From Tarski to Gödel—or how to derive the second incompleteness theorem from the undefinability of truth without self-reference

ALBERT VISSER, *Philosophy, Faculty of Humanities, Utrecht University, Janskerkhof 13, 3512BL Utrecht, The Netherlands.*
E-mail: a.visser@uu.nl

Abstract

In this paper, we provide a fairly general self-reference-free proof of the second incompleteness theorem from Tarski's theorem on the undefinability of truth.

Keywords: formal theories, consistency, self-reference, truth.

1 Prelude

Self-reference is a great and beautiful thing, but it may be interesting to see what one can do without it. In this paper, we provide a self-reference-free proof of the second incompleteness theorem from Tarski's theorem on the undefinability of truth. Thus, we do not aim to eliminate self-reference altogether. Rather, we reduce the total amount of self-reference that is required for the combined proofs of Gödel's and Tarski's theorems.

1.1 Motivation

Why look for alternative proofs of the second incompleteness theorem? As a matter of general methodology, it is always good to have as many proofs as possible of an important theorem. Such alternative proofs increase our insight. Moreover, they could lead to new generalizations of the theorem.

The usual proof of the second incompleteness theorem employs arithmetization, including the verification of the Löb conditions, in combination with a construction of self-reference. Since the second incompleteness theorem presupposes arithmetization in its very formulation, the obvious target of a different proof is to avoid the use of self-reference.

REMARK 1.1
Even if arithmetization cannot be eliminated from the proof of the second incompleteness theorem, there is a next best thing. There could be an arithmetization-free lemma from which the second incompleteness theorem can be derived in a straight-forward, self-reference-free way.

In fact, we have a good candidate for such a lemma for the case of consistent sequential theories axiomatized by a scheme. See [11]. Regrettably, the current proofs of the lemma are via the second

Vol. 29, No. 5, © The Author(s) 2019. Published by Oxford University Press.

Advance Access Publication on 8 March 2019 doi:10.1093/logcom/exz004

incompleteness theorem. Thus, currently, the desired lemma is only a corollary. The problem of finding a direct proof of the relevant result was first formulated by Jan Krajíček.

1.2 On self-reference

There is an interesting *philosophical* discussion about the question of what precisely is involved in self-reference. For example, perhaps speakers or utterances refer and not sentences. Moreover, if sentences do refer at all, then they should contain a term that does the referring. Thus, if a theory—like the usual version of Peano Arithmetic with zero, successor, plus and times—does not contain a term that represents the substitution function, then its Gödel sentence cannot be self-referential. Also, can a theory that officially only refers to numbers in some way refer to sentences? In addition, what about theories, like set theory, that do not have terms at all? Etcetera.

It is clear from our motivation that this kind of discussion is irrelevant for the purposes of the present paper. For us, the usual construction of the Gödel sentence, including the case of, e.g. set theory, is *paradigmatic* for 'self-reference' in the intended sense. We aim for proofs that have sufficient *mathematical* distance from that construction. This means that we also want to avoid proofs using, e.g. Yablo-style constructions or the Grelling paradox. In fact, we want to stay as far as possible from constructions that involve the recursion theorem, diagonalization and majorization.

Of course, it is impossible to draw a sharp boundary. For example, Henryk Kotlarski employs the argument that recursively saturated models are not prime to prove a form of Tarski's theorem on the undefinability of truth. See [6]. This involves consideration of a recursive type that, roughly, makes x not equal to any definable element. Is this diagonalization? I can see arguments for both *yes* and *no*. I do not think we have to settle such issues in the context of the present paper. It is sufficient that the reader recognizes that argument presented here is fairly far removed from the usual self-referential construction.

REMARK 1.2
We employ, in the present paper, the construction of maximally consistent sets at various places. Even here we find some room for doubt. Both Joost Joosten (in conversation) and referee 3 suggest that the stepwise construction of these sets can be seen as a form of diagonalization.

But what would this diagonalization be? I guess one would have to say that it diagonalizes out of the inconsistent (and *ipso facto* complete) sets to produce a consistent complete set. The inconsistent sets are in principle finite objects that can be enumerated. At every step, we make the constructed set different from the inconsistent set considered at that step, if there is one.

This way of fitting the construction of a maximally consistent set into the mold of diagonalization may seem somewhat far-fetched, but lacking an appropriate philosophical framework to evaluate such claims, it is hard to definitely exclude it.

