa(v_i) \wedge t^\kappa(\phi v_i))\urcorner)$ iff $\exists \tau[\tau \overset{i}{\sim} \sigma \wedge \mathrm{Sat}_\tau(\ulcorner \kappa(v_i)\urcorner) \wedge \mathrm{Sat}_\tau(t^\kappa(\ulcorner \phi v_i \urcorner))]$. But $\kappa(v_i)$ is a specific formula, and so we can prove a Sat-sentence for it. In particular, we have:

$$\mathcal{T}^T \vdash \mathrm{Sat}_\tau(\ulcorner \kappa(v_i)\urcorner) \equiv \kappa(\tau_i)$$

But if $\kappa(\tau_i)$, 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 $I\Delta_0 + \omega_1$, so it will be true on the cut given by $\kappa(x)$, so we are done.

The converse is similar. □

5.2. **Peano Arithmetic Is a Special Case (I).** The results proven in the preceding section depend heavily upon the assumption that $\mathcal{T}$ is finitely axiomatized. This is because, as mentioned previously, is that, if $\mathcal{T}$ is infinitely axiomatized, then there is no reason, in general, to suppose that $\mathcal{T}^T$ will prove that all of $\mathcal{T}$'s axioms are true, though it will prove that each of them is. But we do have the following obvious corollary.

Corollary 5.8. *Let $\mathcal{T}$ be an infinitely axiomatized theory. Then $\mathcal{T}^T$ is mutually locally interpretable with $Q + \bigcup\{\mathrm{Con}(\mathcal{U})\}$, where $\mathcal{U}$ is a finite fragment of $\mathcal{T}$.*

Each finite fragment $Q + \mathrm{Con}(\mathcal{U}_1) + \cdots + \mathrm{Con}(\mathcal{U}_n)$ of $Q + \bigcup\{\mathrm{Con}(\mathcal{U})\}$ is interpretable in ${\mathcal{U}_1}^T + \cdots + {\mathcal{U}_n}^T \subseteq \mathcal{T}^T$. Each finite fragment $\mathcal{U}^T$ of $\mathcal{T}^T$ is interpretable in $Q + \mathrm{Con}(\mathcal{U}) \subseteq Q + \bigcup\{\mathrm{Con}(\mathcal{U})\}$.

Corollary 5.9. *If $\mathcal{T}$ is reflexive, then $\mathcal{T}^T$ 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, if $\mathcal{T}$ is reflexive, it contains $Q + \bigcup\{\mathrm{Con}(\mathcal{U})\}$ and so itself locally interprets $\mathcal{T}^T$, and it then follows from Orey's Theorem that $\mathcal{T}$ *globally* interprets $\mathcal{T}^T$. □

So, in particular, we have:[37]

[36] 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.

[37] Stronger versions of this result have been proven by Visser and Enayat. This version seems to be folklore.

Corollary 5.10. *PA^T is intepretable in PA.*

5.3. **Extending Induction.** As I have emphasized, what was shown in Section 5.1 is *not* that $\mathcal{T}^T$ proves that $\mathcal{T}$ is consistent. If $\mathcal{T}$ is a finitely axiomatized (sequential) theory, then $\mathcal{T}^T$ will prove that $\mathcal{T}$'s axioms are true and will prove that the rules preserve truth, but $\mathcal{T}^T$ does not have the induction axiom needed to conclude that all proofs have true conclusions.[38] 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 to replace $\mathcal{T}^T$ with a version that extends the induction axioms to permit semantic predicates—in particular, the truth-predicate—to occur therein.

It is not at all obvious, in general, what it means to 'extend the induction scheme'. The scheme might itself be stated in such a way as to exclude formluae 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 $\text{Sat}_\sigma(t)$ 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). 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 then $\Sigma_1(X)$ if it is of the form $\forall y \phi(y)$, where $\phi(y)$ is $\Delta_0(X)$. And so forth. In our case, if we take Sem to be the set of atomic semantic formulae—$\text{Den}_\sigma(t, x)$, $\text{Sat}_\sigma(x)$, and so forth—then what it means to 'extend induction' in the case of $I\Delta_0$, say, is that we permit induction on $\Delta_0(\text{Sem})$ formulae. The resulting theory is thus $I\Delta_0(\text{Sem})$. Similarly for $I\Sigma_1$, etc.

Definition. Suppose that $\mathcal{T}$ is among $I\Delta_0$, $I\Sigma_n$, $I\Delta_0(X)$, and so forth. Then:

---

[38] Indeed, $\mathcal{T}^T$ cannot *even* prove that all *logically* provable sentences—that is, sentences provable using none of the special axioms of $\mathcal{T}$—are true. Suppose $\mathcal{T}$ proves the following:

> Let the $\mathcal{T}_i$ be the axioms of $\mathcal{T}$. Then if A is a theorem of $\mathcal{T}$, $\bigwedge_i \mathcal{T}_i \rightarrow A$ is logically provable.

I do not know exactly how much is needed for the proof of the result. Not very much, to be sure, but it surely cannot be proven in *Q*. In any event, reason in $\mathcal{T}^T$. Suppose that A is provable in $\mathcal{T}$. Then $\bigwedge_i \mathcal{T}_i \rightarrow A$ is logically provable. By the Tarski clauses, $\text{T}(\bigwedge_i \mathcal{T}_i \rightarrow A)$ iff $\text{T}(\mathcal{T}_1) \wedge \cdots \wedge \text{T}(\mathcal{T}_n) \rightarrow \text{T}(A)$. Since each axiom is true, $\text{T}(A)$ if $\text{T}(\bigwedge_i \mathcal{T}_i \rightarrow A)$. So, if all logically provable sentences are true, every $\mathcal{T}$-provable sentence is true.

Cieśliński (2009) notes that the same result holds even in the case of *PA*, though of course the argument is more complicated, since *PA* is not finitely axiomatizable.

(1) $\mathcal{T}^{D+}$ is the result of (i) adding all T-sentences for the language of $\mathcal{T}$ and (ii) extending the induction scheme in the sense just explained.
(2) $\mathcal{T}^{S+}$ is the result of (i) adding not just the T-sentences for the language of $\mathcal{T}$ but also the 'Sat-sentences', such as

$$\text{Sat}_\sigma(v_0 = v_1)] \equiv \exists x \exists y[\text{val}(\sigma, 0, x) \wedge \text{val}(\sigma, 1, y) \wedge x = y],$$

and (ii) extending the induction scheme.
(3) $\mathcal{T}^{T+}$ is the result of (i) adding a fully compositional truth-theory and (ii) extending the induction scheme.

We'll begin by exploring $\mathcal{T}^{D+}$.

It's well-known that PA^{D+} is a conservative extension of *PA*. Here's a similar result, but stated in terms of interpretability.

Theorem 5.11. *PA^{D+} is interpretable in PA.*

Proof. Let S be a finite subset of the axioms of PA^{D+}. S will contain at most finitely many T-sentences, say for $A_1, \ldots, A_n$. We interpret $\text{T}(x)$ in terms of a 'list-like' theory of truth, that is, as:

$$(x = \ulcorner A_1 \urcorner \wedge A_1) \vee \cdots \vee (x = \ulcorner A_n \urcorner \wedge A_n)$$

Clearly, with $\text{T}(x)$ so defined, *PA* will prove the T-sentences for $A_1, \ldots, A_n$. Moreover, with $\text{T}(x)$ so defined, any extended induction axioms that appear in S will simply become induction axioms of *PA*.

So PA^{D+} is locally interpretable in *PA*. Now apply Orey's Theorem. □

Note that this continues to hold for PA^{S+}, by pretty much the same reasoning. The same argument shows that Q^{D+} and Q^{S+} are locally interpretable in Q.

The proof of 5.11 does *not* obviously extend, however, to sub-systems of *PA* such as $I\Sigma_1$: We cannot show so simply that $I\Sigma_1^{D+}$ is locally interpretable in $I\Sigma_1$. The reason is that the A_i may be of any complexity, and so, if we have an induction axiom for some $\Sigma_1(\text{Sem})$ formula $A(x)$, the result of replacing $\text{T}(x)$ by its 'list-like' definition in $A(x)$ may yield a formula that is not itself Σ_1. But there is a slightly more complicated proof that does work.

Theorem 5.12. *$I\Sigma_n^{D+}$ is locally interpretable in $I\Sigma_n$.*

Proof. Let S be a finite subset of the axioms of $I\Sigma_n^{D+}$. Then S contains only finitely many T-sentences. For illustration, say these are A and B. As before, we interpret $\text{T}(x)$ as: $(x = \ulcorner A \urcorner \wedge A) \vee (x = \ulcorner B \urcorner \wedge B)$. We can then easily prove the T-sentences for A and B. But, of course, S may also contain some extended induction axioms from $I\Sigma_n^{D+}$. We need to see that these are also going to be provable.

