Formalized Interpretability in Primitive Recursive Arithmetic

Joost J. Joosten
Department of Philosophy, University of Utrecht,
Heidelberglaan 8, 3584CS Utrecht, The Netherlands
http://www.phil.uu.nl/~jjoosten/
jjoosten@phil.uu.nl

Abstract. Interpretations are a natural tool in comparing the strength of two theories. In this paper we give a brief introduction to the topic of interpretability and interpretability logics. We will focus on the, so far, unknown interpretability logic of PRA. One research technique will be treated. This technique can be best described as restricting the realizations in the arithmetical semantics.

1 What are interpretations and why study them?

How to interpret "Eli, Eli, lama sabachtani"? Let us consider the concept of interpretation in the previous phrase[1]. What does it actually mean to interpret something. Or more specifically, what do we mean when we say that T interprets some utterance φ of S? Well, in this case T can first *translate* φ to its own language, then *place it in an adequate context* and then somehow *make sense* of it.

The mathematical notion of interpretation is somewhat similar. We say that a theory T interprets another theory S whenever there is some translation such that all translated theorems of S become provable in T. We give a precise definition. Throughout this paper we will stay in the realm of first-order logic.

Definition 1 *$\mathcal{K}$ is a relative interpretation of a theory S into a theory T, we write $\mathcal{K} : T \rhd S$, whenever the following holds. $\mathcal{K}$ is a pair $\langle \delta, F \rangle$. The first component, δ, is a formula in the language of T with a single free variable. This formula is used to specify the domain of our interpretation in a sense that we will see right now. The second component, F, is an*

[1] "Eli, Eli, lama sabachtani" were Jesus' last words. Some scholars translate this to "My God, my God, why hast thou forsaken me?". Others read it as "My God, my God, how thou dost glorify me!".

Proceedings of the Eighth ESSLLI Student Session
Balder ten Cate (editor)

easy (primitive recursive) map that sends formulas ψ in the language of S, to formulas $F(\psi)$ in the language of T. We demand for all ψ that the free variables of ψ and $F(\psi)$ are the same. The map F should commute with the boolean connectives, like $F(\alpha \wedge \beta) = F(\alpha) \wedge F(\beta)$. Moreover F should relativize the quantifiers to our domain specifier δ. Thus, for example $F(\forall x\ \alpha) = \forall x\ (\delta(x) \rightarrow F(\alpha))$.

We think this notion of interpretation is a natural one and comes close to our every day use of the concept of interpretation. And indeed it is a natural tool in comparing the the proof strength of two theories.

A first guess to say what it means that some theory T is at least as strong as some other theory S could be the following. Whenever S sees the truth of a formula ψ, T should also be able to see the truth of ψ. But, S and T might speak different languages. This is where the idea of a translation comes in.

Of course the translation should preserve some structure. Also it seems unreasonable that T should have the same domain of discourse as S. Taking these considerations into account it comes quite natural to say that T is at least as strong as S whenever T interprets S in the sense of Definition 1.

In the mathematical and metamathematical literature the here defined notion of interpretation turns up time and again. Perhaps the most famous example is in the proof of the consistency of non-euclidean geometry. In this proof (see for example [Gre96]) a model for non-euclidean geometry is built in a uniform way inside a model for euclidean geometry. Of course we somehow "know" that euclidean geometry is consistent. This uniform model construction is really nothing but an interpretation.

Tarski, Mostowski and Robinson first studied interpretations as a (meta) mathematical tool in a systematic way in [TMR53]. They also used interpretations to determine the undecidability of certain theories. It is not hard to convince oneself that some consistent theory T is undecidable whenever T interprets some essentially undecidable theory S. We say that S is essentially undecidable if S is undecidable and every consistent extension of T in the same language is also undecidable.