1.3 About the proof

Our proof is roughly the result of combining a proof-plan due to Kreisel [8] with a proof-plan due to Adamowicz and Bigorajska [1].

I give two variants of the proof. Since the two variants share a lot of text, I will give the variants as different threads in the paper. I will represent the first variant as the main thread, and I will give the additions for the second thread `in typewriter font`.

In the first variant, we avoid the use of the third Löb condition, to wit that provability implies the provability of provability. This variant is reasonably robust w.r.t. syntactical details and w.r.t. the representation of the axiom set—as long as it is Σ^0_1.

REMARK 1.3
The robustness of the first variant is analogous to similar robustness of a proof by Mycielski of the second incompleteness theorem for his finitistic theory FIN in an unpublished manuscript 'Finitistic intuitions supporting the consistency of ZF and ZF+ AD'. Mycielski's ideas were adapted by Pavel Pudlák to prove the second incompleteness theorem for weak theories like Q. See [7]. However, where Mycielski and Pudlák replace the third Löb condition by a more flexible condition using two provability predicates, we avoid the third Löb condition altogether.

The second variant employs techniques of refined syntactical analysis that go back to the work of Smullyan [9] and Buss [2].

We think it is interesting to present both approaches. In the first place, one simply wants to see how the proofs play out using these different techniques. In the second place, there is the question whether one of the approaches has limitations that the other doesn't have. It turns out, for the present purposes, that there is no significant difference. The second variant, admittedly, delivers a proof with a somewhat more restricted scope, but this can be repaired. In the third place, the variants could lead to different generalizations.

REMARK 1.4
The ideal would be to study a new proof in a setting that is so general that we do not have the resources to even prove things like the Gödel fixed-point lemma. We do not aspire to this ideal here. The theories studied all have the resources to do diagonalization.

1.4 Prerequisites

The full proof presupposes some knowledge of relevant materials from the text books [4, 5]. For the benefit of the reader who is not acquainted with weak theories we sketch the proof for the case of Peano arithmetic in Section 2. This special case has the advantage that almost all technical complications disappear, while the essence of the proof idea is still visible.

We will employ, for the full proof, the notations, conventions and elementary facts of [13], to which the present paper is a sequel. We will use the interpretation existence lemma. For a careful exposition of this last result, see [14]. Finally, we will make use of a result from [12].

2 The second incompleteness theorem for Peano arithmetic

We prove second incompleteness theorem for Peano arithmetic PA. Let the standard axiomatization of PA be π. We write $\Box_\pi A$ for the arithmetization of the provability of A from π and $\Diamond_\pi A := \neg\Box_\pi\neg A$ for the consistency of (the theory axiomatized by) π extended with A. Here is the second incompleteness theorem for PA:

THEOREM 2.1
$\mathsf{PA} \nvdash \Diamond_\pi \top$.

We will use the following version of Tarski's theorem on the undefinability of truth for the case of PA.

THEOREM 2.2
For any arithmetical formula $A(x)$ (with only x free) $\mathsf{PA}+\{(A(\ulcorner B \urcorner) \leftrightarrow B) \mid B$ is an arithmetical sentence$\}$ is inconsistent.

REMARK AND QUESTION 2.3
Referee 2 prefers to call the following statement 'Tarski's theorem on the undefinability of truth for PA'.

† for all $A(x)$ with only x free, $\mathsf{PA} \nvdash \{(B \leftrightarrow A(\ulcorner B \urcorner)) \mid B$ is an arithmetical sentence$\}$.

This would make my version 'Tarski's theorem on the undefinability of truth for all consistent extensions of PA'.

It is an open question whether we can derive the second incompleteness theorem for PA from (†) in a self-reference free way. However, inspecting the proof in the present paper, one can derive the second incompleteness theorem for PA from the following version, which preserves some of the spirit of (†):

‡ for all $A(x)$ with only x free,

$$\mathsf{PA} \not\rhd (\mathsf{PA} + \{(B \leftrightarrow A(\ulcorner B \urcorner)) \mid B \text{ is an arithmetical sentence}\}).$$

We sketch a proof of the second incompleteness theorem for PA from (‡) in Appendix A. (The appendix employs the notations of Section 2.)