Suppose one of these induction axioms is the axiom for the formula $\phi(x) \vee \mathbf{T}(\mathbf{sb}(\ulcorner \psi(x) \urcorner, x))$, where $\phi(x)$ is itself Σ_n but $\psi(x)$ need not be.[39] The induction axiom in question is thus:

$$\begin{aligned}\phi(0) \vee \mathbf{T}(\mathbf{sb}(\ulcorner \psi(x) \urcorner, 0)) \wedge \\ \forall x[\phi(x) \vee \mathbf{T}(\mathbf{sb}(\ulcorner \psi(x) \urcorner, x)) \rightarrow \phi(Sx) \vee \mathbf{T}(\mathbf{sb}(\ulcorner \psi(x) \urcorner, Sx))] \rightarrow \\ \forall x(\phi(x) \vee \mathbf{T}(\mathbf{sb}(\ulcorner \psi(x) \urcorner, x)))\end{aligned}$$

Under our interpretation of $\mathbf{T}(x)$, this becomes:

$$\begin{aligned}&\big[\phi(0) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, 0) = \ulcorner A \urcorner \wedge A) \vee \phi(0) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, 0) = \ulcorner B \urcorner \wedge B)\big] \wedge \\ &\forall x\big[\phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner A \urcorner \wedge A) \vee \phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner B \urcorner \wedge B) \rightarrow \\ &\phi(Sx) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, Sx) = \ulcorner A \urcorner \wedge A) \vee \phi(Sx) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, Sx) = \ulcorner B \urcorner \wedge B)\big] \rightarrow \\ &\forall x[\phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner A \urcorner \wedge A) \vee \phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner B \urcorner \wedge B)]\end{aligned}$$

(Sorry about that.) The crucial point is that A and B are *sentences*, so the quantifier $\forall x$ cannot bind any variables in A or B . Hence, they can be "pulled out" in the following way.

Abbreviate the formula just displayed as $\Phi(A, B)$. Then it is logically equivalent to:

$$\begin{aligned}[A \wedge B \rightarrow \Phi(\top, \top)] \wedge [A \wedge \neg B \rightarrow \Phi(\top, \bot)] \wedge \\ [\neg A \wedge B \rightarrow \Phi(\bot, \top)] \wedge [\neg A \wedge \neg B \rightarrow \Phi(\bot, \bot)]\end{aligned}$$

where $\top$ is $0 = 0$ and $\bot$ is $0 \neq 0$. By completeness, this equivalence is provable. Now $\Phi(\top, \top)$ is:

$$\begin{aligned}&\phi(0) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, 0) = \ulcorner A \urcorner \wedge \top) \vee \phi(0) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, 0) = \ulcorner B \urcorner \wedge \top) \wedge \\ &\forall x[\phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner A \urcorner \wedge \top) \vee \phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner B \urcorner \wedge \top) \rightarrow \\ &\phi(Sx) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, Sx) = \ulcorner A \urcorner \wedge \top) \vee \phi(Sx) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, Sx) = \ulcorner B \urcorner \wedge \top)] \rightarrow \\ &\forall x[\phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner A \urcorner \wedge \top) \vee \phi(x) \vee (\mathbf{sb}(\ulcorner \psi(x) \urcorner, x) = \ulcorner B \urcorner \wedge \top)]\end{aligned}$$

and that is itself a Σ_n induction axiom. It is therefore provable, and hence so is $A \wedge B \rightarrow \Phi(\top, \top)$. The same goes for the other cases. So the induction axiom in question is provable under our interpretation of $\mathbf{T}(x)$.

Of course, nothing hinges on the details of this particular example. $\square$

As we shall see, corresponding results do *not* hold for $I\Sigma_n^{S+}$.

[39] Here, $\mathbf{sb}(y, x)$ means: The result of substituting the numeral for x for the sole free variable in y. I choose this example because the threat here is that the ability to substitute in this way will allow us to get the induction axiom for $\phi(x) \vee \psi(x)$, which need not be Σ_n.

5.4. **Semantic Consistency Proofs.** If $\mathcal{T}^{T+}$ is going to formalize Tarski's proof of Con($\mathcal{T}$), then it 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 obvious sort of formula to use in the inductive part of the proof is something like:

$$\iota(n) \stackrel{df}{\equiv} \forall z \forall y \forall m < n[(\text{Bew}_{\mathcal{T}}(z, y) \wedge \text{lh}(z, m) \rightarrow \forall \sigma \text{Sat}_{\sigma}(y)]$$

This is Π_1(Sem). Moreover, as a look back at (1)–(5) will show, the formulae involved in the various inductions needed to prove that logic is true are Π_1(Sem)—except for the one concerning denotation, which is Σ_1(Sem). Since $I\Sigma_1$ has induction for Π_1 formulae (Hájek and Pudlák, 1993, p. 63, theorem 2.4), we thus have:[40]

Theorem 5.13. *Suppose $\mathcal{T} \supseteq I\Sigma_1$ and suppose further that $\mathcal{T}^{T+}$ proves that all axioms of $\mathcal{S}$ (which may or may not be $\mathcal{T}$) are true. Then $\mathcal{T}^{T+}$ proves* Con($\mathcal{S}$).

Corollary 5.14. *Suppose $\mathcal{T} \supseteq I\Sigma_1$ is finitely axiomatized. Then $\mathcal{T}^{T+}$ proves* Con($\mathcal{T}$).

This might seem like a nice, neat result. Since $I\Sigma_n$ is finitely axiomatizable, we'll get that $(I\Sigma_1)^{T+}$ proves the consistency of $I\Sigma_1$, that $(I\Sigma_2)^{T+}$ proves Con($I\Sigma_2$), and so forth.

Unfortunately, things do not work out nearly so nicely.

5.5. **How *PA*$^{T+}$ Proves** Con(*PA*)**.** Everyone knows that *PA*$^{T+}$ proves Con(*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 the axioms are true; then you prove that the rules of inference preserve truth; and then you use the extended induction scheme to conclude that all the theorems are true. Since '$0 = S0$' is provably untrue, it isn't a theorem, so *PA* is consistent.

But this sketch fails to address a very important question, namely: How are we supposed to prove that *all* of the axioms of *PA* are true?[41] 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

[40] Henrik Kotlarski (1986) seems to claim that this result can be strengthened to $\mathcal{T} \supseteq I\Delta_0$. This seems doubtful, however. Kotlarski is simply not careful enough about the case of the logical axioms. Enayat and Visser have show that Kotlarski's result can be salvaged in the semantic setting in which he works 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.

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

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. 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:

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

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. Note that $A(v_0, \vec{y})$ is doing duty as a variable with which we are reasoning in *PA*.

We have the induction axiom:

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

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. This will be true just in case the displayed formula is satisfied by every sequence σ. But then, by the clauses for the connectives, that holds just in case, for every sequence σ:

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

This is what we want to prove. What we need to show is:

(1) $\text{Sat}_\sigma(\ulcorner A(0, \vec{y})\urcorner)$ implies $\phi(\ulcorner A(v_0, \vec{y})\urcorner, 0, \sigma)$
(2) $\text{Sat}_\sigma(\ulcorner \forall v_0[A(v_0, \vec{y}) \rightarrow A(Sv_0, \vec{y})]\urcorner)$ implies
$\forall v_0[\phi(\ulcorner A(v_0, \vec{y})\urcorner, v_0, \sigma) \rightarrow \phi(\ulcorner A(v_0, \vec{y})\urcorner, Sv_0, \sigma)]$
(3) $\forall v_0[\phi(\ulcorner A(v_0, \vec{y})\urcorner, v_0, \sigma)]$ implies $\text{Sat}_\sigma(\ulcorner \forall v_0 A(v_0, \vec{y})\urcorner)$

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 $\text{Sat}_\sigma(\psi)$ iff $\text{Sat}_\tau(\psi)$.
(ii) If $\text{Sat}_\sigma(\ulcorner \psi(0)\urcorner)$ and $\text{val}(\sigma, 0, 0)$, then $\text{Sat}_\sigma(\ulcorner \psi(v_0)\urcorner)$.
(iii) If $\text{Sat}_\sigma(\ulcorner \psi(Sv_0)\urcorner)$, and if τ is just like σ except that what it assigns to v_0 is the successor of what σ assigns to v_0, then $\text{Sat}_\tau(\ulcorner \psi(v_0)\urcorner)$.[42]

All of these are provable in PA^{T+} by the usual sorts of arguments.

We get (1) immediately from (ii).

[42] We're assuming, of course, that *all* free occurences of v_0 have been replaced by occurrences of Sv_0.

