Commentationes Mathematicae Universitatis Carolinae

Petr Hájek

Partial conservativity revisited

Commentationes Mathematicae Universitatis Carolinae, Vol. 28 (1987), No. 4, 679--690

Persistent URL: `http://dml.cz/dmlcz/106582`

COMMENTATIONES MATHEMATICAE UNIVERSITATIS CAROLINAE
28,4 (1987)

PARTIAL CONSERVATIVITY REVISITED

Petr HÁJEK

Abstract: We study the notions of partial conservativity and interpretability over fragments of Peano arithmetic.

Key words: Partial conservativity, interpretability, first order arithmetic.

Classification: 03F30, 03F25, 03D35

Introduction. Let S,T be theories, $S \subseteq T$, and let Γ be a class of formulas of T. S is Γ-conservative over T if for each $\psi \in \Gamma$, provability of ψ in S implies provability of ψ in T. In particular, if φ is a T formula then φ is said to be Γ-conservative over T if the theory $(T+\varphi)$ is T-conservative over T. "Partial conservativity" means "Γ-conservativity for some Γ". The first nontrivial example of partial conservativity was exhibited by Kreisel [62]; the first systematic paper on partial conservativity is Guaspari [76] (containing also results by Solovay). Then various people contributed, among them Lindström, Smoryński and the present author (see references). A recent work is Bennet [86]. These papers typically discuss partial conservativity over theories in the language of arithmetic containing PA (Peano arithmetic); but equally typically, contain a remark saying that the assumptions on the underlying theory are in fact too strong and a weaker theory would suffice. One of the advantages of former theories is that for such a theory T, interpretability of $(T+\varphi)$ in T is equivalent to Π_1-conservativity of $(T+\varphi)$ over T and thus results on interpretability can be obtained as corollaries. Here we work systematically with theories containing $I\Sigma_1$ (i.e. arithmetic with induction restricted to Σ_1-formulas) but having possibly a richer language. One basic difference of fragments $I\Sigma_n$ from the whole PA is that they are finitely axiomatizable. As we shall see, this does not affect much properties of partial conservativity but does affect properties of interpretability. For positive results on interpretability we shall heavily use a result due to Pudlák (formulated below) and its strengthening. A second obstacle is the fact that over a weak theory like $I\Sigma_1$, Σ_n formulas are not closed under bounded universal quantification: but it turns out that this can be easily overcome. The result of our investigation is a systematic treatment of partial conservativity over theories containing $I\Sigma_1$ and of its relation to interpretability. We shall give only very sketchy hints on proofs; full proofs are contained in a Czech typescript not for publication; the proofs will also be incorporated in a forthcoming book. The paper is organized as follows: § 1 contains preliminaries, § 2 discusses prominent examples (Gödel's and Rosser's formulas), § 3 presents general theorems on partial conservativity, § 4 contains applications to interpretability and § 5 elaborates a classification of independent Σ_1 sentences.

§ 1. Preliminaries. We say "interpretation" meaning "relative interpretation with absolute equality" (cf. Tarski, Mostowski, Robinson [53]). First recall the old result on interpretability for PA and similar theories.

1.1. Theorem (Orey, Hájek, Guaspari). Let S, T be axiomatized theories in the language of PA and let $S \supseteq T \supseteq PA$. Then the following are equivalent:

(i) S is interpretable in T,

(ii) S is Π_1-conservative over T,

(iii) for each k, $T \vdash Con_{S\upharpoonright k}$,

(iv) there is a binumeration β of S in T such that $T \vdash Con_\beta$.

(Similarly for $S \supseteq T \supseteq ZF$ in the language of ZF and in general for $S \supseteq T$ where T contains PA and proves induction for all T-formulas.)

See Orey [61], Hájek [72], Guaspari [76].

The following lemma is easy but basic for our considerations.

