Lecture 16

Joost J. Joosten

November 25, 2003

1 Gödel 2

We had concluded that PA in a certain sense can not see its own consistency (Gödel's second incompleteness theorem). What do we mean by "in a certain sense"? Well, of course, going all the way to the predicate $\mathsf{Bew}(x)$, there were many choices that we have made. How to represent syntax and syntactical operations? Which axiomatization of PA do we take, and how do we represent this? Although these were all choices to be made, everybody aggreed that they are natural choices.

Of course we could also make non-natural choices. We could say that the sentence $\mathtt{S}0 = \mathtt{S}0$ is a representation of the consistency of PA inside PA. We just fix this representation like this just in the same way that we fixed the representation of the symbol $\perp$ to be $\mathtt{S}0$. Under this representation we see that PA does prove its own consistency.

Of course one can (and should) object here by saying that this representation is not a natural one in the sense that, first of all we do not represent a sufficient part of SYNTAX to be able to compare that part of SYNTAX and its representation. Second, The structure of the consistency statement is by no means[1] reflected in the statement $\mathtt{S}0 = \mathtt{S}0$.

There are, as we shall see, more natural representations for which the second incompleteness theorem fails to hold. Therefor we consider Feferman's variant of Peano Arithmetic. The predicate $\mathsf{Bew}_{\mathrm{PAF}}(x)$ is constructed in precisely the same way as was $\mathsf{Bew}_{\mathrm{PA}}(x)$. The only difference is in the arithmetical representation of the axiom set.

By $\mathrm{PA}{\upharpoonright}y$ we denote the set containing all the axioms of PA with Gödel number at most y. In symbols: $x \in \mathrm{PA}{\upharpoonright}y \Leftrightarrow \mathsf{Ax}_{\mathrm{PA}}(x) \wedge x{\leq}y$. It is now also clear what we mean by $\mathsf{Con}(\mathrm{PA}{\upharpoonright}y)$.

[1] Someone could object here, "But the consistency of PA is true, whence any true statement faithfully represents it". But it is precisely this jump from truth to representation in a formal system that we are allergic to.

Exercise 1.1 *Give this definition. More general, describe how one would get for some theory U the statement $\mathsf{Con}(U)$. Should U necessarily be an arithmetical theory?*

Definition 1.2 $\mathsf{Ax}_{\mathrm{PAF}}(x) := \mathsf{Ax}_{\mathrm{PA}}(x) \wedge \mathsf{Con}(\mathrm{PA}{\restriction}x)$.

It is clear that extensionally PAF and PA are one and the same theories, as PA and certainly every finite part of it is actually consistent. Intensionally we are dealing with a completely different theory namely one that has a built in consistency.

Exercise 1.3 *Suppose that* $\mathrm{PAF} \vdash \mathsf{Pf}_{\mathrm{PAF}}(p, \ulcorner \bot \urcorner)$. *Reason to the effect that* $\mathrm{PA} \vdash \mathsf{Pf}_{\mathrm{PA}}(p, \ulcorner \bot \urcorner)$.

Show that

- $\mathrm{PA} \vdash \Box_{\mathrm{PAF}}\varphi \to \Box_{\mathrm{PA}}\varphi$,[2]
- $\mathrm{PA} \vdash \Box_{\mathrm{PA}}\varphi \to \Box_{\mathrm{PAF}}\varphi \vee \Box_{\mathrm{PA}}\bot$,
- $\mathrm{PA} \vdash \varphi \Leftrightarrow \mathrm{PAF} \vdash \varphi$.
- *It turns out that* $\Box_{\mathrm{PAF}}\varphi \to \Box_{\mathrm{PA}}\Box_{\mathrm{PAF}}\varphi$ *is not for all* φ *provable. Is this not violating provable* Σ_1*-completeness?*

Exercise 1.4 *Show that* $\mathrm{PA} \vdash \mathsf{Con}(\mathrm{PAF})$.

Shavrukov has written a nice paper that fully describes the behavior of these two provability predicates when acting together in one system.[Sha94]

We have just seen that with some effort we can surpass the second incompleteness theorem. The first however remains valid although other fixed points have to be considered. In the case of PAF, one could object that the way of representing the axioms of PAF (the same as those of PA) is not natural. If we do fix a natural representation of our axioms[3] there is no escape from the second incompleteness theorem.