2 Formalized interpretability

In the previous section we have introduced the mathematical notion of interpretability. We have given some arguments to plea that it is a natural and interesting notion to consider. In this section we will add one more argument to our list. We will see that theories can in a certain way speak about interpretations. This insight will provide us with a simple yet expressive formalism in which large parts of metamathematical practise are expressible.

Amongst these are the Model Existence lemma as used in Gödel's Completeness theorem, Gödel's Second Incompleteness theorem, but also the method of relative consistency using interpretations.

Ever since Gödel we know that in theories of some minimal strength we can code syntax and syntactical notions like provability. We write $\Box_T\varphi$ for the very long statement that codes the fact that the sentence φ is provable in the theory T. As usual we denote $\neg\Box\neg\varphi$ by $\Diamond\varphi$. Once we realize that the notion of provability can be coded in a theory, it does not come as a surprise that we can do the same for interpretability.

For, what does it mean that S is interpretable in T? This means that there is a primitive recursive translation such that all translated theorems of S are provable in T. With some sloppy notation this can be written down as $\exists j\,\forall x\,(\Box_S x \to \Box_T x^j)$. Indeed, it turns out that the notion of interpretability can be expressed by a Σ_3-sentence.[2] We will denote the formalized statement of T interprets S by $T \rhd S$.

In this paper we will, for reasons that will become clear below, be mainly interested in interpretability relations between theories that are both finite extensions of some base theory T. Thus, we are interested in statements of the form $(T+\alpha)\rhd(T+\beta)$ which we will abbreviate with $\alpha \rhd_T \beta$. When the base theory T is clear from the context we will even omit sometimes the T in $\rhd_T$ and in $\Box_T$.

After having introduced this notation we see that many interesting properties can be expressed. For example $(i) : \alpha\rhd\beta \to (\Diamond\alpha \to \Diamond\beta)$. The formula (i) expresses that $T+\beta$ is consistent whenever it is interpretable in a consistent theory $T+\alpha$. We would like to say that (i) actually holds for any choice of α and β. One way of doing so is by working with arithmetical realizations and modal logics.

Definition 2 *By* $\mathsf{Form}_{\mathbf{IL}}$ *we denote the set of formulas in the modal language of interpretability logic. This is the smallest set containing* $\bot$, $\top$, *a countable infinite set of propositional variables and being closed under the boolean connectives, a unary modal operator* $\Box$ *and a binary modal operator* $\rhd$. *The* $\Diamond$ *will be an abbreviation for* $\neg\Box\neg$.

Definition 3 *An* arithmetical realization (relative to a theory T) *is a map* $(\cdot)^*$ *that sends any propositional variable* p *to some arithmetical sentence* p^*. *This map is extended to* $\mathsf{Form}_{\mathbf{IL}}$ *by stipulating that it commutes with the boolean connectives and demanding that* $(\Box A)^* = \Box_T A^*$ *and that* $(A \rhd B)^* = A^* \rhd_T B^*$. *An* interpretability principle *of a theory* T *is a formula in* $\mathsf{Form}_{\mathbf{IL}}$ *that is provable in* T *under any arithmetical realization. By the* interpretability logic of T *we mean the set of all interpretability principles of* T *or some system generating this set. We write* $\mathbf{IL}(\mathrm{T})$.

[2] A Σ_3-sentence is one that starts with a sequence of an existential- then universal- and then again existential quantifier to be followed by some formula only containing bounded quantification.

Note that by $\rhd$ we might now denote either the modal operator or the formalized notion of interpretability. We are confident however that this will not cause any confusion. Also note that the interpretability logic of a theory is the interpretability behaviour of that theory as seen by itself.

The modal language we have just introduced is rather expressive. Gödel's Second Incompleteness theorem can be written down as $\Diamond\top \rightarrow \neg\Box\Diamond\top$. Some reflection learns us that $\Diamond\top \rightarrow \neg(\top \rhd \Diamond\top)$ can be seen as a generalized version of Gödel's Second Incompleteness theorem; Under the assumption of the consistency, the consistency itself is not only just not provable, but not even interpretable!