1.2. Lemma. Let $\varphi(x)$ be a Σ_n-formula (whose free variables are x and possibly others. There is a Σ_n-formula $\psi(y)$ such that

(1) $(\forall k)\ I\Sigma_1 \vdash \psi(k) \equiv (\forall x \leq k)\varphi(x)$,

(2) $I\Sigma_1 \vdash \psi(y) \rightarrow (\forall x \leq y)\varphi(x)$.

Proof. Trivial for n=0. For $n \geq 1$ and $\varphi(x) \equiv (\exists u)\alpha(x,u)$ let $\psi(y)$ be $(\exists s)(Seq(s) \& (\forall x \leq y)\alpha(x,(s)_x))$.

1.3. Notation. The formula $\psi(y)$ from the previous lemma will be denoted $[(\forall x \leq y)\varphi(x)]^{*,\Sigma,n}$ or simply $[(\forall x \leq y)\varphi(x)]^*$. Dually, for each Π_n-formula $\varphi(x)$ we have a Π_n-formula $\psi(y)$ (denoted by $[(\exists x \leq y)\varphi(x)]^{*,\Pi,n}$) such that

$$(\forall k)I\Sigma_1 \vdash \psi(k) \equiv (\exists x \leq k)\varphi(x),$$
$$I\Sigma_1 \vdash (\exists x \leq y)\varphi(x) \rightarrow \psi(y).$$

1.4. Discussion and definition (proofs from true formulas; herbrandian proofs). It is well known that in $I\Sigma_1$ we may define partial truth predicates: for each n we have a Σ_n truth definition for all Σ_n sentences satisfying the usual Tarski's conditions and similarly for Π instead of Σ. If Γ is Σ_n or Π_n then $Tr_\Gamma(x)$ means the corresponding truth predicate; $I\Sigma_1 \vdash \varphi \equiv Tr\,(\overline{\varphi})$ for each $\varphi \in \Gamma$. We make the following definition (y is a proof of x from a true Γ-formula):

$$Prf_\Gamma(x,y) \equiv (\exists z \leq y)(z \in \Gamma \,\&\, Tr_\Gamma(z) \,\&\, Prf(z \rightarrow x,y)).$$

Clearly, if Γ is Σ_n then Prf_Γ is Σ_n in $I\Sigma_1$: a trivial transformation shows that if Γ is Π_n then Prf_Γ is Π_n in $I\Sigma_1$. Here Prf is the proof predicate for a given theory defined in "Hilbert style". We shall also use the "Herbrand style" proof predicate HPrf investigated by Pudlák [85], but only for finitely axiomatized theories T. In words, a herbrandian proof of φ in T is a propositional proof of a disjunction of instances of the quantifier-free part of $He(\bigwedge T \rightarrow \varphi)$, where, for any ψ, $He(\psi)$ is the purely existential Herbrand form of ψ.
Similarly we define $HPrf_\Gamma$ for Γ being Σ_n or Π_n.

Pr(x) and HPr(x) are provability predicates (usual and herbrandian); clearly, $I\Sigma_1 \vdash (\forall x)(Pr(x) \equiv HPr(x))$. But this equivalence does not relativize to definable cuts - see next definition.

1.5. Definition. (1) Let $T \supseteq Q$ (Q is Robinson's arithmetic). A formula J(x) with one free variable defines a cut in T if T proves the following:

$$J(0) \,\&\, (\forall x)(J(x) \rightarrow J(x+1) \,\&\, (\forall x,y)(y < x \,\&\, J(x) \rightarrow J(y)).$$

Note that if T is PA then $T \vdash (\forall x)J(x)$ for each such J; but this is not the case for any fragment $I\Sigma_n$ - there are cuts J such that $(\forall x)J(x)$ is unprovable.

(2) A theory $T \supseteq Q$ is sequential if it has coding of finite sequences of arbitrary objects, i.e. has a predicate Seq and functions $(s)_x$, lh(s) such that the following is provable in T: there is an empty sequence (of length 0), the length of each sequence is a number and for each sequence s of a length x and an arbitrary object z there is a sequence s' of length x+1 prolonging s by z. (See Pudlák [85]).

