Estratto da

Atti degli incontri di logica matematica Volume 6, Siena 28-31 maggio 1989.

Disponibile in rete su http://www.ailalogica.it

Modal logic and interpretability

(extended abstract)

Alessandro Berarducci[1]

1. Introduction. We give an exposition of some results obtained while the author was a Ph.D. student at the University of California at Berkeley writing a dissertation under the direction of Prof. R. Solovay. Complete proofs can be found in the forthcoming paper [Berarducci]. Our results concern the notion of "intepretability" of a first order theory into another first order theory (in the sense of [Tarski] and [Feferman]). We consider in particular finite extensions of Peano Arithmetic (PA), namely those first order theories which have the form PA + φ, where PA is Peano Arithmetic and φ is a sentence in the language of PA. The main result gives an extension of Solovay's modal analysis of the notion of "provability" [Solovay] to the case of "interpretability".

2. Interpretability.

2.1. Definition. Let L and L' be first order languages. Assume for simplicity that L and L' are relational languages without equality. A "translation" of L into L' consists of a formula U(x) of L', called the universe of the translation, together with a map f which associates to each n-ary relation symbol R of L a formula f(R) of L' having exactly n free variables (say the first n variables of L').

[1]Current address: Universita' di L'Aquila, Dipartimento di Matematica Pura e Applicata, Coppito, L'Aquila, Italy.

2.2. Definition. Let F = (f, U) be a translation of L into L'. For each formula C of L we define inductively a formula C^F of L' by replacing each occurrence of an atomic formula $R(x_1, \ldots, x_n)$ in C with its translation $f(R)(x_1, \ldots, x_n)$, and each occurrence of a quantifier $\forall x$ with its relativized version $\forall x(U(x) \rightarrow \ldots)$. That is:

1) $R(x_1, \ldots, x_n)^F = f(R)(x_1, \ldots, x_n)$ if R is a relation symbol of L;

2) $(A \wedge B)^F = A^F \wedge B^F$;

3) $(\neg A)^F = \neg(A^F)$;

4) $(\forall x\, A)^F = \forall x\,(Ux \rightarrow A^F)$.

To avoid unwanted conflicts of bounded variables we assume that before defining C^F all the bounded variables occurring in the formulas f(R) and in the formula U(x) have been renamed so that none of them occurs (free or bound) in the formula C.

2.3. Definition. Given two first order theories T and S, in the languages L[T] and L[S] respectively, we say that T interprets S iff:

$\exists$ (f,U) such that F = (f,U) is a translation of L[S] into L[T] and:
$\forall$ A $\in$ Axioms of S
$\exists$ p : p is a proof of A^F from the axioms of T.

An interpretation of S in T gives us a canonical way of constructing a model M^F of S, starting from a model M of T: the underlying set of M^F is the subset of M consisting of all the elements satisfying U(x), and the relations on M^F are so defined that $M^F \models C(a_1, \ldots, a_n)$ iff $M \models C^F(a_1, \ldots, a_n)$. It is clear from

the definition of interpretability that for recursively axiomatized theories (in a finite language) the notion "T interprets S" can be formalized as a Σ^0_3 -formula in the language of arithmetic (uniformly in T and S).

2.4. Definitions.

1) $Interp_{PA}(x, y)$ is the Σ^0_3 -formula formalizing the assertion "x and y are (codes of) sentences of PA such that the theory PA∪ {x} interprets the theory PA∪ {y}". (Since we have defined interpretations only for relational languages we assume that PA has been formulated in a relational language.)

2) $Prov_{PA}(x)$ the Σ^0_1-formula expressing "x is (the code of) a sentence which is a theorem of PA".

3) $Prov_{PA,y}(x)$ is the Σ^0_1-formula (in the two variables x, y) asserting "there is a proof of x from PA which employs only axioms with Gödel numbers less than y".

The following theorem of Orey (cfr. [Feferman]) says that intepretability over PA is definable in terms of restricted provability:

2.5. Theorem. PA + φ interprets PA + Ψ iff for every finite subtheroy U of PA + Ψ , PA + φ proves the consistency of U. Moreover this equivalence can be proven in PA.