An interpretation of S in T provides in an obvious way a uniform procedure to define a model of S within any model of T. Thus, the formula $\Diamond A \rhd A$ expresses in a certain sense the Model Existence lemma; whenever A is consistent in T, we can make a model of $T + A$.

Now consider a theory T. What is the modal characterization of its interpretability logic? For two classes of theories the answer to this question is known. If T is finitely axiomatizable **IL**(T) is known to be **IL**P as defined below. If T is essentially reflexive[3] **IL**(T) is also known. It is **IL**M, which is defined below. (See for an overview of these results [Vis97].)

We now present a logic **IL** that generates interpretability formulas that are interpretability principles for any reasonable theory. The logic **IL** is the smallest set of formulas in $\mathsf{Form}_{\mathbf{IL}}$ that is closed under the necessitation rule $A/\Box A$ and under Modus Ponens that contains all propositional tautologies and all instantiations of the following axiom schemata.

$$
\begin{array}{ll}
\mathsf{L}_1: & \Box(C \rightarrow D) \rightarrow (\Box C \rightarrow \Box D) \\
\mathsf{L}_2: & \Box A \rightarrow \Box\Box A \\
\mathsf{L}_3: & \Box(\Box A \rightarrow A) \rightarrow \Box A \\
\mathsf{J}_1: & \Box(C \rightarrow D) \rightarrow C \rhd D \\
\mathsf{J}_2: & (C \rhd D) \wedge (D \rhd E) \rightarrow C \rhd E \\
\mathsf{J}_3: & (C \rhd E) \wedge (D \rhd E) \rightarrow C \vee D \rhd E \\
\mathsf{J}_4: & C \rhd D \rightarrow (\Diamond C \rightarrow \Diamond D) \\
\mathsf{J}_5: & \Diamond A \rhd A
\end{array}
$$

The logic that arises from only the provability schemes L_1-L_3 is often called **GL** after Gödel and Löb. In this logic we have only formulas which are built up using the $\Box$ modality. We call this class of formulas $\mathsf{Form}_{\mathbf{GL}}$.

Two other prominent principles are $\mathsf{M}: A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C$ and $\mathsf{P}: A \rhd B \rightarrow \Box(A \rhd B)$. The logic that arises by adding more axiom schemes to **IL** is denoted by **IL** with the names of the principles postfixed to it.

[3] A theory is reflexive if it proves the consistence of any finitely axiomatized subtheory. It is essentially reflexive if all its finite extensions are reflexive.

For no theory T that is neither finitely axiomatizable nor essentially reflexive, **IL**(T) is known. PRA is such a theory.

3 What is PRA?

Primitive Recursive Arithmetic, we will write PRA, is a theory that has been studied extensively in the literature. We can think of PRA as the theory with minimal strength that can do basic reasoning about primitive recursion. In a rudimentary form PRA was first introduced by Skolem in 1923. (See for a translation [Sko67].) The emergence of PRA is best understood in the light of Hilbert's programme and finitism (see [Tai81]).

The precise formulation is not very much to our interest in this paper but the reader may think of it as $\mathrm{I}\Delta_0$ (see for example [HP93]) together with the Σ_1 induction rule. The latter allows one to conclude $\forall x\, \sigma(x)$ from $\sigma(0)$ and $\forall x\, (\sigma(x) \to \sigma(x+1))$ whenever σ is a Σ_1-formula.

It is well known that PRA is a reflexive theory but not essentially reflexive. However, any extension of PRA by Σ_2-sentences is reflexive (see [Bek97]). This important feature of PRA is reflected in our treatise of a lowerbound of **IL**(PRA). It is worth noting that we use no specific properties of PRA in providing an upperbound for **IL**(PRA) and indeed our results hold for a large class of theories.