PROOF OF THEOREM 2.1 FROM THEOREM 2.2. We assume Theorem 2.2. Suppose, in order to arrive at a contradiction, that $\mathsf{PA} \vdash \Diamond_\pi \top$.

Let $\mathcal{S}$ be a maximal set of Σ^0_1-sentences such that $\mathsf{PA}+\mathcal{S}$ is consistent. Let $\mathcal{M}$ be a model of $\mathsf{PA}+\mathcal{S}$.

We apply the Henkin construction as described in e.g. [3] (see also [14]), to construct an $\mathcal{M}$-internal Henkin model $\mathsf{H}(\mathcal{M})$ of PA based on $\Diamond_\pi \top$. We claim that $\mathcal{M}$ and $\mathsf{H}(\mathcal{M})$ satisfy the same Σ^0_1-sentences, to wit the Σ^0_1-sentences of $\mathcal{S}$.

Consider any Σ^0_1-sentence S and suppose $\mathcal{M} \models S$. It is easily seen that $\mathsf{H}(\mathcal{M})$ is an end-extension of $\mathcal{M}$ and, hence, $\mathsf{H}(\mathcal{M}) \models S$. [By inspection of the Henkin construction, we find that $\mathsf{PA} \vdash \Box_\pi B \to B^{\mathsf{H}}$, for all arithmetical sentences B. Here B^{H} is the translation of B via the translation function associated with the Henkin interpretation H. Since, by PA-verifiable Σ^0_1-completeness, we have $\mathsf{PA} \vdash S \to \Box_\pi S$, we find $\mathsf{PA} \vdash S \to S^{\mathsf{H}}$. Thus, it follows that $\mathsf{H}(\mathcal{M}) \models S$.]

We may conclude that $\mathsf{H}(\mathcal{M}) \models \mathcal{S}$. On the other hand, by the maximality of $\mathcal{S}$, the models $\mathcal{M}$ and $\mathsf{H}(\mathcal{M})$ cannot satisfy more Σ^0_1-sentences than those in $\mathcal{S}$.

Since $\mathsf{H}(\mathcal{M})$ again will contain $\Diamond_\pi \top$, we can repeat the construction H to obtain $\mathsf{HH}(\mathcal{M})$.

We claim that $\mathsf{H}(\mathcal{M})$ and $\mathsf{HH}(\mathcal{M})$ are elementary equivalent. The reason is as follows. The Henkin construction is based on yes–no decisions that depend on the truth or falsity of Σ^0_1-questions. On the standard level it, thus, depends on the truth or falsity of Σ^0_1-sentences. Since $\mathcal{M}$ and $\mathsf{H}(\mathcal{M})$ satisfy the same Σ^0_1-sentences we find that, on standard levels, the same choices are made in the Henkin construction, and hence that $\mathsf{H}(\mathcal{M})$ and $\mathsf{HH}(\mathcal{M})$ are elementary equivalent.

The Henkin construction provides an internal truth-predicate $\mathfrak{H}$ such that we have $\mathsf{PA} \vdash \mathfrak{H}(\ulcorner B \urcorner) \leftrightarrow B^{\mathsf{H}}$. So we find:

$$\begin{aligned} \mathsf{H}(\mathcal{M}) \models B \quad &\Leftrightarrow \quad \mathsf{HH}(\mathcal{M}) \models B \\ &\Leftrightarrow \quad \mathsf{H}(\mathcal{M}) \models B^{\mathsf{H}} \\ &\Leftrightarrow \quad \mathsf{H}(\mathcal{M}) \models \mathfrak{H}(\ulcorner B \urcorner). \end{aligned}$$

Thus, $\mathsf{H}(\mathcal{M})$ is a model of $\mathsf{PA} + \{\mathfrak{H}(\ulcorner B \urcorner) \leftrightarrow B \mid B$ is an arithmetical sentence$\}$. But this contradicts Theorem 2.2 on the undefinability of truth. □

REMARK 2.4
In case PA proves its own consistency, our construction effectively yields a Δ^0_2-truth-predicate $\mathfrak{H}$ such that $\mathsf{PA} + \{\mathfrak{H}(\ulcorner B \urcorner) \leftrightarrow B \mid B$ is an arithmetical sentence$\}$ is consistent. We note that the intensional details of the precise choice of the axiomatization of PA are essentially used in the construction of $\mathfrak{H}$ and the verification of its relevant properties.