For (2), $\text{Sat}_\sigma(\ulcorner \forall v_0[A(v_0, \vec{y}) \rightarrow A(Sv_0, \vec{y})]\urcorner)$ is equivalent to:

$$\forall \chi \overset{0}{\sim} \sigma\, [\text{Sat}_\chi(\ulcorner A(v_0, \vec{y})\urcorner) \rightarrow \text{Sat}_\chi(\ulcorner A(Sv_0, \vec{y})\urcorner)] \tag{5.1}$$

What we want to show is:

$$\forall v_0\{\exists \tau[\tau \overset{0}{\sim} \sigma \wedge \text{val}(\tau, 0, v_0) \wedge \text{Sat}_\tau(\ulcorner A(v_0, \vec{y})\urcorner)] \rightarrow$$
$$\exists \tau[\tau \overset{0}{\sim} \sigma \wedge \text{val}(\tau, 0, Sv_0) \wedge \text{Sat}_\tau(\ulcorner A(v_0, \vec{y})\urcorner)]\}$$

So suppose $\exists \tau[\tau \overset{0}{\sim} \sigma \wedge \text{val}(\tau, 0, v_0) \wedge \text{Sat}_\tau(\ulcorner A(v_0, \vec{y})\urcorner)]$. Then $\text{Sat}_\tau(\ulcorner A(Sv_0, \vec{y})\urcorner)$, by (5.1). Now let υ be just like τ except that it assigns v_0 the successor of what τ assigns v_0. Then, by (iii), $\text{Sat}_\upsilon(\ulcorner A(v_0, \vec{y})\urcorner)$. And so, generalizing, we have

$$\exists \tau[\tau \overset{0}{\sim} \sigma \wedge \text{val}(\tau, 0, Sv_0) \wedge \text{Sat}_\tau(\ulcorner A(v_0, \vec{y})\urcorner)]\},$$

as wanted.

The argument for (3) is similar.

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 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, such as primitive recursive arithmetic, then there is no reason whatsoever to expect that $\mathcal{T}^{T+}$ should prove that all of $\mathcal{T}$'s axioms are true.

So, as Visser once put it, the fact that PA^{T+} proves Con(*PA*) is something of a happy accident. Too happy, as we are about to see.

5.6. **An Unfortunate Result.**

Lemma 5.15. $(I\Sigma_1)^{T+}$ *proves that all axioms of PA are true.*

Proof. The argument given in the last section needed precisely one extended instance of induction, that for the formula $\phi(a, \sigma, z)$. This is Σ_1.[43] The other thing we need to check is that the general principles (i)–(iii) on which we relied can be proven in $(I\Sigma_1)^{T+}$. 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 $I\Sigma_1$; the universal quantifier over sequences makes the relevant claims Π_1. So the reasoning in the last section can all be carried out in $(I\Sigma_1)^{T+}$. □

Corollary 5.16. $(I\Sigma_1)^{T+}$ *proves* Con(*PA*).[44]

Proof. By Theorem 5.13. □

[43] As mentioned in note 11, $\tau \overset{0}{\sim} \sigma$ can be made to be Δ_1.

[44] Kotlarski (1986) claims that $(I\Delta_0)^{T+}$ proves $\text{Con}(PA)$. It can be shown that $(I\Delta_0)^{T+}$ proves that all axioms of *PA* are true, but, as mentioned in note 40, Kotlarski does not address the question how we are supposed to show that 'logic is true', and it does not appear that $(I\Delta_0)^{T+}$ can prove that logic is true.

Since Con(*PA*) is a single theorem of PA^{T+}, the full power of PA^{T+} can't be needed for the proof; only finitely many axioms of PA^{T+} will be needed, so Con(*PA*) has to be provable in $(I\Sigma_n)^{T+}$, for some n. In that sense, Corollary 5.16 is no surprise. Nonetheless, I take it 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 sort of semantic consistency proof for $I\Sigma_1$. It's disappointing if the answer turns out to be, "One that proves Con(*PA*)".[45]

It's worth noting that we get a similar phenomenon in $(I\Delta_0)^{S+}$.

Theorem 5.17. $(I\Delta_0)^{S+}$ *contains PA.*

The argument is similar in spirit to the one for Lemma 5.15. The difference is that, in this case, we are proving only that $(I\Delta_0)^{S+}$ proves *each* of the induction axioms for *PA*, rather than that it proves that they are all true. This makes things rather easier.

Proof. Let $A(v_0, v_1)$ be a formula; extension to the case of extra free variables is straightforward. Now consider the formula:

$$\phi(z) \stackrel{df}{\equiv} \exists\tau < t(\sigma, z)\left[\tau \stackrel{0}{\sim} \sigma \wedge \mathbf{val}(\tau, 0, z) \wedge \mathbf{Sat}_\tau(\ulcorner A(v_0, v_1)\urcorner)\right]$$

Here, $t(\sigma, z)$ is a term I shall not attempt to describe that appropriately bounds the initial quantifier.[46] So this is $\Delta_0(\mathbf{Sem})$, so $(I\Delta_0)^{S+}$ has induction for it. The induction axiom is:

$$\begin{aligned}
&\exists\tau < t(\sigma, 0)\left[\tau \stackrel{0}{\sim} \sigma \wedge \mathbf{val}(\tau, 0, 0) \wedge \mathbf{Sat}_\tau(\ulcorner A(v_0, v_1)\urcorner)\right] \wedge \\
\forall v_0\{&\exists\tau < t(\sigma, v_0)[\tau \stackrel{0}{\sim} \sigma \wedge \mathbf{val}(\tau, 0, v_0) \wedge \mathbf{Sat}_\tau(\ulcorner A(v_0, v_1)\urcorner)] \rightarrow \\
&\exists\tau < t(\sigma, Sv_0)[\tau \stackrel{0}{\sim} \sigma \wedge \mathbf{val}(\tau, 0, Sv_0) \wedge \mathbf{Sat}_\tau(\ulcorner A(Sv_0, v_1)\urcorner)]\} \rightarrow \\
&\forall v_0\exists\tau < t(\sigma, v_0)\left[\tau \stackrel{0}{\sim} \sigma \wedge \mathbf{val}(\tau, 0, v_0) \wedge \mathbf{Sat}_\tau(\ulcorner A(v_0, v_1)\urcorner)\right]
\end{aligned}$$

But the Sat-sentence for $A(x, y)$ will give us:

$$\mathbf{Sat}_\tau(\ulcorner A(v_0, v_1)\urcorner) \equiv \exists v_0\exists v_1[\mathbf{val}(\tau, 0, v_0) \wedge \mathbf{val}(\tau, 1, v_1) \wedge A(v_0, v_1)]$$

So $\phi(z)$ is equivalent to:

$$\exists v_1[\mathbf{val}(\tau, 1, v_1) \wedge A(z, v_1)]$$

[45] Of course, *PA* itself proves Con($I\Sigma_1$), 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 we are discussing.

[46] Since τ can be taken to be σ with the first entry changed to z, we can actually calculate what τ is, given σ.

Then the induction axiom reduces to:

$$\exists v_1[\mathbf{val}(\sigma, 1, v_1) \wedge A(0, v_1)] \wedge$$
$$\forall v_0 \left[\exists v_1[\mathbf{val}(\sigma, 1, v_1) \wedge A(v_0, v_1)] \rightarrow \exists v_1[\mathbf{val}(\sigma, 1, v_1) \wedge A(Sv_0, v_1)]\right] \rightarrow$$
$$\forall v_0 \exists v_1[\mathbf{val}(\sigma, 1, v_1) \wedge A(v_0, v_1)]$$

and this holds for any σ.

Now suppose $A(0, v_1)$ and $\forall v_0(A(v_0, v_1) \rightarrow A(Sv_0, v_1))$. Then there is a sequence χ such that $\mathbf{val}(\chi, 1, v_1)$. Hence, for this χ, we have: $A(v_0, v_1) \equiv \exists v_1(\mathbf{val}(\chi, 1, v_1) \wedge A(v_0, v_1))$. So the induction axiom becomes:

$$A(0, v_1) \wedge \forall v_0(A(v_0, v_1) \rightarrow A(Sv_0, v_1)) \rightarrow \forall v_0(A(v_0))$$

as wanted. □

This last result is relevant to Volker Halbach's (2001a) claim that the 'uniform disquotation scheme'—our $(\cdot)^{S+}$—is plausibly analytic, since PA^{S+} is a conservative extension of *PA*. What we have just seen, however, is that this result depends crucially upon the choice of *PA* as base theory. Whether one takes conservativity to be required for analyticity or regards it as merely indicative of it, the uniform disquotation scheme appears to be logically quite strong, transforming a theory interpretable in Q into one that contains *PA*. It is only in very special cases that it gives us nothing we did not already have.

6. Disentangling Syntax from the Object-Language

6.1. **Reviving an Old Approach to Truth-theories.** What's responsible for the unfortunate Corollary 5.16?

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 extended 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. We need to know that all the axioms of the object-theory are true, and the idea is to get their truth from them: We assume the axioms themselves and derive their truth from their T-sentences.

As we have seen, however, that is not at all how things work in the case of *PA*. The truth of *all* the axioms of *PA* is not derived from the axioms of *PA* in that way, 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 axiom, but that's it. This is what leads to Corollary 5.16: The truth of all the axioms of *PA* is a consequence, not of those axioms, but of a single instance of induction. Speaking more generally, the problem is that a single theory is playing both of the roles I just distinguished: In $(I\Sigma_1)^{T+}$, $I\Sigma_1$ is *both* the

underlying syntax *and* what provides us with the axioms of the theory we had meant to be reasoning about. So induction axioms that were introduced to allow us to formalize certain sorts of syntactic arguments have instances that entail the truth of principles in the object-language that go beyond what we'd meant to be assuming.