Now, one still can object, that there were a lot of choices that were made in the course of fixing our coding techniques. Might it be so that if we would have chosen another Gödel numbering or another pairing function or another representation of finite sequences, that then in such a case, PA does prove its own consistency? To answer this question would include a description of a possible coding technique. And really some demands should be made upon the coding. For again there are simple[4] but somehow pathological codings

[2] We have returend to sloppy notation here. What would be the proper notation?

[3] One could for example demand syntactical constraints like the formula representing it being Σ_1^b or just Σ_1 for that matter.

[4] Primitive recursive!

known with really strange behavior.

In our program we now come quite naturally to the notion of interpretability. For as we shall see, somehow a natural coding can be seen as an interpretation. Interpretations do have an independent interest as is described in the other handout.

2 Interpretations and their arithmetizations.

In the other handout we have defined the notion of interpretability. Somehow this needs to be made a bit more precise. We will see that it is quite understandable that nowhere in the literature anyone has ever fully worked out the arithmetization process. This is just because it is a complete hassle but obviously executable.

We will also not give a fully detailed treatment. However, we find it instructive to reflect on some features. Again, what is our goal: to give an arithmetization of $U \rhd V$, the fact that there is a relative interpretation $\langle \delta, F \rangle$ such that F and δ are indeed of the required form and that U proves any translated axiom of V.

Thus it seems reasonable to construct formulas:

- $\mathsf{Domainspecifier}_U(x)$ that is to hold on (codes of) formulas that have one free variable and is provably (in U) non-empty.[5]

- $\mathsf{Translation}(x)$ that is to hold on (codes of) formulas that have one free variable and meet all the conditions that a translation F imposes.

- $\mathsf{Interpretation}(x)$ that expresses that x is a pair comprising a domain specifier and a translation.

- A Σ definiable function $\mathsf{Termtrans}(t, F, \varphi)$ that translates the term t with n free variables to a formula φ with the same n variables plus an additional one, representing the "read-off variable". In this formula φ, every occurence of a symbol of our language is replaced in the right way with the corresponding formula that is provided by our translation F. We will often refer to the value of the corresponding Σ-pterm as t^F or even t^j for some interpretation j.

- A Σ definable function $\mathsf{trans}(\varphi, F, \delta, \varphi^{\mathsf{tr}})$ that translates the formula φ with n free variables to a formula φ^{tr} with precisely the same free n

[5] From now on we will omit the subscript U and the statement "(codes of)" and other trivialities of this sort.

variables free such that every occurrence of a symbol of our language is replaced in the right way with the corresponding formula that is provided by our translation F and moreover, every quantifier is relativized using δ. We will often refer to the value of the corresponding Σ-pterm as φ^j.

If you think of how such arithmetical statements can best be made, immediately one observes a problem of "variable clashes" where in the translation variables that were not bound become bound in the translation or the other way round. These are technical matters concerning α conversion. Sometimes one moves to a fresh set of variables that are only used during the translation process.

Exercise 2.1 *Which of the formulas above is likely to be the most involved one? Explain your choice.*

In providing all these arithmetizations we should first realize how our theories U and V are actually represented. In the general case, this is rather more complicated than the case we are in. We commit ourselves to only consider extensions of PA of the form PA + α. This leads us to the following (not hard to write down in full detail (arithmetized!)) definition of an interpretation.

Definition 2.2 *An interpretation is a pair $\langle \delta, F \rangle$ where δ is (the code of, Ooooh no, we promised not to say these sort of things any more) a formula with one variable that is actually a domain specifier. F is a quadruple of pairs, consecutively written as*
$\langle \ulcorner 0 \urcorner, \mathsf{Zero} \rangle$
$\langle \ulcorner \mathtt{S} \urcorner, \mathsf{Successor} \rangle$
$\langle \ulcorner + \urcorner, \mathsf{Addition} \rangle$
$\langle \ulcorner \cdot \urcorner, \mathsf{Multiplication} \rangle$.
Here Zero, Successor, Addition *and* Multiplication *are all formulas with the required properties. For example for* Zero *we require that it be a formula with precisely one variable which provably holds for precisely one element.*[6]

Now that we have become more concrete it is easier to think of how the necessary formulas are composed.

[6]We have chosen to translate = to =. This is not a real restriction. If we would not have done so, it seems strange to demand that there is just one element satisfying Zero. They will only have to lie in the same equivalence class.