REMARK 2.5
Our proof works if based on the weaker assumption that $\mathsf{PA} + \mathcal{S} \vdash \Diamond_\pi \top$. It therefore follows that $\Box_\pi \bot$ is in any maximal set of Σ^0_1-sentences $\mathcal{S}$ such that $\mathsf{PA} + \mathcal{S}$ is consistent.

3 Statement of two theorems

In this section, we state both second incompleteness theorem and Tarski's theorem on the undefinability of truth in the general forms we consider in the present paper.

3.1 Our version of the second incompleteness theorem

We work with S^1_2 as our basic basic weak arithmetic. See [2] or [4]. We use $\Box_\sigma$ for the arithmetization of provability from axiom set σ and $\Diamond_\sigma$ for $\neg\Box_\sigma\neg$.

We will prove the following version of the second incompleteness theorem.

THEOREM 3.1
Suppose U is a recursively enumerable theory. Let σ be a Σ^0_1-formula [Σ^{b}_1-`formula`] that defines an axiom set of U in the standard model. Suppose that $N : U \rhd (\mathsf{S}^1_2 + \Diamond_\sigma \top)$. Then U is inconsistent.

The basic idea for a version of the second incompleteness theorem that uses interpretability is due to Feferman [3]. Of course, Feferman, around 1960, was thinking of PA as base theory rather than S^1_2.
The variant where σ is Σ^0_1 is stronger than the one where σ is Σ^{b}_1. However, there are easy arguments to reduce the second incompleteness theorem for Σ^0_1-axiomatizations to the second incompleteness theorem for Σ^{b}_1-axiomatizations. See [13].

3.2 Our version of Tarski's theorem

Let Θ be a signature. Suppose $N : U \rhd \mathsf{R}$, where R is the very weak arithmetic introduced in [10]. Let A be a Θ-formula with only x free. We take $\mathsf{TB}^{N,A}_{\Theta}$ to be the set of all Tarski biconditionals $A(\ulcorner B \urcorner) \leftrightarrow B$, for B a Θ-sentence and $\ulcorner B \urcorner$ an N-numeral.

REMARK 3.2
We note that we generally need an interpretation like N to have numerals at all. The signature Θ could, after all, contain only a binary predicate $\in$. In such a case there are no self-explanatory numerals. Also, in the minimal context we are considering, we could have $N : U \rhd \mathsf{R}$ and $N' : U \rhd \mathsf{R}$, such that $\mathsf{TB}_{\Theta}^{N,A}$ and $\mathsf{TB}_{\Theta}^{N',A}$ are not provably equivalent over the given theory U.[1]

Here is our version of Tarski's theorem:

THEOREM 3.3
Let U be a theory of signature Θ. Suppose that $N : U \rhd \mathsf{R}$ and that, for some U-formula A with only x free, we have $U \vdash \mathsf{TB}_{\Theta}^{N,A}$. Then, U is inconsistent.

We can reduce our version of Tarski's theorem to the following special case. This reduction uses the recursion theorem, so since we want to exclude anything that even smells of self-reference we, probably, have to exclude this reduction from the main thread of the argument. This is a pity since it would have been nice to reduce all instances of the second incompleteness theorem to one single application of self-reference.

THEOREM 3.4
Let $\mathcal{A}$ be the signature of arithmetic extended with a fresh unary predicate T. Then, $\mathsf{R} + \mathsf{TB}_{\mathcal{A}}^{\mathsf{ID},\mathsf{T}}$ is inconsistent.

Here is our argument for the reduction of Theorem 3.3 to Theorem 3.4. Suppose $N : U \rhd \mathsf{R}$ and $U \vdash \mathsf{TB}_{\Theta}^{N,A}$. Let ν be the translation associated with N. We interpret $\mathsf{R} + \mathsf{TB}_{\mathcal{A}}^{\mathsf{ID},\mathsf{T}}$ in U via the translation $\nu^{\star}$ which is ν on the arithmetical vocabulary and which is $A(\mathsf{tr}_{\nu^{\star}}(x))$ on T. Here, $\mathsf{tr}_{\nu^{\star}}$ is the arithmetization of the function $B \mapsto B^{\nu^{\star}}$. We evidently need the recursion theorem to make our definition work.