The solution to the problem is therefore obvious: We need to disentangle the syntax from the object-language. And, interestingly enough, this is how Tarski himself proceeds in "The Concept of Truth in Formalized Languages". I quoted Tarski's description of the meta-theory in which he proposes to define truth earlier. Here is his description of the meta-*language*:

> 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. 210–11)

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, in section 3, of the calculus of classes (Tarski, 1958, pp. 172ff). Tarski was of course aware—at least by the time his paper was published—that the syntactic theory 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 familiar maneuver. So the basic idea of separating the syntax from the object-theory is old, even if the application I propose to make of it is somewhat new.

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 our syntax. The most natural choice for $\mathcal{S}$, and the one that would be closest to Tarksi's original intentions, would be a theory of concatenation (Quine, 1946; Tarski et al., 1953; Grzegorczyk, 2005); this would also have the advantage that what follows would be independent of issues about how we code expressions. To keep things familiar, however, we shall take $\mathcal{S}$ to be isomorphic to the language of arithmetic.[47] (Think of $\mathcal{S}$ as the language of arithmetic written in boldface, or something

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

of the sort.) Our theory of syntax can then be taken be $\boldsymbol{Q}$, or $I\Delta_0$, or whatever we wish.

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. But the quantifiers in sentences of $\mathcal{S}$ do not range over all sets, even if (perhaps) they range over some of them. So we need to keep the domains of $\mathcal{S}$ and $\mathcal{L}$ separate somehow. There are various ways to do this. Perhaps the simplest 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.[48]

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:

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

As before, $\beta \overset{v}{\sim} \alpha$ abbreviates: $\forall w\,[\mathbf{var}(w) \wedge v \neq w \rightarrow \mathbf{val}(\beta, w) = \mathbf{val}(\alpha, w))]$. What this says is thus 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.

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

v: $\quad \mathbf{var}(v) \rightarrow \mathbf{Den}_\alpha(v, \mathbf{x}) \equiv \mathbf{x} = \mathbf{val}(\alpha, v)$

$\wedge$: $\quad \mathbf{Sat}_\alpha(\ulcorner A \wedge B \urcorner) \equiv \mathbf{Sat}_\alpha(A) \wedge \mathbf{Sat}_\alpha(B)$

$\forall$: $\quad \mathbf{Sat}_\alpha(\ulcorner \forall \mathbf{v}_i A(\mathbf{v}_i) \urcorner) \equiv \forall\beta[\beta \overset{i}{\sim} \alpha \rightarrow \mathbf{Sat}_\beta(\ulcorner A(\mathbf{x}) \urcorner)$

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:

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

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

[48] 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.

0: $\quad \mathrm{Den}_\alpha(\text{`}0\text{'}, \mathbf{x}) \equiv \mathbf{x} = 0$

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

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:

Definition. Let $\mathcal{T}$ be an arithmetical theory. Then $\widehat{\mathcal{T}}^{T_\mathcal{L}}$ is the semantics for $\mathcal{L}$ we have just described.

We can think of $\widehat{\xi}^{T_\eta}$ as a two-place functor: Given a theory and a language, it returns a new theory that constitutes a semantics for that language based upon the original theory as syntax. Our interest is in the properties of this functor.

Note that we are not (yet) extending any induction scheme that might be present in $\mathcal{T}$. So $\widehat{\mathcal{T}}^{T_\mathcal{L}}$ is not going to be formalizing semantic consistency proofs of the sort discussed in Section 5.4. More generally, induction in $\widehat{\mathcal{T}}^{T_\mathcal{L}}$ does not apply to statements involving assignments, or semantics, or the object-language. The induction axioms must be 'purely syntactical'.

6.2. **The Weakness of Compositional Truth-theories.** We now prove a strong generalization of Corollary 3.5, which told us that we can formalize a materially adequate theory of truth for the language of arithmetic in the weak theory we called ${Q_{\text{seq}}}^T$. It turns out that we can do this for *any* finite language, and we can do it in a theory interpretable in $\boldsymbol{Q}$.[49]

Lemma 6.1. *$\widehat{Q}^{T_\mathcal{L}}$ is a materially adequate theory of truth for $\mathcal{L}$. That is, $\widehat{Q}^{T_\mathcal{L}}$ proves*

$$\boldsymbol{T}(\ulcorner A \urcorner) \equiv A$$

for each sentence A of $\mathcal{L}$.

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

The first lesson we learn here, then, is 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.

This result plays an important role in Craig and Vaught's (1958) proof that every axiomatizable theory that has no finite models has a finitely

[49] I think we can do something similar even for non-finite languages. In this case, it'll be local interpretability that's of interest, rather than interpretability. But I'm not sure about this.

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**, more or less—a weak theory of assignments, and the Tarski clauses for the language of $\mathcal{T}$. That's enough to 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).

Thus, $\widehat{Q}^{T_{\mathcal{L}}}$ is not a mere curiosity but is of actual mathematical utility. It is also as weak as it is possible for it to be.

Lemma 6.2. *$\widehat{Q}^{T_{\mathcal{L}}}$ 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 is then completely trivial: $\mathrm{val}(v, \mathrm{x})$ will always be true, for each v and x. A semantic theory for the language, so interpreted, is then easily constructed. □

Lemmas 6.1 and 6.2 give us a first indication of why it is worth disentangling syntax from the object-theory. Together, they imply that there is a materially adequate truth-theory for the language of arithmetic that is as weak as it could plausibly be: It is interpretable in **Q**. If, on the other hand, 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 ${Q_{\mathrm{seq}}}^{T}$. And it follows from Theorem 5.3 that ${Q_{\mathrm{seq}}}^{T}$ is *not* interpretable in **Q**. So Lemma 6.1 is an improvement on Theorem 5.3.[50]

6.3. **The Strength of Compositional Truth-theories, and the Weakness of Disquotational Ones.** We know, then, from Lemma 6.2, that $\widehat{Q}^{T_{\mathcal{L}}}$ is very weak. Unfortunately, however, this does not really help us to characterize the strength of truth-theories. For one thing, the interpretation of $\widehat{Q}^{T_{\mathcal{L}}}$ in Q wreaks havoc on the meanings of the primitives of $\mathcal{L}$: It

[50]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, for reasons similar to those mentioned in connection with the problem of 'essential richness', that this would credit ordinary speakers with far too much tacit knowledge. But knowing such a theory need involve no more than knowing $\widehat{Q}^{T_{\mathcal{L}}}$, and the logical strength of that theory derives entirely from its syntactic component.

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 naturally interpret Tarski as taking the object-theory to play something like this role. (Though we do not need to suppose, as Tarski may have, that $\mathcal{L}$ must in any sense consist of 'meaning postulates'.)[51] 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 $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ compares, in logical strength, to $\mathcal{T}$. From this point of view, Lemma 6.2 concerns the special case where $\mathcal{T}$ is the null theory.

As was explained in Section 2.1, there are different ways of comparing theories, so we can ask various sorts of questions about the relationship between $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ and $\mathcal{T}$. One question is whether $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ is a conservative extension of $\mathcal{T}$. And we have, in fact, already seen that it is: That is the result of Craig and Vaught's (1958) mentioned earlier, in a slightly different form.

But there is a different, and ultimately more interesting, question we can also ask, namely, whether $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ is interpretable in $\mathcal{T}$. And to this question, the answer is "no", at least if $\mathcal{T}$ is finitely axiomatized.

Theorem 6.3. *Let $\mathcal{T}$ be a consistent theory in $\mathcal{L}$. Then $\widehat{Q}^{T_\mathcal{L}}$ plus 'all axioms of $\mathcal{T}$ are true' proves the consistency of $\mathcal{T}$ on a cut.*

Corollary 6.4. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$. Then $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ proves the consistency of $\mathcal{T}$ on a cut and so is not interpretable in $\mathcal{T}$.*[52]

The proof of Theorem 6.3 is essentially the same as that of Theorem 5.3. Corollary 6.4 then follows from Corollary 4.7 and the obvious analogue of Corollary 3.3.

It's worth emphasizing that the *only* role $\mathcal{T}$ plays in the proof is in allowing us to prove that all $\mathcal{T}$'s axioms are true. The work is all done in $\widehat{Q}^{T_\mathcal{L}}$.

We also get an analogue of Corollary 5.6.

Theorem 6.5. *Let $\mathcal{T}$ be a finitely axiomatized, consistent theory in $\mathcal{L}$. Then $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ is mutually interpretable with $Q + \mathrm{Con}(\mathcal{T})$.*

Proof. That $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ interprets $Q + \mathrm{Con}(\mathcal{T})$ follows from Theorem 6.3. The other direction is just a minor modification of Theorem 5.5. □

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

[52] Note that it follows that the finitely axiomatizable theory Craig and Vaught show is a conservative extension of $\mathcal{T}$ is not interpretable in $\mathcal{T}$.