4 A specific research tool: restricting the possible arithmetical realizations.

As we mentioned before, it is unknown what is **IL**(PRA). In this situation lower and upper bounds are already quite informative. This section makes some comments on these bounds. Also we shall reflect a bit on one technique that is used in determining upperbounds.

A lowerbound PRA certainly is a reasonable theory according to [JV00]. From [JV00] we thus get for free that $\mathbf{IL}\mathsf{M}_0\mathsf{P}_0\mathsf{W} \subseteq \mathbf{IL}$(PRA). With these letters we refer to the corresponding schemata:

$$\begin{aligned}
&\mathsf{M}_0 : A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C\\
&\mathsf{P}_0 : A \rhd \Diamond B \to \Box(A \rhd B)\\
&\mathsf{W} : A \rhd B \to A \rhd B \wedge \Box\neg A
\end{aligned}$$

In [Joo03] two more interpretability principles of PRA are formulated.

$$\begin{aligned}
&\mathsf{B} : A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C\\
&\mathsf{Z} : (A \rhd B) \wedge (B \rhd A) \to A \rhd A \wedge B
\end{aligned}$$

In B we require that A be an ES_2 (essentially Σ_2) formula. In Z we require that both A and B be ED_2 (essentially Δ_2) formulas. These two classes of formulas are defined as follows.

$$\begin{array}{lcl} ES_2 & := & \Box\mathsf{Form}_{\mathbf{IL}} \mid \neg\Box\mathsf{Form}_{\mathbf{IL}} \mid ES_2 \wedge ES_2 \mid ES_2 \vee ES_2 \mid \neg(ES_2 \rhd \mathsf{Form}_{\mathbf{IL}}) \\ ED_2 & := & \Box\mathsf{Form}_{\mathbf{IL}} \mid \neg ED_2 \mid ED_2 \wedge ED_2 \mid ED_2 \vee ED_2 \end{array}$$

Consequently $\mathbf{IL}\mathsf{BM_0P_0WZ}$ is also a lowerbound for $\mathbf{IL}$(PRA).

An upperbound In our Definition 3 we defined $\mathbf{IL}$(T) to be the set of all interpretability principles of T. An interpretability principle of T is a modal formula in $\mathsf{Form}_{\mathbf{IL}}$ that is provable in T under any arithmetical realization.

Let $\mathsf{Sub}(\Gamma)$ be the set of realizations that take their values in Γ. We define the Γ-interpretability logic of T to be set of all formulas in $\mathsf{Form}_{\mathbf{IL}}$ that are provable in T under any realization in $\mathsf{Sub}(\Gamma)$. We denote this logic by $\mathbf{IL}_\Gamma$(T). Clearly we have that $\mathbf{IL}_\Delta(\mathrm{T}) \subseteq \mathbf{IL}_\Gamma(\mathrm{T})$ whenever $\Gamma \subseteq \Delta$. This observation can be used to obtain a rough upperbound for $\mathbf{IL}$(PRA). In order to do so, we first calculate the Γ-*provability* logic of PRA for a specific Γ. This is defined completely analogously to its interpretability variant and is denoted by $\mathbf{PL}_\Gamma$(PRA).

First we define the set $\mathcal{B}$ of arithmetical sentences as follows.

$$\mathcal{B} := \bot \mid \top \mid \Box(\mathcal{B}) \mid \Diamond(\mathcal{B}) \mid \mathcal{B} \to \mathcal{B} \mid \mathcal{B} \vee \mathcal{B} \mid \mathcal{B} \wedge \mathcal{B}$$

Definition 4 *The logic* **RGL** *is obtained by adding the linearity axiom schema* $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ *to* **GL**. *Here* $\boxdot B$ *is an abbreviation of* $B \wedge \Box B$.

The logic **RGL** (the **R** stands for restricted) has been considered before in the literature. It is the system J in Chapter 13 of Boolos' book [Boo93]. Ever since Solovay (see [Sol76]) we know that **GL** is the provability logic of any strong enough theory and certainly for PRA.