(3) ACA_0 is the usual (fully) conservative second order extension of PA, i.e. ACA_0 has two sorts of variables (numbers and sets), axioms of PA for numbers with the induction scheme replaced by a single axiom

$0 \in X \& (\forall x)(x \in X \rightarrow x+1 \in X)$ and with comprehension for all formulas containing no set quantifiers. (Note that ACA_0 is sequential; ACA_0 is related to PA as GB to ZF.)

1.6. Theorem (Pudlák [85]). (1) Let T be a consistent sequential theory, S a finitely axiomatized theory. S is interpretable in T iff there is a definable cut J in T such that $T \vdash HCon^J(S)$, i.e. T proves that J does not contain any Herbrandian proof of the contradiction from S.

(2) In particular, if T is consistent, finitely axiomatizable and sequential then there is a definable cut J in T such that $T \vdash HCon^J(T)$.

(3) On the contrary, for such a T there is no cut J such that $T \vdash Con^J(T)$.

Point (2) can be strengthened as follows:

1.7. Theorem. Let T be finitely axiomatizable, sequential and let $T \vdash I\Sigma_1$. Then there is a cut J in T such that

$T \vdash (\forall u)(Tr_{\Sigma_1}(u) \rightarrow HCon^J((T+u)))$.

(Proof by inspection of Pudlák [85] - tedious.)

§ 2. Some prominent examples

2.1. We shall investigate the properties of Gödel's consistency formula Con_T and Rosser's formula ρ_T, for T being either an extension of PA in the same language or a finitely axiomatized extension of $I\Sigma_1$. (In the former case we assume a fixed Δ_1 binumeration of the axioms to be given; in the latter we work with the natural binumeration just listing the axioms.) Both formulas may be constructed using either Prf or HPrf; for Con_T this is immaterial (see above), but for the Rosser's formula (which we assume in Σ_1 form, i.e. saying "there is a proof y of my negation such that no $z < y$ is a proof of me") there may be differences: some results below hold only for the Rosser's formula based on HPrf, say, the H-Rosser formula.

2.2. **Convention:** If T is a theory and φ a formula of T we say that φ is interpretable in T meaning that the theory $(T+\varphi)$ is interpretable in T.

2.3. **Theorem.** Let $T \supseteq I\Sigma_1$ be consistent (and axiomatizable).

(1) Gödel's formula Con is not interpretable in T and its negation $\neg$Con is interpretable in T.

(2) $\neg$Con is Π_1 conservative; Con is Σ_1-conservative iff T is Σ_1-sound (i.e. each provable Σ_1-formula is true in N).

(3) Rosser's formula ρ is Π_1-nonconservative; $\neg\rho$ is Σ_1-conservative

iff T is Σ_1-sound. (The same holds for the H-Rosser formula.)

(4) If T is sequential and has induction for all formulas then neither φ nor $\neg\varphi$ is interpretable. (The same for H-Rosser formula.)

(5) But if T is sequential and finitely axiomatizable and φ is the H-Rosser formula then both φ and $\neg\varphi$ are interpretable.

Comments on proofs. (1) For the first claim see Feferman [60] and Švejdar [78]; for the second see Feferman [60] where $T \supseteq PA$ is assumed. The same result can be proved:

(a) for $T \supseteq B\Sigma_2$ using Low Basis Theorem (see Clote [83]) and the corresponding Low Arithmetized Completeness Theorem,

(b) for $T \supseteq I\Sigma_1$ finitely axiomatized using Second Gödel's incompleteness theorem and Pudlák's theorem 1.6 and

(c) in full generality i.e. for any $T \supseteq I\Sigma_1$ using a version of Low Basis Theorem in $I\Sigma_1$ (see Hájek and Kučera [∞]).

(2) The first claim is due to Kreisel [68] and is the first example of non-trivial partial conservativity. Checking for $T \supseteq I\Sigma_1$ is immediate. The second claim is due to Smoryński [80].