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 $\mathcal{T}$—and add it in a way that is guaranteed not to 'infect' $\mathcal{T}$ itself—the result is a theory that is logically stronger than $\mathcal{T}$ in the sense that it is not interpretable in $\mathcal{T}$.[53]

Even from a purely technical point of view, then, $\widehat{Q}^{T_\mathcal{L}}$ is an interesting theory. It is as weak as it can be, yet $\widehat{Q}^{T_\mathcal{L}}$ 'upGödels' any finitely axiomatized theory $\mathcal{T}$ that you care to give it.[54] We can think of Pudlák's form of the incompleteness theorem as defining a map on theories: Given a consistent, sequential theory $\mathcal{T}$ containing Q, it hands us $Q + \mathrm{Con}(\mathcal{T})$, which is guaranteed to be logically stronger than $\mathcal{T}$ in the sense that it is not interpretable in $\mathcal{T}$. In effect, what we have found is that, for finitely axiomatized theories, $\widehat{Q}^{T_\mathcal{L}}$ can be used to define the same functor, *modulo* interpretabilty: Given a finitely axiomatized theory $\mathcal{T}$ in $\mathcal{L}$, $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ is mutually interpretable with $Q + \mathrm{Con}(\mathcal{T})$.[55]

By contrast, the T-sentences themselves have *no* logical power, *even if* we extend whatever induction scheme might happen to be available. Once again, getting a completely general version of this result is hard, because schemes can come in so many different forms. But if, as before, we focus just on the case of the usual hierarchy, then the claim can be stated precisely. And again, the results here significantly improve the corresponding results of Section 5.3, which is yet another reason to want to disentangle syntax from the object-theory.

Definition. $\widehat{\mathcal{T}}^{D+_\mathcal{A}}$ is the theory of truth for the language of arithmetic that is similar to $\widehat{\mathcal{T}}^{T_\mathcal{A}}$ but, instead of containing a compositional theory of truth contains the T-sentences for $\mathcal{A}$ and extends the induction scheme to permit the presence of the truth-predicate.

By essentially the same argument as for Theorem 5.12, we have:

Proposition 6.6. $\widehat{I\Sigma_n}^{D+_\mathcal{A}}$ *is locally interpretable in* $I\Sigma_n$.

It is clear that we thus also have:

[53] It would, I think, be well worth investigating such theories as $\widehat{I\Sigma_1}^{T_\mathcal{L}}$. It's of course immediate that $\widehat{I\Sigma_1}^{T_\mathcal{L}} + \mathcal{T}$ is not interpretable in $\mathcal{T}$, since even $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$ isn't. But is there some nice characterization of exactly how strong $\widehat{I\Sigma_1}^{T_\mathcal{L}} + \mathcal{T}$ is? In general, one would suppose it is stronger than $\widehat{Q}^{T_\mathcal{L}} + \mathcal{T}$; surely it isn't interpretable in $Q + \mathrm{Con}(\mathcal{T})$. On the other hand, one would suppose that $\widehat{I\Sigma_1}^{T_\mathcal{L}} + \mathcal{T}$ is weaker than $\widehat{I\Sigma_1}^{T^+_\mathcal{L}} + \mathcal{T}$ (for which, see below) and, in particular, that it does not interpret $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$. So where precisely does it sit? And what of intermediate theories, like $\widehat{I\Delta_0}^{T_\mathcal{L}} + \mathcal{T}$?

[54] Thanks to Visser for the wonderful neologism.

[55] Note that if we use a theory of concatenation as our base theory, then this result is coding-free and so gives us a co-ordinate-free account of what consistency statements are, like the central result of Visser's paper on the second incompleteness theorem (Visser, 2009a, theorem 4.1).

Proposition 6.7. *$\widehat{I\Sigma_n}^{D+\mathcal{A}} + \mathcal{T}$ is locally interpretable in $I\Sigma_n + \mathcal{T}$.*

And so, in particular:

Corollary 6.8. *$\widehat{I\Sigma_n}^{D+\mathcal{A}} + I\Sigma_m$ is locally interpretable in $I\Sigma_k$, where $k = \max(m, n)$.*

Proof. $\widehat{I\Sigma_n}^{D+\mathcal{A}} + I\Sigma_m$ is locally interpretable in $I\Sigma_n + I\Sigma_m$, where these two theories are formulated in disjoint copies of the language of arithmetic. But $I\Sigma_n + I\Sigma_m$ will obviously be interpretable in $I\Sigma_{\max(m,n)}$. □

So we get an analogue of Theorem 5.11.

Corollary 6.9. *$\widehat{PA}^{D+\mathcal{A}} + PA$ is interpretable in PA.*

Proof. Any finite fragment of this theory is contained in one or another of the $\widehat{I\Sigma_n}^{D+\mathcal{A}} + I\Sigma_m$. So each finite fragment is interpretable in $I\Sigma_n$, for some n, and so in *PA*. That establishes local interpretability, and now we invoke Orey's Theorem. □

We'll see shortly that something even stronger is true.[56]

6.4. **Semantic Consistency Proofs, Again.** We have seen that $\widehat{Q}^{T_{\mathcal{L}}} + \mathcal{T}$ 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 the usual semantic proof of the consistency of $\mathcal{T}$. That leaves us more or less where we were at the end of Section 5.1. 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 5.4, 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, this answer turned out to be correct but disappointing. It's true that $(I\Sigma_1)^{T+}$ proves the consistency of $I\Sigma_1$, but it also proves the consistency of *PA*.

The work we did in the last section puts us in a position to resolve this problem. What we need to add is, indeed, something along the lines of 'induction for Σ_1 formulae'. But the problem that infected our earlier efforts has now been resolved: We can 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 a number-theoretic 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. 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 $\mathcal{T}^{T+}$.

[56] What happens if we consider theories related to $\widehat{\mathcal{T}}^{D+\mathcal{A}}$ as $\mathcal{T}^{S+}$ is related to $\mathcal{T}^{D+}$?

Definition. $\widehat{\mathcal{T}}^{T^+_{\mathcal{L}}}$ is $\widehat{\mathcal{T}}^{T_{\mathcal{L}}}$ with the induction axioms in $\mathcal{T}$ extended to permit semantic vocabulary and reference to assignments.

As before, this definition isn't perfectly general. But we know how to apply it to the cases that matter here: $\widehat{I\Sigma_n}^{T^+_{\mathcal{L}}}$ is really $\widehat{I\Sigma_n(\mathrm{Sem})}^{T_{\mathcal{L}}}$, where Sem is the set of atomic formulae of the forms: $\mathrm{Den}_\alpha(t, \mathbf{x})$; $\mathrm{Sat}_\alpha(x)$; $\mathrm{T}(x)$; and $\mathrm{val}(\sigma, x)$.

It's important to appreciate that, in extending induction in this way, we are not extending it nearly as far as we might extend it. The only new induction axioms we are allowing are ones that contain the semantic predicates mentioned. For example, suppose that $\mathcal{L}$ is the language of set-theory, and consider the following formulae (which are chosen more or less randomly):

$$\exists \mathbf{x} \mathrm{Den}_\sigma(t, \mathbf{x})$$
$$\mathbf{x} \in \mathbf{y} \wedge \mathrm{val}(\sigma, v) = \mathbf{x} \wedge \mathrm{Sat}_\sigma(z)$$

We do *not* have induction for such formulae in $\widehat{I\Sigma_n}^{T^+_{\mathcal{L}}}$, as I am understanding it. The first is ruled out because it contains the quantifier '$\exists$x', which ranges not over numbers but over sets. The second is ruled out because it contains the predicate $\in$.

I am not terribly happy about this restriction. By imposing it, we force ourselves to operate with what might seem like an unnaturally weak theory, and the significance of the results we shall prove about what it can or, more importantly, cannot do might therefore be questioned. It's my hope that there will prove to be some natural way of loosening these restrictions and allowing induction over the formulae mentioned, and others of their general kind, without adding (significant?) strength to $\widehat{I\Sigma_n}^{T^+_{\mathcal{L}}}$. Part of the difficulty is that it is hard to know how to integrate quantifiers over sets (or, more generally, whatever the object-language talks about) into our measure of logical complexity: What sorts of formulae count as Σ_n, in the relevant sense, if these formulae may contain quantifiers over both numbers and sets? But let us leave such questions aside for now.