In the proof below we will make use of the standard modal semantics for **GL**. A **GL**-frame F is a pair $\langle W, R\rangle$ where W is a finite non-empty set of worlds and R is a transitive conversely well-founded relation on it. A **GL**-model is a triple $\langle W, R, \Vdash\rangle$. Here $\Vdash$ is a relation on $W \times \mathsf{Form}_{\mathbf{GL}}$ such that for all $m \in M$ the set $\{A \in \mathsf{Form}_{\mathbf{GL}} \mid m \Vdash A\}$ is a maximal **GL**-consistent one. Moreover we demand $m \Vdash \Box A \Leftrightarrow \forall n\ (mRn \to n \Vdash A)$. We write $M \models A$ and say that A holds on M if for all $m \in M$ we have $m \Vdash A$. For F a frame we write $F \models A$ if A holds on any model that has F as its underlying frame. It is well known that $\mathbf{GL} \vdash A$ if and only if A holds on all finite transitive and conversely well-founded models.

Theorem 5 $\mathbf{PL}_{\mathcal{B}}(\mathrm{PRA}) = \mathbf{RGL}$

PROOF OF THEOREM 5. Let L_n be the linear frame with n elements. For convenience we call the bottom world $n{-}1$ and the top world 0. It is well known that $\mathbf{RGL} \vdash A \Leftrightarrow \forall n\ (L_n \models A)$. Our proof will thus consist of showing that $\forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash A^* \Leftrightarrow \forall n\ (L_n \models A)$.

For the $\Leftarrow$ direction we assume that $\exists * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \nvdash A^*$ and show that for some $m{\in}\omega$, $L_m \not\models A$. So, fix a $*$ for which $\mathrm{PRA} \nvdash A^*$. The arithmetical formula A^* can be seen as a formula in the closed fragment of **GL**. By the completeness of **GL** we can find a **GL** model such that $M, x \Vdash \neg A^*$. By $\rho(y)$ we denote the rank of y, that is, the length of the longest R-chain that starts in y. Let $\rho(x) = n$. As the valuation of $\neg A^*$ at x solely depends on the rank of x (see for example [Boo93], Chapter 7, Lemma 3), we see that $L_{n+1}, n \Vdash \neg A^*$ for every possible valuation on L_{n+1} (we also denote this by $L_{n+1}, n \models \neg A^*$). We define $\mathbf{L}_{n+1}, m \Vdash p \Leftrightarrow L_{n+1}, m \models p^*$. It is clear that $\mathbf{L}_{n+1}, n \Vdash \neg A$.

For the $\Rightarrow$ direction we fix some $n{\in}\omega$ such that $L_n \not\models A$ and construct a $*$ in $\mathsf{Sub}(\mathcal{B})$ such that $\mathrm{PRA} \nvdash A^*$. Let $\mathbf{L}_n$ be a model with domain L_n such that $\mathbf{L}_n, n{-}1 \Vdash \neg A$. Instead of applying the Solovay construction we can assign to each world m the arithmetical sentence

$$\varphi_m := \Box_{\mathrm{PRA}}^{m+1}\bot \wedge \Diamond_{\mathrm{PRA}}^{m}\top.$$

(We define $\Box_{\mathrm{PRA}}^{0}\bot{:=}\bot$ and $\Box_{\mathrm{PRA}}^{n+1}\bot{:=}\Box_{\mathrm{PRA}}(\Box_{\mathrm{PRA}}^{n}\bot)$. From now on we will omit the subscript PRA.) It is easy to see that

1. $\mathrm{PRA} \vdash \varphi_l \to \neg\varphi_m \quad$ if $l \neq m$,
2. $\mathrm{PRA} \vdash \varphi_l \to \Box(\bigvee_{m<l} \varphi_m)$,
3. $\mathrm{PRA} \vdash \varphi_l \to \bigwedge_{m<l} \Diamond\varphi_m$.