(3) First claim is due to Kreisel [62], second to Švejdar (unpublished)

(4) Follows easily from (3) and from 1.1.

(5) Pudlák's theorem 1.6 gives an interpretation of $(T+\varphi)$ in $(T+\neg\varphi)$ and vice versa.

Problem: does (5) also hold for the usual (non-herbrandian) Rosser formula?

2.4. **Corollary.** If T is sequential and finitely axiomatized then there is a Π_1 formula φ such that both φ and $\neg\varphi$ are interpretable in T.(For PA and similar theories there is no such φ .) In particular, if T is ACA_0 then $(ACA_0+\varphi)$ is interpretable in ACA_0 but $(PA+\varphi)$ is not interpretable in PA. Similarly for GB and ZF instead of ACA_0 and PA. First example of such a φ was constructed by Solovay (unpublished, cf. Hájek [81]).

§ 3. **General theorems on partial conservativity.** We shall present several theorems on partial conservativity. Their proofs use various generalizations of Rosser's formula. In the whole section $T \supseteq I\Sigma_1$ is a fixed consistent axiomatized theory (and we assume a Δ_1 binumeration of T to be fixed).

3.1. **Notation.** (1) Let $\alpha(u)$, $\beta(u)$ be two T-formulas, let Δ be $(\exists u)\alpha(u)$ and ∇ be $(\exists u)\beta(u)$. (α, β may contain parameters.) Following

Guaspari we denote by $\Delta \prec \nabla$ the formula

$$(\exists u)(\alpha(u) \& (\forall v \leq u) \neg \beta(V))$$

(there is a witness for α less than each witness for β).

(2) If β is Π_n and β' is the Σ_n formula naturally equivalent to $\neg\beta$ then $\Delta \prec^* \nabla$ will denote the formula

$$(\exists u)(\alpha(u) \& [(\forall v \leq u)\ \beta'(v)]^{*,\Sigma,n})$$

(assuming that n is clear from the context). Similarly, $\Delta \preccurlyeq^* \nabla$ is $(\exists u)(\alpha(u) \& [(\forall v < u)\ \beta'(v)]^{*,\Sigma,n})$.

3.2. **Remark.** We shall investigate selfreferential formulas satisfying

$$T \vdash \xi \equiv \Delta(\neg\overline{\xi}) \prec^* \nabla(\overline{\xi})$$

or, more generally, for each k,

$$T \vdash \xi(\overline{k}) \equiv \Delta(\neg\overline{\xi}, \overline{k}) \prec^* \nabla(\xi, \overline{k}).$$

Observe that if α is Σ_n and β is Π_n then ξ is Σ_n in T.

(2) If Δ_i is $(\exists u)\alpha_i(u)$ and ∇_i is $(\exists v)\beta_i(v)$ (i=1,2) then $(\Delta_1 \vee \Delta_2) \prec (\nabla_1 \vee \nabla_2)$ means the formula saying "there is a witness for $\alpha_1 \vee \alpha_2$ less than each witness for $\beta_1 \vee \beta_2$"; similarly for $\prec^*$ instead of $\prec$.

3.3. **Definition.** (1) φ is <u>hereditarily</u> Γ-<u>conservative</u> over T if, for each T_0 such that $I\Sigma_1 \subseteq T_0 \subseteq T$, φ is Γ-conservative over T_0.

(2) φ is <u>doubly</u> Γ-<u>conservative</u> over T if φ is Γ-conservative over T and $\neg\varphi$ is $\check{\Gamma}$-conservative over T (where $\check{\Gamma}$ is the dual class of Γ).

We shall now formulate three general theorems on partial conservativity.

3.4. **Theorem.** For each $n \geq 1$ there is (1) a hereditarily Π_n-conservative Σ_n-sentence, (2) a hereditarily Σ_n-conservative Π_n-sentence, (3) a doubly Π_n-conservative Σ_n-sentence (its negation is thus a doubly $\Sigma_n$conservative Π_n-sentence).