4 The Henkin construction

The main tool of our proof will be the interpretation existence lemma. In this section we state this lemma and collect the relevant facts around it.

Let X be a primitive unary predicate symbol and let $\eta[X]$ be the Henkin translation based on $\mathsf{S}^1_2 + \Diamond_X \top$. Consider an arithmetical formula A with only a designated variable x free. We write $\eta[A]$ for the result of replacing X by A in the formulas that make up $\eta[X]$.

THEOREM 4.1
Suppose σ is a Σ^0_1-formula that represents the axioms of U in the standard model. We have $\mathsf{H}[\sigma] : (\mathsf{S}^1_2 + \Diamond_\sigma \top) \rhd U$. Here $\mathsf{H}[\sigma]$ is the Henkin interpretation based on $\eta[\sigma]$.

We have, inside $\mathsf{S}^1_2 + \Diamond_\sigma \top$, a truth predicate $\mathfrak{H}[\sigma]$ for $\mathsf{H}[\sigma]$ that satisfies the commutation conditions for $\mathsf{H}[\sigma]$ on a definable cut $\mathsf{J}[\sigma]$. It follows that $\mathsf{S}^1_2 + \Diamond_\sigma \top \vdash \mathfrak{H}[\sigma](\ulcorner A \urcorner) \leftrightarrow A^{\mathsf{H}[\sigma]}$.

For a proof of this result see [14] which also discusses the history of the result.

Let $\Sigma^0_{1,1}$ be the class of formulas of the form $\exists x\, \forall y \leq t(x)\, \exists z\, S_0(x,y,z,\vec{u})$, where S_0 is Δ_0. Inspection shows that $\mathsf{prov}_\sigma(x)$, where σ is Σ^0_1, can be written as a $\Sigma^0_{1,1}$-formula. [Inspection shows that $\mathsf{prov}_\sigma(x)$, where σ is Σ^{b}_1, can be written as a $\exists\Sigma^{\mathsf{b}}_1$-formula. This employs

[1] If U is sequential, then $\mathsf{TB}_{\Theta}^{N,A}$ and $\mathsf{TB}_{\Theta}^{N',A}$ are provably equivalent over U.

the Σ_1^{b}-collection (or: Σ_1^{b}-replacement) principle. See [2, Theorem 14, p. 53].] We have the following insight.

THEOREM 4.2
Suppose σ is a Σ_1^0-formula [Σ_1^{b}-formula] that represents the axioms of U in the standard model. Suppose $\mathcal{N}_0$ and $\mathcal{N}_1$ are models of $\mathsf{S}_2^1 + \Diamond_\sigma \top$. Suppose further that $\mathcal{N}_0$ and $\mathcal{N}_1$ are $\Sigma_{1,1}^0$-elementary equivalent [$\exists\Sigma_1^{\mathsf{b}}$-elementary equivalent]. Then $\mathsf{H}[\sigma](\mathcal{N}_0)$ and $\mathsf{H}[\sigma](\mathcal{N}_1)$ are elementary equivalent.

PROOF. Inspecting the Henkin construction, we see that it fully depends on $\Sigma_{1,1}^0$-decisions [$\exists\Sigma_1^{\mathsf{b}}$-decisions]. For standard formulas only the standard $\Sigma_{1,1}^0$-sentences [$\exists\Sigma_1^{\mathsf{b}}$-sentences] true or false in our models are relevant. □

Consider any theory U of signature Θ. Suppose the axiom set of U is represented by σ in Σ_1^0 [in $\exists\Sigma_1^{\mathsf{b}}$]. Let $N : U \rhd (\mathsf{S}_2^1 + \Diamond_\sigma \top)$. We write H for $\mathsf{H}[\sigma]$ and $\mathfrak{H}$ for $\mathfrak{H}[\sigma]$. We write $\mathfrak{H}^N$ for the N-translation of $\mathfrak{H}$.

THEOREM 4.3
Consider $\mathcal{M} \models U$ and suppose $N(\mathcal{M})$ and $N\mathsf{H}N(\mathcal{M})$ are $\Sigma_{1,1}^0$-elementary equivalent [$\exists\Sigma_1^{\mathsf{b}}$-elementary equivalent]. Then $\mathsf{H}N(\mathcal{M}) \models \mathsf{TB}_\Theta^{N,\mathfrak{H}^N}$.