Definition 2.3 *Supose we had all the above formulas at our disposal. We then define* $\mathsf{Int}(\alpha, \beta)$*, we will more often write* $\alpha \rhd \beta$*, to be the formula*[7]

$$\exists j(\mathsf{interpretation}(j) \wedge \forall x\ (\mathsf{Ax}_{\mathrm{PA}+\beta}(x) \to \exists p\ (\mathsf{Pf}_{\mathrm{PA}+\alpha}(p, x^j))))$$

Note that this is a Σ_3 notion.

Exercise 2.4 *Give the definition (in the style of Definition 2.3) of* theorems interpretability.

Exercise 2.5 *Give a precise formulation of* $\mathsf{Ax}_{\mathrm{PA}+\beta}(x)$ *using* $\mathsf{Ax}_{\mathrm{PA}}(x)$. *Show how theorem interpretability can be reduced to axiom interpretability.*

Exercise 2.6 *The* identity interpretation *is precisely what we think it should be. Namely the one that does not do anything essential. Give this identity interpretation (including the domain specifier).*

Exercise 2.7 *Suppose that* $\mathrm{PA} \vdash \alpha \to \beta$. *Prove that* $\mathrm{PA} \vdash \alpha \rhd \beta$.

The previous exercise can be formalized giving rise to the following.

Exercise 2.8 *Prove that* $\mathrm{PA} \vdash \Box(\alpha \to \beta) \to \alpha \rhd \beta$.

In the latter exercise α and β play no specific role so that indeed it holds for any substitution and we have an interpretability principle of PA. Actually we just made the first step in an arithmetical soundness proof of $\mathbf{IL}\mathsf{M}$.

We can think of uniform operations on pairs of interpretations that are useful. One of these constructions is the *composition* as is illustrated in the next exercise.

Exercise 2.9 *Prove that* $\mathrm{PA} \vdash (\alpha \rhd \beta) \wedge (\beta \rhd \gamma) \to \alpha \rhd \gamma$. *Do this by defining the composition operator. Describe how it should work rather than giving the definition in full arithmetical detail.*

Another operation on interpretations is the *disjunction* that is to be defined and used in the next exercise.

Exercise 2.10 *Prove that* $\mathrm{PA} \vdash (\alpha \rhd \gamma) \wedge (\beta \rhd \gamma) \to \alpha \vee \beta \rhd \gamma$.

The next exercise deals with an arithmetization of the fact that relativized interpretations give rise to relative consistency proofs.

Exercise 2.11 *Prove that* $\mathrm{PA} \vdash (\alpha \rhd \beta) \wedge \mathsf{Con}\alpha \to \mathsf{Con}\beta$.

[7] Note that we have chosen the notion of so-called *axioms interpretability* rather than the notion of *theorems interpretability*.

3 More exercises

Exercise 3.1 *Show that* $\mathbf{IL} \nvdash (A \rhd B) \wedge (A \rhd C) \to A \rhd B \wedge C$.

Exercise 3.2 *Show that* $\mathbf{IL} \nvdash (A \rhd B) \vee (B \rhd A)$.

Exercise 3.3 *Show that* $\mathbf{IL} \nvdash A \rhd B \to \Box(A \to B)$.

Exercise 3.4 *Show that* $\mathbf{IL} \nvdash A \rhd B \to \Box(A \to B \vee \Diamond B)$.

Exercise 3.5 *Show that* $\mathbf{IL} \nvdash A \rhd \Diamond A$

Exercise 3.6 *Describe* $(t = x)^j$.

Exercise 3.7 *Show that the following statements are all provable in* **IL***:*

1. $A \rhd A$,
2. $\Diamond\Diamond A \rhd \Diamond A$,
3. $(A \rhd B) \wedge \Box(B \to C) \to A \rhd C$,
4. $A \rhd (A \wedge \Box\neg A) \vee (A \wedge \neg\Box\neg A)$,
5. $\Diamond A \leftrightarrow \Diamond(A \wedge \Box\neg A)$,
6. $A \rhd A \wedge \Box\neg A$,
7. $A \rhd A \vee \Diamond A$,
8. $A \vee \Diamond A \rhd A$,
9. $A \vee B \rhd C \to (A \rhd C) \wedge (B \rhd C)$.

References

[Sha94] V.Yu. Shavrukov. A smart child of Peano's. *Notre Dame Journal of Formal Logic*, 35(2):161–185, 1994.