We set $p^* := \bigvee_{\mathbf{L}_n, m \Vdash p} \varphi_m$. Notice that $*$ is in $\mathsf{Sub}(\mathcal{B})$. Using 1, 2 and 3 we can prove a truth lemma, that is, for all m

$$\begin{array}{ll} \mathbf{L}_n, m \Vdash C \Rightarrow \mathrm{PRA} \vdash \varphi_m \to C^* & \text{and} \\ \mathbf{L}_n, m \nVdash C \Rightarrow \mathrm{PRA} \vdash \varphi_m \to \neg C^*. & \end{array}$$

By this truth-lemma, $\mathbf{L}_n, n-1 \Vdash \neg A \Rightarrow \mathrm{PRA} \vdash \varphi_{n-1} \to (\neg A)^*$ and consequently $\mathrm{PRA} \vdash \Diamond\varphi_{n-1} \to \neg\Box A^*$. Thus $\mathbb{N} \models \Diamond\varphi_{n-1} \to \neg\Box A^*$. As φ_{n-1} is consistent with PRA we see that $\mathbb{N} \models \Diamond\varphi_{n-1}$ whence $\mathbb{N} \models \neg\Box A^*$ and thus $\mathrm{PRA} \nvdash A^*$. QED

Definition 6 *The logic* **RIL** *is obtained by adding the linearity axiom schema* $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ *to* **ILW**.

Theorem 7 $\mathbf{RIL} = \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$

PROOF OF THEOREM 7. We will expose a translation from formulas φ in $\mathsf{Form}_{\mathbf{IL}}$ to formulas φ^{tr} in $\mathsf{Form}_{\mathbf{GL}}$ such that

$$\mathbf{RIL} \vdash \varphi \Leftrightarrow \mathbf{RGL} \vdash \varphi^{\mathsf{tr}} \quad (*)$$
and
$$\mathbf{RIL} \vdash \varphi \leftrightarrow \varphi^{\mathsf{tr}}. \quad (**)$$

If we moreover know $(***)$: $\mathbf{RIL} \vdash \varphi \Rightarrow \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash \varphi^*$ we would be done. For then we have by $(**)$ and $(***)$ that

$$\forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash \varphi^* \leftrightarrow (\varphi^{\mathsf{tr}})^*$$

and consequently

$$\begin{array}{ll} \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash \varphi^* & \Leftrightarrow \\ \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash (\varphi^{\mathsf{tr}})^* & \Leftrightarrow \\ \mathbf{RGL} \vdash \varphi^{\mathsf{tr}} & \Leftrightarrow \\ \mathbf{RIL} \vdash \varphi. & \end{array}$$

We first see that $(***)$ holds. From our remarks concerning a lowerbound of $\mathbf{IL}(\mathrm{PRA})$ we know that $\mathbf{ILW} \subseteq \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. Thus it remains to show that $\mathrm{PRA} \vdash \Box(\Box A^* \to B^*) \vee \Box(\boxdot B^* \to A^*)$ for any formulas A and B in $\mathsf{Form}_{\mathbf{IL}}$ and any $* \in \mathsf{Sub}(\mathcal{B})$. As any formula in the closed fragment of $\mathbf{ILW}$ is equivalent to a formula in the closed fragment of $\mathbf{GL}$ (see [HŠ91]), Theorem 5 gives us that indeed the linearity axiom holds for the closed fragment of $\mathbf{GL}$.