An immediate consequence of Orey's theorem is that the complexity of the formula $Interp_{PA}(x, y)$ can be reduced from Σ^0_3 to Π^0_2 (but not further, cfr. [Solovay2] and [Lindström]). Note that if instead of PA we consider a finitely axiomatized theory, like GB, then the notion of interpretability has complexity Σ^0_1. The behavior of GB with respect to interpretability has been studied in [Visser] and differs significantly from the one of PA (while both PA and GB share the same modal logic of provability by Solovay's result).

3. Modal logic. The way modal logic has been used to study formal provability and intepretability is through the introduction in the language of modal logic of modal operators $\Box$, $\Box_x$, and $\rhd$ whose intended meaning are $Prov_{PA}$, $Prov_{PA,x}$ and $Interp_{PA}$ respectively (other operators have also been considered with interesting applications, cfr. [Visser]). So for example Orey's theorem can be expressed by the modal formula (*): $A \rhd B \leftrightarrow \forall n \, \Box(A \to \Box_n B)$ showing in particular that $\rhd$ is definable in terms of $\Box$ and $\Box_x$. Since Orey's theorem is true (as Orey proved it) we say that the corresponding modal formula (*) is "valid". Moreover since the proof of Orey's theorem can be formalized in PA we say that (*) is not only valid but also "PA-valid". Another example of a valid modal formula is $\Box A \leftrightarrow (\neg A) \rhd \bot$ which says that provability can be defined in terms of intepretability. Gödel's second incompleteness theorem provides a third example of a valid formula: $\neg\Box\neg A \to \neg\Box(A \to \neg\Box\neg A)$. We can read this as an expression of the fact that if PA+A is consistent, then PA+A does

not prove its own consistency. The reflection principle for PA can also be expressed by a valid modal formula: $\forall n \, \Box(A \to \neg\Box_n \neg A)$, i.e. the theory PA+A proves the consistency of every finite subtheory of itself. Our last example of a valid modal formula is a principle discovered by F. Montagna, which is an expression of the fact that Σ^0_1-formulas are preserved under interpretations: $A \rhd B \to (A \wedge \Box D \rhd B \wedge \Box D)$. Montagna's principle would fail to be valid if we replaced the base theory PA with GB.

Considered the wealth of classical examples, a natural question is whether there is a decision procedure to test the validity of a modal formula. The first such decision procedure was obtained by [Solovay] for the restricted class of modal formulas containing only the provability operator $\Box$, propositional variables (standing for arbitrary sentences of PA), and boolean connectives (including a propositional constant $\bot$ for falsehood). The modal formula expressing Gödel's second incompleteness theorem is an example of such a formula, so this restricted class is already quite expressive.

3.1. Open problem: does such a decision procedure exists for the language containing all of the above mentioned operators, namely $\Box$, $\Box_x$, and $\rhd$ (with the possibility of quantifying over the variable x in $\Box_x$) ?

Our main result is that we still have a decision procedure for valid modal formulas in the language with both $\Box$ and $\rhd$ (but without $\Box_x$). To state this precisely we need:

3.2. Definition. Consider the modal language containing $\Box$, $\rhd$, boolean connectives, and propositional variables. Let H be a map which assigns to each propositional variable A a sentence A^H of PA. We extend H to all the modal formulas by preserving the boolean connectives and defining:

$(\bot)^H \equiv (0=1)$

$(\Box A)^H \equiv Prov_{PA}([A^H])$;

$(A \rhd B)^H \equiv Interp_{PA}([A^H], [B^H])$

where $[\varphi]$ is the numeral for the Gödel number of the PA-formula φ.

3.3. Definition. Let A be a modal formula. We say that A is PA-valid, if for all maps H as above, PA $\vdash A^H$. We say that A is ω-valid if for all H, $\omega \models A^H$.

Every PA-valid formula is clearly also ω-valid. An example of an ω-valid formula which is ω-valid but not PA-valid is the modal formula expressing the soundness of PA: $\Box A \to A$. Another example is the formula expressing the consistency of PA: $\neg\Box\bot$. Note that this latter formula does not have any propositional variable, so it corresponds to a single sentence of PA rather then to a scheme. Clearly A is PA-valid iff $\Box A$ is ω-valid.