Examples (Γ is Σ_n, Λ is Π_n):

(1) ξ such that $I\Sigma_1 \vdash \xi \equiv Pr_\Gamma(\neg\overline{\xi}) \prec^* Pr(\overline{\xi})$,

(2) $(\neg\xi)$ such that $I\Sigma_1 \vdash \xi \vdash Pr(\neg\overline{\xi}) \prec^* Pr_\Lambda(\overline{\xi})$,

(3) ξ such that $I\Sigma_1 \vdash \xi \leq Pr_\Gamma(\neg\overline{\xi}) \prec^* Pr_\Lambda(\xi)$.

If T is Σ_n-sound we may take in (1)a ξ such that $I\Sigma_1 \vdash \xi \equiv Pr_\Gamma(\neg\overline{\xi})$.

3.5. **Theorem** (on non-separability). Let Γ be Σ_n or Π_n $(n \geq 1)$, Th be the set of all theorems of T, Consv(Γ) and hConsv(Γ) the set of all Γ-conservative and hereditarily Γ-conservative sentences respectively , NRef the set of all the sentences non-refutable in T. Then obviously

$$\mathrm{Th} \subseteq \mathrm{hConsv}(\Gamma) \subseteq \mathrm{Consv}(\Gamma) \subseteq \mathrm{Consv}(\Sigma_1) \cap \mathrm{Consv}(\Pi_1) \subseteq \mathrm{Nref}$$

and there is no set X such that

(1) X is Δ_1 and $\mathrm{Th} \subseteq X \subseteq \mathrm{NRef}$ (classical!),

(2) X is Π_1 and $\mathrm{Th} \subseteq X \subseteq \mathrm{Consv}(\Gamma)$,

(3) X is Σ_1 and $\mathrm{hConsv}(\Gamma) \subseteq X \subseteq \mathrm{NRef}$,

(4) X is Σ_2, $\Gamma \supseteq \Sigma_1$ and $\mathrm{hConsv}(\Gamma) \subseteq X \subseteq \mathrm{Consv}(\Sigma_1)$,

(4´) X is Σ_2, $\Gamma \supseteq \Pi_1$ and $\mathrm{hConsv}(\Gamma) \subseteq X \subseteq \mathrm{Consv}(\Pi_1)$.

3.6. **Theorem** (Π_2-completeness). For each $n \geq 1$ and $\Gamma = \Sigma_n$ or Π_n, both $\mathrm{Consv}(\Gamma)$ and $\mathrm{hConsv}(\Gamma)$ is Π_2-complete.

3.7. **Remark.** Theorem 3.4 was obtained for $T \supseteq PA$ by Guaspari and Solovay, see Guaspari [67]; their examples are more complicated than ours.
Theorem 3.5:
(1) is very classical, (2) seems to be new. (3) was first proved in the particular case T=ZF, $\Gamma = \Pi_1$, interpretability instead of partial conservativity in Hájek [71]; Lindström [84] generaliz for $T \supseteq PA$, the present generalization is mine. (4) is contained (implicitly) in Lindström [84] for $T \supseteq PA$.
Theorem 3.6:
For T=ZF and $\Gamma = \Pi_1$ (and interpretability) Solovay; his proof works for PA. For PA, $\Gamma = \Pi_n$ and Cons see Hájek [79], for PA, $\Gamma = \Sigma_n$ and Consv see Quinsey [81]. In full generality but for $T \supseteq PA$ see Lindström [84]; generalization to $T \supseteq I\Sigma_1$ is mine.

We shall present two general fixed point theorems that form the main means of proofs of the preceding theorems.

3.8. **Sheperdson-Smoryński´s fixed point theorem.** Let Φ, Ψ be Σ_1 formulas.

(1) Let $I\Sigma_1 \vdash \xi \equiv [(Pr(\neg\overline{\xi}) \vee \Phi) \prec (Pr(\xi) \vee \Psi)]$. Then

(i) $T \vdash \xi$ iff $N \models \Phi \prec \Psi$ iff $N \models \xi$;

(ii) $T \vdash \neg\xi$ iff $N \models \Psi \preceq \Phi$.