It is clear that we can now adapt the arguments given in Section 5.4 to our new framework. In particular, we will be able to formalize a semantic proof of $\mathrm{Con}(\mathcal{T})$ is $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$, where $\mathcal{L}$ of course is the language of $\mathcal{T}$. So we have:

Theorem 6.10. *Let $\mathcal{T}$ be a theory in a finite relational language $\mathcal{L}$, and suppose that $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$ proves that all axioms of $\mathcal{T}$ are true. Then $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T} \vdash \mathrm{Con}(\mathcal{T})$.*

The restriction to relational languages is essential, because, we cannot prove that every term has a denotation. The obvious way to do so would be by induction on $\exists \mathbf{x} \mathrm{Den}_\sigma(t, \mathbf{x})$, but, for the reasons just mentioned, we

do not have induction for this predicate in $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}}$. As earlier, however, the restriction can then be lifted, since $I\Sigma_1$ will prove that a theory $\mathcal{T}$ stated in a non-relational language can always be interpreted in its relational counterpart $\mathcal{T}_R$ and so that, if $\mathrm{Con}(\mathcal{T}_R)$, then $\mathrm{Con}(\mathcal{T})$. So we have:

Corollary 6.11. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language $\mathcal{L}$. Then $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T} \vdash \mathrm{Con}(\mathcal{T})$.*

There is thus a sense in which Corollary 5.16, though surprising in a way, is in another way natural. The syntax needed to carry out semantic consistency proofs is no more than can be formalized in $I\Sigma_1$. If Corollary 5.16 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 the axioms of *PA* are true. Well, we don't. 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... for the establishment of all" axioms of *PA*.

What we want to see now, then, is that our way of disentangling syntax from the object-language really does solve the problem Corollary 5.16 revealed. What we would like to be able to show is that, although $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + I\Sigma_1$ proves $\mathrm{Con}(I\Sigma_1)$, it does *not* prove $\mathrm{Con}(PA)$ or even $\mathrm{Con}(I\Sigma_2)$. To show this, we will establish a sort of converse of Corollary 6.11.

Theorem 6.12. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is interpretable in $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$.*

Before we begin the proof, let me note a couple important corollaries.

Corollary 6.13. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is mutually interpretable with $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$.*

This follows immediately, since $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$ contains $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$. We can generalize yet further.

Corollary 6.14. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then, if $n \geq 1$, $\widehat{I\Sigma_n}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is mutually interpretable with $I\Sigma_n + \mathrm{Con}(\mathcal{T})$.*

We'll prove Corollary 6.14 after we prove Theorem 6.12.

We are going to need a version of the so-called arithmetized completeness theorem (Hájek and Pudlák, 1993, pp. 104–5), which is provable in $I\Sigma_1$. There are two different ways one often sees this theorem stated, and the proof of Theorem 6.12 rests upon the way these two statements of it relate to one another.

Theorem 6.15 (Arithmetized Completeness Theorem)**.** *Let $\mathcal{T}$ be a recursively axiomatized theory. Then:*

(1) $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$ *interprets* $\mathcal{T}$.
(2) $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$ *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.[57] 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.[58] That the model *is* a model of $\mathcal{T}$ amounts to its being provable, in $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$, that the axioms of $\mathcal{T}$ are, in the sense of truth associated with the model, true, that is, that they are true in the model.

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 understand it very well.[59] I will explain why the complexity of the model matters—note that it is independent of $\mathcal{T}$—and why its being LL_1 is enough for the proof of Theorem 6.12.

Proof of Theorem 6.12. If we are going to intepret $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$ in $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$, we need to deal with three things: (i) $\mathcal{T}$; (ii) the semantic theory for $\mathcal{L}$, including the theory of assignments; and (iii) the underlying syntax, $I\Sigma_1$. A significant part of the last will be no problem, since we already have $I\Sigma_1$ 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 $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$ 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 Q. Then '0' is a term, and among the axioms of

[57] It is not 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 also 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 $I\Sigma_1$ proves the Tarski clauses for Σ_n formulae and therefore proves, for each Σ_n formula $A(x)$ the Sat-sentence: $\mathrm{Sat}_{n,\sigma}(\ulcorner A(v_0)\urcorner) \equiv A(\mathrm{val}(\sigma, 0))$. One can therefore use Σ_n formulae as codes for Σ_n-definable sets when working in $I\Sigma_1$ (Hájek and Pudlák, 1993, §I.1(d), esp. p. 60, Remark 1.80).

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

[59] The definition is on p. 85 of Hajék and Pudlák's book, for those who would like to explore it. Thanks to Ali Enayat for making it a *little* clearer to me.

$\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + Q$ that we need to interpret are these two:

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

The first comes from **Q** 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 $I\Sigma_1 + \mathrm{Con}(Q)$ both interpets **Q** and gives us a semantics for the language of **Q** 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 0.

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: $<x, y> \in S$.[60] And, of course, you restrict the quantifiers to the domain of the model. The fact that $\mathcal{T}$ is provably true in the model 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 '$S0$'. Some fiddling may be necessary here and there to get everything completely in sync, but this is merely tedious.

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, $I\Sigma_1$. As noted earlier, much of that is trivial, since we're working in $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$ and so have $I\Sigma_1$ readily available. So if we were just trying to interpret $\widehat{I\Sigma_1}^{T_{\mathcal{L}}} + \mathcal{T}$, we'd be done. What we're actually trying to interpret, however, is $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}} + \mathcal{T}$, and so what we lack at this point—all we lack—is a demonstration that the extended induction axioms can be proven in $I\Sigma_1 + \mathrm{Con}(\mathcal{T})$, 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 $\mathrm{Sat}_\alpha(x)$ were, say, Σ_2, then we'd have no hope whatsoever of

[60] 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; and we then use that very formula to construct the translation of $R(x, y)$.

proving the translations of induction axioms containing $\mathrm{Sat}_\alpha(x)$ in $I\Sigma_1$.[61] But we know that $\mathrm{Sat}_\alpha(x)$ and its friends are LL_1. The induction axioms we're trying to prove are, therefore, of the form $\exists x\phi(x)$, where $\phi(x)$ is built from atomic arithmetical formulae and the translations of our atomic semantic formulae: $\mathrm{Den}_\alpha(t, \mathrm{x})$; $\mathrm{Sat}_\alpha(x)$; $\mathrm{T}(x)$; and $\mathrm{val}(\sigma, x)$. Since these are at worst LL_1, the induction axioms we're trying to prove are $\Sigma_1(LL_1)$. And it just so happens that $I\Sigma_1$ proves induction for $\Sigma_1(LL_1)$ formulae (Hájek and Pudlák, 1993, p. 85, lemma 2.78). □

It's just beautiful the way this works out: LL_1 is *precisely* what the model needs to be for that last step to work.

Proof of Corollary 6.14. There are various ways of proving this. One is to note that, in $I\Sigma_2$, we get a better bound on the complexity of the model: It's *low* Δ_2. So then the question is whether $I\Sigma_n$ proves induction for $\Sigma_n(\Delta_2)$ formulae, when $n \geq 2$. It does (Hájek and Pudlák, 1993, p. 82, theorem 2.67). □

It's a nice question whether this also extends to *PA*—that is, to the case where *PA* is our theory of *syntax*, rather than the object-theory. We clearly have this:

Corollary 6.16. *Let $\mathcal{T}$ be a finitely axiomatized theory in a finite language. Then $\widehat{PA}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is mutually locally interpretable with $PA + \mathrm{Con}(\mathcal{T})$.*

Proof. Each finite fragment of $PA + \mathrm{Con}(\mathcal{T})$ is contained in one of the $I\Sigma_n + \mathrm{Con}(\mathcal{T})$, which is interpretable in $\widehat{I\Sigma_n}^{T^+_{\mathcal{L}}} + \mathcal{T}$ and so in $\widehat{PA}^{T^+_{\mathcal{L}}} + \mathcal{T}$. The converse is similar. □

The reflexivity of *PA* entails that of $PA + \mathrm{Con}(\mathcal{T})$,[62] so $\widehat{PA}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is globally interpretable in $PA + \mathrm{Con}(\mathcal{T})$, by Orey's Theorem. It is not at all obvious, however—to me, anyway—that $\widehat{PA}^{T^+_{\mathcal{L}}} + \mathcal{T}$ must be reflexive. It would be nice if it was, though, since then we could remove "locally" from Corollary 6.16.

6.5. **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 6.17. *$\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + PA$ is interpretable in PA.*

[61] Note, though, that it's nonetheless clear that this is going to work at some level or other, given that the complexity of the model is independent of $\mathcal{T}$. In the situation just mentioned, for example, we'd be perfectly fine at $I\Sigma_2$.

[62] Mostowski shows that every extension of *PA* that does not expand the language is reflexive.

Proof. Any finite fragment of $\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + PA$ is contained in one of the theories: $\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + I\Sigma_n$ and so by Corollary 6.14 is interpretable in $I\Sigma_m + \mathrm{Con}(I\Sigma_n)$. But ***PA***, being reflexive, contains every such theory. So every finite fragment of $\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + PA$ is interpretable in ***PA***, which shows that $\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + PA$ is locally interpretable in ***PA***. Now invoke Orey's Theorem. □

Corollary 6.18. *$\widehat{PA}^{T^+_{\mathcal{L}}} + PA$ is interpretable in **PA**.*

Proof. Any finite fragment of $\widehat{PA}^{T^+_{\mathcal{L}}} + PA$ is contained in one of the theories: $\widehat{I\Sigma_m}^{T^+_{\mathcal{L}}} + PA$. So $\widehat{PA}^{T^+_{\mathcal{L}}} + PA$ is locally interpretable in ***PA*** and so is interpretable in ***PA***. □

On the other hand:

Corollary 6.19. *$\widehat{PA}^{T^+_{\mathcal{L}}}$ + plus 'all axioms of **PA** are true' proves* $\mathrm{Con}(PA)$. *Indeed, $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}}$ plus 'all axioms of **PA** are true' proves* $\mathrm{Con}(PA)$.