PROOF. From the assumptions of the theorem, we find, by Theorem 4.2, that $\mathsf{H}N(\mathcal{M})$ and $\mathsf{H}N\mathsf{H}N(\mathcal{M})$ are elementary equivalent. Hence,

$$\begin{aligned} \mathsf{H}N(\mathcal{M}) \models A &\Leftrightarrow \mathsf{H}N\mathsf{H}N(\mathcal{M}) \models A \\ &\Leftrightarrow \mathsf{H}N(\mathcal{M}) \models A^{\mathsf{H}N} \\ &\Leftrightarrow \mathsf{H}N(\mathcal{M}) \models \mathfrak{H}^N(\ulcorner A \urcorner). \end{aligned}$$

□

[Here is one final fact about the Henkin construction that we will use for the second thread in the main proof in Section 5.

THEOREM 4.4
Suppose σ is a Σ_1^{b}-formula that represents the axioms of U in the standard model. Then we have $\mathsf{S}_2^1 + \Diamond_\sigma \top \vdash \Box_\sigma A \to \mathfrak{H}[\sigma](A)$.[2]

The proof is by inspection of the Henkin construction.]

5 Proof of the main theorem

Suppose σ is a Σ_1^0-formula that represents the axioms of U in the standard model. Suppose further that $N : U \rhd (\mathsf{S}_2^1 + \Diamond_\sigma \top)$.

Theorem 4.3 tells us that, in order to prove Theorem 3.1, it is sufficient to provide a model $\mathcal{M}$ of U so that $N\mathsf{H}[\sigma]N(\mathcal{M})$ is $\Sigma_{1,1}^0$-elementary equivalent [$\exists\Sigma_1^{\mathsf{b}}$-elementary equivalent] with $N(\mathcal{M})$.

In the first thread, we replace the originally given N by a new $K : U \rhd (\mathsf{S}_2^1 + \Diamond_\sigma \top)$. We prove the desired elementary equivalence for K in the role of N. We use the following fact.

[2] A much stronger result is true. As a consequence, the result also holds for, e.g. Σ_1^0-formulas. See [14].

THEOREM 5.1
We have

$$\mathsf{S}^1_2 \rhd_{\mathsf{loc,cut}} \mathsf{W} := \mathsf{S}^1_2 + \{S \to S^I \mid S \in \Sigma^0_{1,1} \text{ and } I \text{ is a definable cut}\}.$$

Here $\rhd_{\mathsf{loc,cut}}$ is local interpretability on definable cuts. For a treatment of this insight, see [12].[3]

Since $\Pi^0_{1,1}$-formulas are downwards preserved to cuts, we have $(\mathsf{S}^1_2 + \Diamond_\sigma \top) \rhd_{\mathsf{loc,cut}} (\mathsf{W} + \Diamond_\sigma \top)$. It follows that $U \rhd_{\mathsf{loc}} (\mathsf{W} + \Diamond_\sigma \top)$. Since U proves its own consistency with respect to N, it follows that U proves every restricted consistency statement $\Diamond^N_{\mathsf{W}+\Diamond_\sigma\top,n}\top$. Here the restriction n means that we consider the sub-theory of $\mathsf{W} + \Diamond_\sigma \top$ given by the axioms of $\mathsf{W} + \Diamond_\sigma \top$ that are $\leq n$. The Orey–Hájek characterization, tells us that we find a $K : U \rhd (\mathsf{W} + \Diamond_\sigma \top)$. We note that the Orey–Hájek characterization is itself based on the interpretation existence lemma. See [11] or [14].

We now take $\mathcal{S}$ to be a maximal set of $\Sigma^0_{1,1}$-sentences such that $U + \mathcal{S}^K$ is consistent. Let $\mathcal{M}$ be a model of $U + \mathcal{S}^K$. Let $\mathcal{J}$ be $K(\mathcal{M})$-definable $K(\mathcal{M})$-cut (closed under ω_1) such that there is a $K(\mathcal{M})$-definable isomorphism of $\mathcal{J}$ with a cut $\mathcal{J}'$ of $K\mathsf{H}K(\mathcal{M})$, where $\mathsf{H} := \mathsf{H}[\sigma]$. We know that such a cut exists by the results of [7], noting that $K(\mathcal{M})$ is a model of S^1_2 and, hence, sequential. We have

$$\begin{aligned} S \in \mathcal{S} \quad &\Rightarrow \quad K(\mathcal{M}) \models S && (1)\\ &\Rightarrow \quad \mathcal{J} \models S && (2)\\ &\Rightarrow \quad K\mathsf{H}K(\mathcal{M}) \models S. && (3) \end{aligned}$$