(2) More generally, let, for i=1,2, $T_i \supseteq I\Sigma_1$, let Pr_i be the proof predicate based on a fixed Δ_1 binumeration of T_i. Let

$$I\Sigma_1 \vdash \xi \equiv [(Pr_1(\neg\overline{\xi}) \vee Pr_2(\neg\xi) \vee \Phi) \prec (Pr_1(\xi) \vee Pr_2(\xi) \vee \Psi)].$$ Then

(i) $T_1 \vdash \xi$ iff $T_2 \vdash \xi$ iff $N \models \Phi \prec \Psi$ iff $N \models \xi$;

(ii) $T_1 \vdash \neg\xi$ iff $T_2 \vdash \neg\xi$ iff $N \models \Psi \preccurlyeq \Phi$.

For proofs see Shepherdson [60], Smoryński [80].

3.9. **Lindström's fixed point theorem.** (Let T be as above.)

(1) Let $\chi(y)$ be Σ_n and let

$I\Sigma_1 \vdash \xi \equiv Pr_{\Sigma_n}(\neg\overline{\xi}) \prec^{*} (\exists y)\neg\chi(y)$. Then

(i) for each m, $(T+\xi) \vdash \chi(m)$,

(ii) for each $I\Sigma_1 \subseteq T_0 \subseteq T$ and each Π_n-sentence π, $(T_0+\xi) \vdash \pi$ implies $T_0+\{\chi(m)|m\} \vdash \pi$.

(2) Let $\chi(y)$ be Π_n and let

$I\Sigma_1 \vdash \xi \equiv (\exists y)\neg\chi(y) \prec^{*} Pr_{\Pi_n}(\xi)$. Then

(i) for each m, $(T+\neg\xi) \vdash \chi(m)$,

(ii) for each $I\Sigma_1 \subseteq T_0 \subseteq T$ and each Σ_n-sentence σ, $(T_0+\neg\xi) \vdash \sigma$ implies $T_0+\{\chi(m)|m\} \vdash \sigma$.

For $T \supseteq PA$ see Lindström [84].

3.10. **Remark.** Both fixed point theorems may be <u>parametrized</u> (by replacing ξ, Φ, Ψ, $\chi(y)$ by $\xi(\overline{k})$, $\Phi(\overline{k})$, $\Psi(\overline{k})$, $\chi(y,\overline{k})$ respectively); details are evident. Shepherdson-Smoryński theorem is used to prove 3.5 (1),(2) (and is very useful at many other occasions); Lindström's theorem is used for the rest of 3.5 and for 3.4. The proof of 3.6 uses Lindström's theorem and the following consequence of Shepherdson-Smoryński theorem: if X is a Σ_1 set and $I\Sigma_1 \subseteq T_0 \subseteq T_1$ then there is a Σ_1 formula $\sigma(x)$ and Π_1 formula $\pi(x)$ such that both σ and π numerate X both in T_0 and in T_1.

§ 4. **Applications to interpretability.** It follows from 1.1 and 3.6 that if T is consistent sequential and has full induction then the set of all φ interpretable in T (i.e. such that the theory $(T+\varphi)$ is interpretable in T) is a complete Π_2 set. We focus our attention to finitely axiomatized theories containing $I\Sigma_1$. Note that if. T is finitely axiomatizable then the set $Intp_T$ of all φ interpretable in T is Σ_1.

4.1. **Theorem.** Let T be finitely axiomatized, $T \supseteq I\Sigma_1$ and let $\Gamma = \Sigma_n$ or $\Gamma = \Pi_n$ $(n \geq 1)$. Then Consv(Γ)-Int p_T is non-empty and contains a $\check{\Gamma}$-sentence.