Proof. From Theorem 6.10. □

It follows, obviously, that $\widehat{PA}^{T^+_{\mathcal{L}}} + PA$ does not prove that all axioms of ***PA*** are true. What this means is that, once we have disentangled the syntax from the object-language, the 'happy accident' that permits PA^{T+} to prove $\mathrm{Con}(PA)$ 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 PA^{T+} proves $\mathrm{Con}(PA)$.

The combination of Corollary 6.18 and Corollary 6.19 is notable for another reason, as well. Deflationists about truth typically hold that the only legitimate use for the truth-predicate is as a 'device of generalization'. Precisely what that is supposed to mean has never been made terribly clear. But one thing one *might* have thought it meant, or at least implied, was something like: Assuming we have the T-sentences for some language $\mathcal{L}$, then a theory consisting of the sentences in some set $\mathcal{S}$ is in some natural sense equivalent to the theory containing the single statement "All sentences in $\mathcal{S}$ are true".[63] Indeed, the one attempt known to me to explain what it might mean to "use the truth-predicate as a device of generalization" proceeds along precisely these lines (Halbach, 1999). Considered as additions to $\widehat{PA}^{T^+_{\mathcal{L}}}$, however, or even as additions to $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}}$, the two theories consisting of the axioms of ***PA***, on the one hand, and the single statement "All axioms of ***PA*** are true", on the other,

[63] It sometimes seems to be supposed that the fact that the truth-predicate is a 'device of generalization' is suitably explained by the fact that the truth-predicate allows us to form sentences like the one just mentioned, which is of course a generalization. But "All axioms of PA contain the symbol '→'" is a generalization, too. *Every* predicate allows us to form generalizations we could not form without it. In what sense, then, is the truth-predicate supposed *just* to be a 'device of generalization'?

have very different logical properties: The latter is a *lot* stronger than the former.

That's not to say, of course, that there's not some other way of explaining what it means to 'use the truth predicate merely as a device of generalization'. But I don't know what that would be.

7. Concluding Philosophical Reflections

Taken together, the results proven in Section 6.4 show that, where $\mathcal{T}$ is a finitely axiomatized theory in $\mathcal{L}$, $\widehat{\mathcal{U}}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is mutually interpretable with $\mathcal{U} + \mathrm{Con}(\mathcal{T})$, for a very wide range of choices for $\mathcal{U}$. This holds, in particular, for ***Q***, for $I\Sigma_n$, so long as $n \geq 1$, and even for ***PA*** (though in this last case we have only mutual local interpretability). It would be nice to know if more cases could be added to the list. I do not know whether $\widehat{I\Delta_0}^{T^+_{\mathcal{L}}} + \mathcal{T}$ is mutually interpretable with $I\Delta_0 + \mathrm{Con}(\mathcal{T})$. But even without an answer to that question, or even if the answer turns out to be negative, the fact that these theories are mutually interpretable in so many cases gives us reason to suppose that the connection between truth-theories and consistency-statements we have been exploring is robust. More precisely, what it shows is that a compositional truth-theory amounts to a kind of abstract consistency statement: If you build a truth-theory for $\mathcal{L}$ on top of an appropriate syntax $\mathcal{S}$, and then hand it a finitely axiomatized theory $\mathcal{T}$ in the language it concerns, it hands you back $\mathcal{S}$ plus the consistency statement for $\mathcal{T}$.

This is despite the fact that, as we have seen, there is another sense in which a compositional truth-theory adds nothing at all to the underlying syntax: $\widehat{I\Sigma_1}^{T^+_{\mathcal{L}}}$ is interpretable in $I\Sigma_1$, by the same argument as given for Lemma 6.2.[64] What that shows is that one needs to be very careful about how one measures the strength of truth-theories.

This observation is relevant to another issue that comes up in the literature on deflationism. Facts about what happens when one adds various semantic assumptions to ***PA*** play a critical role in the discussion of the so-called conservativeness argument, championed by Stewart Shapiro (1998) and Jeffrey Ketland (1999). The argument emerges from the thought that a deflationary truth-predicate, being in some sense 'insubstantial', ought not to allow us to prove anything we cannot prove without it. That is: $\mathcal{U}^{T+}$ ought to be a conservative extension of $\mathcal{U}$. But, of course, PA^{T+} proves $\mathrm{Con}(PA)$ and so isn't a conservative extension of ***PA***.

In his response to this argument, Hartry Field places very heavy weight upon the fact that the non-conservativity result depends essentially upon the presence of the new induction axioms and is *not* due

[64] Indeed, it would appear that, so long as $\mathcal{T} \supseteq Q$, $\widehat{\mathcal{T}}^{T^+_{\mathcal{L}}}$ is going to be interpretable in $\mathcal{T}$.

simply to the presence of a compositional truth-theory.[65] In particular, if we add a compositional truth-theory to *PA* without extending induction, then the result is a conservative extension of *PA* (Parsons, 1981, pp. 213–15). And so Field writes:

> Since truth can be added in ways that produce a conservative extension. . . , there is no need to disagree with Shapiro when he says that "conservativeness is essential to deflationism". . . . Shapiro's position, however, is that a deflationist must hold that adding 'true' to number theory in the full-blooded way that involves [extending the induction axioms also] produces a conservative extension. (1999, p. 536)

Field goes on to argue that a deflationist need hold no such thing. At most, the deflationist should hold that the principles that are "essential to truth"—that flow from its disquotational nature—are conservative over number theory. But, Field claims, the induction principles are not "essential to truth" in the relevant sense: Their truth flows not from the nature of truth, but from the nature of the natural numbers. They are not semantical but arithmetical in character.

I more or less agree with this last point, though what Field ought to have said is that the induction axioms are *syntactic* in character, not arithmetical. What our discussion here shows, however, is that Field's emphasis on induction is misplaced.

In what seems to me the crucial passage, Field quotes Shapiro's (1998, p. 499) question: "How thin can the notion of arithmetical truth be if, by invoking it, we can learn more about the natural numbers?" and then replies:

> . . . [T]he way in which we "learn more about the natural numbers by invoking truth" is that in having that notion we can rigorously formulate a more powerful arithmetical theory than we could rigorously formulate before. There is nothing very special about truth here: using any other notion not expressible in the original language we can get new instances of induction, and in many cases these lead to nonconservative extensions. (Field, 1999, p. 536)

There are two respects in which this is at best misleading.

What does Field mean by "using [a] notion not expressible in the original language"? The natural way to read him would be as talking about definability, about what happens if we add a new predicate that allows us to define a set not definable in the original language. In that case, Field would be saying this:

[65] See also Halbach's (2001b) treatment of the argument.

Base: *PA*	No New Induction	Extend Induction
Add the T-sentences	Conservative Interpretable	PA^{D+} Conservative Interpretable
Add a fully compositional truth-theory	PA^{T} Conservative Interpretable	PA^{T+} Non-conservative Not Intepretable

TABLE 1. Some Mathematical Facts

> If we add a new predicate that defines a set not definable in the original language, we can get new instances of induction, which may lead to new theorems in the original language.

That is of course right. We *can* get new instances that *may* lead to new theorems. But the case of the truth-predicate is precisely not one of those cases. Tarski's Theorem tells us that the set of truths of the language of *PA* is not definable in the language of arithmetic. This has nothing to do with whether we add a fully compositional truth-theory, as opposed just to adding just the T-sentences. Either way, we will be able to define a set we could not previously define: It will be defined by $\mathrm{T}(x)$. Let me say that again. If we add a truth-predicate $\mathrm{T}(x)$ to the language of *PA* and extend *PA* by adding the T-sentences, then that is enough to guarantee that $\mathrm{T}(x)$ defines the set of truths of the language of *PA* and so defines a set not definable in the original language. But even if we extend induction, the result is still a conservative extension of *PA*. The moral, then, is supposed to be this: The non-conservativity result is *not* due just the presence of "new instances of induction" formulated using a "notion not expressible in the original language". The presence of a fully compositional truth-theory is essential. Indeed, what I should like to say is that what is most responsible for the non-conservativity result is the compositional truth-theory, not the extension of induction.

When truth-theories are considered simply as additions to *PA*, it is essentially impossible to disentangle the contributions being made by the truth-theory, on the one hand, and the extension of induction, on the other. The mathematical facts are summarized in table 1. So long as we do not *both* add a compositional truth-theory *and* extend the induction scheme, the resulting theory is conservative over *PA* and interpretable in *PA*. How are we supposed to choose whom to blame, then? Does it even make sense to blame one rather than the other?