Our translation will be the identity translation except for $\rhd$. In that case we define

$$(A \rhd B)^{\mathsf{tr}} := \Box(A^{\mathsf{tr}} \to (B^{\mathsf{tr}} \vee \Diamond B^{\mathsf{tr}})).$$

We first see that we have $(**)$. It is sufficient to show that $\mathbf{RIL} \vdash p \rhd q \to \Box(p \to (q \vee \Diamond q))$. We reason in $\mathbf{RIL}$. An instantiation of the linearity axiom gives us $\Box(\Box\neg q \to (\neg p \vee q)) \vee \Box((\neg p \vee q) \wedge \Box(\neg p \vee q) \to \neg q)$. The first disjunct immediately yields $\Box(p \to (q \vee \Diamond q))$.

In case of the second disjunct we get by propositional logic $\Box(q \to \Diamond(p \wedge \neg q))$ and thus also $\Box(q \to \Diamond p)$. Now we assume $p \rhd q$. By W we get $p \rhd q \wedge \Box\neg p$. Together with $\Box(q \to \Diamond p)$, this gives us $p \rhd \bot$, that is $\Box\neg p$. Consequently we have $\Box(p \to (q \vee \Diamond q))$.

We now prove $(*)$. By induction on $\mathbf{RIL} \vdash \varphi$ we see that $\mathbf{RGL} \vdash \varphi^{\mathsf{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in $\mathbf{GL}$. The only axioms where the $\Box A \to \Box\Box A$ axiom scheme is really used is in J_2 and J_4. To prove the translation of W we also need L_3.

If $\mathbf{RGL} \vdash \varphi^{\mathsf{tr}}$ then certainly $\mathbf{RIL} \vdash \varphi^{\mathsf{tr}}$ and by $(**)$, $\mathbf{RIL} \vdash \varphi$.

QED

We thus see that **RIL** is an upperbound for **IL**(PRA). Using the translation from the proof of Theorem 7, it is not hard to see that both the principles P and M are provable in **RIL**. Choosing larger Γ will generally yield a smaller $\mathbf{IL}_\Gamma$(PRA) and thus a sharper upperbound.

Finally we remark that if $\mathbf{RGL} \nvdash \varphi$, then φ is certainly not a provability principle. But in this case we can find a counterexample with a "clear (meta)mathematical" content.

I would like to thank Lev Beklemishev for many enligthening discussions and for pointing out an error in an earlier version of this paper. Also I would like to thank an anonymous reviewer who helped improving the readability of this paper.

Bibliography

[Bek97] L.D. Beklemishev. Induction rules, reflection principles, and provably recursive functions. *Annals of Pure and Applied Logic*, 85:193–242, 1997.

[Boo93] G. Boolos. *The Logic of Provability*. Cambridge University Press, Cambridge, 1993.

[Gre96] M.J. Greenberg. *Euclidean and Non-Euclidean Geometries, 3rd edition*. Freeman, 1996.

[HP93] P. Hájek and P. Pudlák. *Metamathematics of First Order Arithmetic*. Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[HŠ91] P. Hájek and V. Švejdar. A note on the normal form of closed formulas of interpretability logic. *Studia Logica*, 50:25–38, 1991.

[Joo03] J.J. Joosten. The closed fragment of the interpretability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. Logic Group Preprint Series 128, University of Utrecht, February 2003.

[JV00] J.J. Joosten and A. Visser. The interpretability logic of *all* reasonable arithmetical theories. *Erkenntnis*, 53(1–2):3–26, 2000.

[Sko67] T. Skolem. The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains. In J. van Heijenoort, editor, *From Frege to Gödel*, pages 302–333. iUniverse, Harvard, 1967.

[Sol76] R.M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 28:33–71, 1976.

[Tai81] W. Tait. Finitism. *Journal of Philosophy*, 78:524–546, 1981.

[TMR53] A. Tarski, A. Mostowski, and R. Robinson. *Undecidable theories*. North–Holland, Amsterdam, 1953.

[Vis97] A. Visser. An overview of interpretability logic. In M. Kracht, M. de Rijke, and H. Wansing, editors, *Advances in modal logic '96*, pages 307–359. CSLI Publications, Stanford, CA, 1997.