4.2. **Remark.** In particular, let T be ACA_0 and $\Gamma = \Pi_1$. We get a Σ_1-formula φ which is Π_1-conservative over ACA_0 but $(ACA_0+\varphi)$ is not interpretable in ACA_0. Since φ is Π_1-conservative over ACA_0 it is Π_1-conservative over PA and thus $(PA+\varphi)$ is interpretable in PA. Compáre this with 2.4: we had there a Π_1 formula such that $(ACA_0+\varphi)$ is interpretable in ACA_0 but $(PA+\varphi)$ is not in PA. Similarly for GB and ZF.

The first example of a Σ_2 formula φ such that $ZF+\varphi$ is interpretable in ZF but $(GB+\varphi)$ is not in GB was constructed in Hájek [71] under the assumption of ω-consistency; this assumption was removed in Hájek and Hájková [72]. A Σ_1 formula of the desired properties was first constructed by Solovay.

4.3. We shall investigate the situation more closely. In the sequel let T be a finitely axiomatized consistent theory containing $I\Sigma_1$. We focus our attention to independent Σ_1 sentences. First observe that each independent (nonprovable and nonrefutable) Σ_1 sentence is false and trivially it is Σ_1-nonconservative. For each such Σ_1 sentence σ we may ask

- whether σ is Π_1-conservative,
- whether σ is interpretable,
- whether $\neg\sigma$ is interpretable.

The formula $\neg\sigma$ is Π_1 and hence Π_1-nonconservative; if T is Σ_1-sound then $\neg\sigma$ is Σ_1-conservative. For Σ_1-ill theories the Σ_1-conservativity of $\neg\sigma$ is a reasonable question but we shall not discuss it.

Our three questions admit eight combinations of answers, say, eight types of independent Σ_1 formulas.

4.4. **Theorem.** For each type, there is an independent Σ_1 formula of that type.

4.5. We shall describe examples in a rather uniform way. We shall use formulas HPr and HPr_{Σ_1}, i.e. herbrandian provability and herbrandian provability from a true Σ_1 sentence. Further we shall use a Σ_1 formula Intp(x) formally expressing (i.e. numerating-in-$I\Sigma_1$) interpretability of a formula in T. All examples have the form of a self-referential formula ξ such that

$$I\Sigma_1 \vdash \xi \equiv \Delta(\neg\overline{\xi}) \prec^{*} \nabla(\xi)$$

where $\Delta(x)$ has one of the following forms:

$HPr(x)$, $HPr_{\Sigma_1}(x)$, $HPr(x) \vee Intp(x)$, $HPr_{\Sigma_1}(x) \vee Intp(x)$.

$\nabla(x)$ has one of the following forms:

$HPr(x)$, $HPr(x) \vee Intp(x)$.

This gives eight possibilites that exactly give our eight examples. For several claims it is immaterial whether we use Pr or Hpr; but for some claims (using 1.7) it is not.

4.6. **Theorem.** (1) All eight examples are independent Σ_1 sentences.
(2) ξ is interpretable iff ∇ does not contain Intp(x).
(3) $\neg\xi$ is interpretable iff Δ does not contain Intp(x).
(4) ξ is Π_1-conservative iff Δ contains $HPr_{\Sigma_1}(x)$.

The proof uses 1.6, 1.7, provability of Herbrand's theorem in $I\Sigma_1$ and the following very important theorem (more precisely, its corollary 4.8):

4.7. **Second Lindström's fixed point theorem.** Let $T \supseteq I\Sigma_1$, $n,m \geq 0$, $\Gamma = \Sigma_n$, $\Lambda = \Pi_m$. Let $\varphi(x,y)$ be a Γ-formula, $\Theta(x,y)$ a Λ-formula.
Let

$$T \vdash \xi \equiv [(Pr_\Gamma(\neg\bar{\xi}) \vee (\exists y)\varphi(\neg\bar{\xi},y)) \prec^* (Pr_\Lambda(\bar{\xi}) \vee (\exists u)\Theta(\xi,u))].$$