We have seen, however, that *PA* is a special case. We have also seen that the usual way of formalizing truth-theories can lead to peculiar

Base: $\mathcal{U}+\mathcal{T}$	No New Induction	Extend Induction
Add the T-sentences	Locally Interpretable	$\widehat{\mathcal{U}}^{D+_{\mathcal{A}}}+\mathcal{T}$ Locally Interpretable
Add a fully compositional truth-theory	$\widehat{\mathcal{U}}^{T_{\mathcal{L}}}+\mathcal{T}$ Not interpretable	$\widehat{\mathcal{U}}^{T^{+}_{\mathcal{L}}}+\mathcal{T}$ Not interpretable, and stronger still

TABLE 2. Some Other Mathematical Facts ($\mathcal{U} = I\Sigma_n, PA$)

phenomena. And if we now look again, with these lessons in mind—if we focus not on conservativity but on interpretability;[66] if we make sure the object-theory is finitely axiomatized; and if we disentangle the syntax from the object-theory—then the facts, summarized in table 2, look very different.[67] What we see is that adding the compositional principles results in an increase in logical strength *whether or not* we extend the induction axioms.[68] Extending the induction axioms, on the other hand, results in an increase in strength only in the presence of a fully compositional truth-theory. That suggests, to me, anyway, that it is the compositional truth-theory that is doing the work here.

By themselves, of course, these observations do not pose a serious problem for anyone, deflationists included. They do, however, make it clear that, as a matter of mathematical fact, the compositional principles have substantial logical strength. That makes it worth asking how deflationists intend to earn a right to them: The various principles comprising a compositional truth-theory cannot be regarded as a collection of trivialities on the order of the T-sentences.[69] But that issue, I shall have to discuss on another occasion.[70]

[66]Do the conservativity results still hold when we consider weaker theories? Is ${I\Sigma_1}^T$ a conservative extension of $I\Sigma_1$? The proof of this, in the case of *PA*, is far from trivial, and I have no idea whether it works for weaker theories.

[67]I'm assuming here, of course, that $\mathcal{T}$ is not a theory whose consistency is independently provable in $I\Sigma_1$, e.g., *Q*: The methods used above show that $\widehat{I\Sigma_1}^{T^{+}_{\mathcal{L}}}+Q$ is interpretable in $I\Sigma_1$, since it is interpretable in $I\Sigma_1+\mathrm{Con}(Q)$, and $I\Sigma_1$ proves $\mathrm{Con}(Q)$.

[68]Since these results hold even with $\mathcal{U} = PA$, we do not need to restrict the *syntax* in any way to get these sorts of results. What matters is that the *object-theory* should be finitely axiomatized. But that is simply because we can't hope to prove that *all* the theory's axioms are true otherwise.

[69]In my view (Heck, 2004), the T-sentences themselves are not trivialities, either, but this is a separate issue.

[70]Thanks to Volker Halbach and Jeff Ketland for conversations early in the history of this paper, and to Josh Schechter for conversations later on, that helped greatly. Comments on a draft of the paper from Cezary Cieśliński and Ali Enayat did much to improve it. A talk based upon the paper was given at a conference on philosophical logic, organized by Delia Graff Fara and held at Princeton University in April 2009. The paper

References

Boolos, G. (1993). *The Logic of Provability*. New York, Cambridge University Press. 9

Burgess, J. P. (2005). *Fixing Frege*. Princeton NJ, Princeton University Press. 7, 16, 17

Cieśliński, C. (2009). Truth, conservativeness, and provability. Forthcoming in *Mind*. 27

Craig, W. and Vaught, R. L. (1958). 'Finite axiomatizability using additional predicates', *Journal of Symbolic Logic* 23: 289–308. 36, 37, 38, 39

DeVidi, D. and Solomon, G. (1999). 'Tarski on 'essentially richer' metalanguages', *Journal of Philosophical Logic* 28: 1–28. 12

Feferman, S. (1960). 'Arithmetization of metamathematics in a general setting', *Fundamenta Mathematicae* 49: 35–92. 3, 5, 19

Field, H. (1999). 'Deflating the conservativeness requirement', *Journal of Philosophy* 96: 533–40. 49

Grzegorczyk, A. (2005). 'Undecidability without arithmetization', *Studia Logica* 79: 163–230. 35

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First-order Arithmetic*. New York, Springer-Verlag. 7, 15, 16, 19, 27, 30, 43, 44, 46

Halbach, V. (1999). 'Disquotationalism and infinite conjunction', *Mind* 108: 1–22. 47

was also discussed at a meeting of the New England Logic and Langauge Colloquium in May 2011, and it was presented at the Philosophy of Mathematics Seminar at Oxford University, also 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, Hartry Field, Volker Halbach, Daniel Isaacson, Charles Parsons, Agustín Rayo, and Lionel Shapiro. Special thanks to my commentator at Princeton, Josh Dever, whose comments did a lot to improve the presentation.

I owe the greatest debt, however, to Albert Visser. This paper simply could not have been written but for his generous assistance. Just as my ideas were starting to come together, Albert helped me to disentangle different threads in what I was trying to do. Once we'd managed that, he then made a series of observations based upon work he was doing at the time (Visser, 2009a,b) that transformed the direction of my project: These appear here as Theorem 5.3 and Corollary 6.4. And, finally, Albert has generously, and patiently, answered question after question after question as I've struggled to make sure all the details were right. So thanks, Albert! I've learned a ton.

Given Albert's extensive influence on this paper, it is perhaps worth my saying a word about where I take my own contribution to lie. A few of the key results are mine: Corollary 6.13 and Corollary 6.18, for example; so is the proof of Theorem 5.3, in particular, the way it resolves the problems posed by the logical axioms, which Albert and I independently discovered had been neglected in the existing literature. But I take my main contribution to lie in the general approach taken here: The idea that we should investigate truth-theories over weak base theories; the realization that, if we proceed in the usual way, then this investigation doesn't go as one might have hoped; the consequent suggestion that we ought to disentangle syntax from the object-theory; and the realization that this investigation goes a good deal better, indeed, almost as well as one could have hoped.

—— (2001a). 'Disquotational truth and analyticity', *Journal of Symbolic Logic* 66: 1959–1973. 34
—— (2001b). 'How innocent is deflationism?', *Synthese* 126: 167–94. 49
Heck, R. G. (2004). 'Truth and disquotation', *Synthese* 142: 317–52. 51
—— (2005). 'Reason and language', in C. MacDonald and G. MacDonald (eds.), *McDowell and His Critics*. Oxford, Blackwells, 22–45. 38
—— (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. 38
Ketland, J. (1999). 'Deflationism and Tarski's paradise', *Mind* 108: 69–94. 48
Kleene, S. (1952). 'Finite axiomatizability of theories in the predicate calculus using additional predicate symbols', *Memoirs of the American Mathematical Society* 10: 27–68. 14, 38
Kotlarski, H. (1986). 'Bounded induction and satisfaction classes', *Zeitschrift für Mathematische Logik* 32: 531–544. 30, 32
Mostowski, A. (1952). 'On models of axiomatic systems', *Fundamenta Mathematicae* 39: 133–58. 5
Nelson, E. (1986). *Predicative Arithmetic*. Mathematical Notes 32. Princeton NJ, Princeton University Press. 7
Parsons, C. (1981). 'Sets and classes', in *Mathematics in Philosophy*. Ithaca NY, Cornell University Press, 209–20. 49
Pudlák, P. (1985). 'Cuts, consistency statements and interpretations', *Journal of Symbolic Logic* 50: 423–41. 18
Quine, W. V. O. (1946). 'Concatenation as a basis for arithmetic', *Journal of Symbolic Logic* 11: 105–14. 35
Ray, G. (2005). 'On the matter of essential richness', *Journal of Philosophical Logic* 34: 433–57. 12, 13
Shapiro, S. (1998). 'Proof and truth: Through thick and thin', *Journal of Philosophy* 95: 493–521. 48, 49
Tarski, A. (1944). 'The semantic conception of truth and the foundations of semantics', *Philosophy and Phenomenological Research* 4: 341–75. 1, 12
—— (1958). 'The concept of truth in formalized languages', in J. Corcoran (ed.), *Logic, Semantics, and Metamathematics*. Indianapolis, Hackett, 152–278. 13, 14, 35, 43
Tarski, A., Mostowski, A., and Robinson, A. (1953). *Undecidable Theories*. Amsterdam, North-Holland Publishing. 4, 5, 11, 35
Visser, A. (1991). 'The formalization of interpretability', *Studia Logica* 50: 81–105. 7
—— (2006). 'Categories of theories and interpretations'. Wellesley MA, A. K. Peters, 77–136. 5
—— (2008). 'Pairs, sets and sequences in first-order theories', *Archive for Mathematical Logic* 47: 299–326. 9, 12

—— (2009a). 'Can we make the second incompleteness theorem coordinate free?', *Journal of Logic and Computation* 21: 543–60. 19, 40, 52

—— (2009b). 'The predicative Frege hierarchy', *Annals of Pure and Applied Logic* 160: 129–53. 25, 52

Wang, H. (1952). 'Truth definitions and consistency proofs', *Transactions of the American Mathematical Society* 73: 243–275. 9, 11, 12, 30

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

DEPARTMENT OF PHILOSOPHY, BROWN UNIVERSITY, PROVIDENCE RI 02912