3.4. Main theorem. It is decidable whether a modal formula (in the language with both $\Box$ and $\rhd$) is PA-valid. Similarly it is decidable whether any such modal formula is ω-valid.

This result has been obtained independently and at about the same by Shavrukov [Shavrukov]. Both proofs use earlier work of

Visser, De Jongh and Veltman on this problem, who provided us with the right conjecture, namely that the PA-valid formulas are exactly the theorems of the modal theory ILM (cfr. [Visser]), together with the necessary Kripke models to prove the decidability of ILM (cfr. [De Jongh-Veltman]).

3.5. Definition. The axioms of the theory ILM are all the boolean tautologies (including those containing $\Box$ and $\rhd$) plus the following axiom schemes (where $\Diamond$ stands for $\neg\Box\neg$):

1) $\Box(A\to B) \to (\Box A \to \Box B)$;

2) $\Box A \to \Box\Box A$;

3) $\Box(\Box A\to A) \to \Box A$;

4) $\Box(A\to B) \to A \rhd B$;

5) $(A \rhd B \wedge B \rhd C) \to (A \rhd C)$;

6) $(A \rhd C \wedge B \rhd C) \to (A \vee B \rhd C)$;

7) $A \rhd B \to \Diamond A \to \Diamond B$;

8) $\Diamond A \rhd A$;

9) $A \rhd B \to A \wedge \Box D) \rhd (B \wedge \Box D)$.

The rules of inference are modus ponens and necessitation: $A / \Box A$.

To prove our main result we show:

3.6. Theorem. The PA-valid formulas are exactly the theorems of ILM.

3.7. Theorem. The ω-valid formulas are exactly the theorems of the theory ILM^ω which is defined like ILM except that we omit the rule of inference $A/\Box A$ and we add the axiom scheme $\Box A \to A$.

Moreover ILM^{ω} can be many one reduced to the decidable theory ILM so it is still a decidable theory.

The reduction of ILM^{ω} to ILM can be described as follows: $ILM^{\omega} \vdash C$ iff ILM $\vdash T(C) \rightarrow C$ where T(C) is the conjunction of: 1) all the formulas of the form $\Box\neg A \rightarrow \neg A$ such that for some B, $A \rhd B$ is a subformula of C; 2) all the formulas of the form $\Box A \rightarrow A$ such that A is a subformula of C.

The proof of 3.6 and 3.7 is constructive in the sense that it can be used to find sentences of PA with a preassigned behavior with respect to interpretability and provability whenever such PA-formulas exist (and to decide if they do exist): for example we can prove the non-validity of the formula $A \rhd B \rightarrow \Box(A \rhd B)$ by explicitly constructing two sentences A and B of PA which falsify it, namely such that the theory PA+A intepretes PA+B but it does so in such a nonconstructive way that PA is not able to formalize the proof that PA+A interprets PA+B (this phenomenon would not be possible if we replaced PA with the finitely axiomatized GB). Even for such a simple example it would not be easy to prove that such sentences exist without resorting to the general theorem.

References.

[Berarducci] A. Berarducci, The interpretability logic of Peano Arithmetic, to appear in the Journal of Symbolic logic.

[De Jongh-Veltman] D. De Jongh and F. Veltman, Provability logic for relative interpretability. Proceedings of Heyting '88 (Bulgaria).

[Feferman] S. Feferman, Arithmetization of methamatematics in a general setting, Fundamenta Mathematicae, 49, pp. 33-92, 1960.

[Linström] Provability and interpretability in theories containing arithmetic, In Atti degli incontri di logica matematica 2, pp. 431-451, Siena, Italy, 1984.

[Shavrukov] V. Y. Shavrukov, Logic of relative interpretability over Peano Arithmetic, Preprint 5 of Steklov Math. Institute, Dec. 1988.

[Smorynski] C. Smorynski, Self Reference and Modal Logic, Springer-Verlag, 1985.

[Solovay] R. M. Solovay, Provability Interpretations of Modal Logic, Israel Journal of Mathematics, 25, pp. 287-304, 1976.

[Solovay2] R. M. Solovay, Interpretability in set theories, unpublished manuscript, 1976.

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

[Visser] A. Visser, Preliminary notes on Interpretability Logic, Logic Group Preprint Series No. 29, Department of Philosophy, University of Utrecht, 1988.