Step (2) holds since $K(\mathcal{M}) \models W$. We have Step (3), since the truth of $\Sigma^0_{1,1}$-sentences is preserved by the isomorphism from $\mathcal{J}$ to $\mathcal{J}'$ and then is upwards preserved from the cut $\mathcal{J}'$ to $K\mathsf{H}K(\mathcal{M})$.

By the maximality of $\mathcal{S}$, it follows that $S \in \mathcal{S}$ iff $K(\mathcal{M}) \models S$ and $S \in \mathcal{S}$ iff $K\mathsf{H}K(\mathcal{M}) \models S$. We may conclude that $K(\mathcal{M})$ and $K\mathsf{H}K(\mathcal{M})$ are $\Sigma^0_{1,1}$-elementary equivalent. Thus, we have our desired result with K in the role of N.

[We treat the second thread. Suppose σ is a Σ^{b}_1-formula that represents the axioms of U in the standard model. Suppose $N : U \rhd (\mathsf{S}^1_2 + \Diamond_\sigma \top)$. Let $\mathcal{S}$ be a maximal set of $\exists\Sigma^{\mathsf{b}}_1$-sentences such that $U + \mathcal{S}^N$ is consistent. Let $\mathsf{H} := \mathsf{H}[\sigma]$. We have, by verifiable $\exists\Sigma^{\mathsf{b}}_1$-completeness in S^1_2, for $S \in \mathcal{S}$ in combination with Theorem 4.4, that

$$\begin{aligned} U + \mathcal{S}^N \quad &\vdash \quad \Box^N_\sigma S^N\\ &\vdash \quad S^{N\mathsf{H}N}. \end{aligned}$$

Now let $\mathcal{M}$ be a model of $U + \mathcal{S}^N$. It follows that $\mathsf{H}N(\mathcal{M}) \models \mathcal{S}^N$. By maximality, it follows that, for any $S \in \exists\Sigma^{\mathsf{b}}_1$, we have

$$N(\mathcal{M}) \models S \Leftrightarrow S \in \mathcal{S} \quad \text{and} \quad N\mathsf{H}N(\mathcal{M}) \models S \Leftrightarrow S \in \mathcal{S}.$$

So we are done.]

[3] The theory W is interpretable in an ω_1-cut of the theory Peano Basso of [12].

We note that the second thread looks somewhat more efficient. However, the first thread avoids the more refined syntactic analysis that is the basis of the second thread.

OPEN QUESTION 5.2
Our argument is presented as a model-theoretic argument. So, it is itself not obviously formalizable in a weak theory. However, it seems to me that the models can be eliminated from the argument. They mainly function as a heuristic tool. So, the question is how much resources do we need to internalize our argument in a theory. Is S^1_2 sufficient?

Acknowledgements

I thank Fedor Pakhomov, for his comments on an earlier version of the paper and for his encouragement; Joost Joosten, for his critical remarks on the paper; and the three referees whose comments helped me to improve the paper.

References

[1] Z. Adamowicz and T. Bigorajska. Existentially closed structures and Gödel's second incompleteness theorem. *The Journal of Symbolic Logic*, **66**, 349–356, 2001.

[2] S. R. Buss. *Bounded Arithmetic*. Bibliopolis, Napoli, 1986.

[3] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, **49**, 35–92, 1960.

[4] P. Hájek. and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1993.

[5] R. Kaye. Models of Peano Arithmetic. Oxford Logic Guides. Oxford University Press, 1991.

[6] H. Kotlarski. The incompleteness theorems after 70 years. *Annals of Pure and Applied Logic*, **126**, 125–138, 2004.

[7] P. Pudlák. Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, **50**, 423–441, 1985.

[8] C. Smoryński. The incompleteness theorems. In *Handbook of Mathematical Logic*, J. Barwise ed., pp. 821–865. North-Holland, Amsterdam, 1977.

[9] .R. M. Smullyan. *Theory of Formal Systems*, Annals of Mathematics Studies, vol. 47. Princeton University Press, Princeton, New Jersey, 1961.

[10] A. Tarski, A. Mostowski and R. M. Robinson. *Undecidable Theories*. North-Holland, Amsterdam, 1953.

[11] A. Visser. Can we make the Second Incompleteness Theorem coordinate free. *Journal of Logic and Computation*, **21**, 543–560, 2011. First published online August 12, 2009, doi: 10.1093/logcom/exp048.

[12] A. Visser. Peano Corto and Peano Basso: a study of local induction in the context of weak theories. *Mathematical Logic Quarterly*, **60**, 92–117, 2014.

[13] A. Visser. Another look at the second incompleteness theorem. Logic Group Preprint Series 339, Faculty of Humanities, Philosophy, Utrecht University, Janskerkhof 13, 3512 BL Utrecht, https://lgps.sites.uu.nl, 2017.

[14] A. Visser. The interpretation existence lemma. In *Feferman on Foundations, Outstanding Contributions to Logic 13, pp. 101–144. Springer*, 2017.

A Interpretability of Tarski's biconditionals over PA

Let $\mathbb{S}$ be the set of A such that $\mathsf{PA}+\mathcal{S}\vdash A$, for all maximal sets of Σ^0_1-sentences $\mathcal{S}$ such that $\mathsf{PA}+\mathcal{S}$ is consistent.[4] Let

$$\mathsf{TB}^{\mathfrak{H}} := \{\mathfrak{H}(\ulcorner B\urcorner)\leftrightarrow B \mid \text{B is an arithmetical sentence}\},$$

where $\mathfrak{H}$ is the truth predicate associated with the Henkin interpretation H based on $\Diamond_\pi\top$.

THEOREM A.1
Suppose $\mathsf{PA}\vdash\Diamond_\pi\top$. Then, $\mathsf{PA}\rhd(\mathsf{PA}+\mathsf{TB}^{\mathfrak{H}})$.

PROOF. Suppose $\mathsf{PA}\vdash\Diamond_\pi\top$. Inspection of the proof in Section 2 shows that we have: (†) $\mathsf{H}:\mathbb{S}\rhd(\mathsf{PA}+\mathsf{TB}^{\mathfrak{H}})$.

We first show that (‡) $\mathsf{PA}\rhd_{\mathsf{loc}}\mathbb{S}$. Suppose $\mathbb{S}\vdash A$. We write $\Box$ for provability in predicate logic in the signature of arithmetic. Suppose $\Box\neg A$ is consistent with PA. Then, for some maximally consistent set $\mathcal{S}$ of Σ^0_1-sentences over PA, we have $\Box\neg A\in\mathcal{S}$. It follows, by the essential reflexivity of PA, that $\mathsf{PA}+\mathcal{S}\vdash\neg A$. This contradicts the assumption that $\mathbb{S}\vdash A$ and, hence, that $\mathsf{PA}+\mathcal{S}\vdash A$, in view of fact that $\mathsf{PA}+\mathcal{S}$ is consistent. Thus, it follows that $\mathsf{PA}\vdash\Diamond A$. We may conclude, by the interpretation existence lemma, that $\mathsf{PA}\rhd A$.

Combining (†) and (‡), we find that $\mathsf{PA}\rhd_{\mathsf{loc}}(\mathsf{PA}+\mathsf{TB}^{\mathfrak{H}})$. Since, PA is essentially reflexive and $\mathsf{PA}+\mathsf{TB}^{\mathfrak{H}}$ is recursively axiomatized, we find $\mathsf{PA}\rhd(\mathsf{PA}+\mathsf{TB}^{\mathfrak{H}})$.[5] □

REMARK A.2
I think there is some hope that we can prove a result analogous to Theorem A.1 for arbitrary recursively enumerable theories U. However, this would ask for some further adaptation of the argument.

OPEN QUESTION A.3
Can one prove Theorem A.1 directly (in a self-reference-free way), without the detour over maximally consistent sets of Σ^0_1-sentences?

Received 4 February 2019

[4] This is the set of formulas that I call $\mathsf{semco}(\mathsf{PA})$ in an as yet unpublished paper.

[5] The set $\mathbb{S}$ is not recursively enumerable. Elsewhere, we will show that it is complete Π^1_1.