Then
(1) for each m, $(T+\xi) \vdash (\exists y \leq \bar{m})\Theta(\bar{\xi},y) \to (\exists y \leq \bar{m})\varphi(\neg\xi,y)$,
(2) for each m, $(T+\neg\xi) \vdash (\exists y \leq \bar{m})\varphi(\neg\bar{\xi},y) \to (\exists y \leq \bar{m})\Theta(\xi,y)$,
(3) each $\check{\Gamma}$-sentence provable in $(T+\xi)$ is provable in $T+\{\neg\Theta(\bar{\xi},\bar{m}) \mid m\}$,
(4) each $\check{\Gamma}$-sentence provable in $(T+\neg\xi)$ is provable in $T+\{\neg\varphi(\bar{\xi},\bar{m}) \mid m\}$.
(5) The above remains true if Pr is replaced by HPr in all occurences.

4.8. **Corollary.** If $(\exists y)\Theta(x,y)$ defines a set $X \subseteq$ NRef (of non-refutable sentences) and $(\exists y)\varphi(x,y)$ defines a set Y of non-refutable sentences then ξ satisfies the following:

ξ is Π_n-conservative, $\neg\xi$ is Σ_n-conservative, $\xi \notin X$, $(\neg\xi) \notin Y$.

4.9. **Remark.** Note that Lindström has 4.7 and 4.8 for $T \supseteq PA$. The generalization of constructions of partially conservative sentences presented up to now (from $T \supseteq PA$ to $T \supseteq I\Sigma_1$) could make the reader think that everything generalizes smoothly. It is indeed remarkable that in most constructions the replacement of $\preccurlyeq$ by $\preccurlyeq^*$ works. But there are some things that do not generalize: for example, the implication

$$(\Delta \vee \nabla) \to ((\Delta \prec \nabla) \vee (\nabla \preccurlyeq \Delta))$$

(provable in PA) does not immediately generalize to $I\Sigma_1$ using *. Another example is in Lindström [79]: If T=PA and $X \subseteq$ NRef then there is a Δ_1 binumeration τ of PA such that $(PA+\neg Con_\tau)$ is interpretable in PA but $(ACA_0+\neg Con_\tau)$ is not interpretable in ACA_0. It is not clear how to generalize this from ACA_0 to any finitely axiomatizable $T \supseteq I\Sigma_1$.

Lindström's papers contain (for $T \supseteq PA$) various stronger and more detailed theorems on partial conservativity. It has been our task to demonstrate the possibility and ways of generalization to $T \supseteq I\Sigma_1$ on the most important theorems rather than to cover everything.

References

C. BENNET [86]: On some orderings of extensions of arithmetic (thesis), University of Göteborg 1986.

P. CLOTE [83]: Partition relations in arithmetic, Proc. Sixth Latin-American Symp. on Math. Logic, Venezuela 1983.

P. CLOTE [85]: Applications of the low-basis theorem in arithmetic, Proceedings of Recursion-theory week at Oberwolfach, Springer 1985.

S. FEFERMAN [60]: Arithmetization of metamathematics in a general setting, Fund. Math. 49(1960), 35-92.

D. GUASPARI [79]: Partially conservative extensions of arithmetic, Trans. Amer. Math. Soc. 254(1979), 47-68.

P. HÁJEK [71]: On interpretability in set theories, Comment. Math. Univ. Carolinae 12(1971), 73-79.

P. HÁJEK [72]: On interpretability in set theories II, Comment. Math. Univ. Carolinae 13(1972), 445-455.

P. HÁJEK [79]: On partially conservative extensions of arithmetic, in: Logic Colloquium 78, North-Holland Pub. (1979), 225-234.

P. HÁJEK [81]: On interpretability in theories containing arithmetic II, Comment. Math. Univ. Carolinae 22(1981), 617-688.

P. HÁJEK [84]: On a new notion of partial conservativity, in: Logic Colloquium 83, Lect. Notes in Math. vol. 1104.

P. HÁJEK, M. HÁJKOVÁ [72]: On interpretability in theories containing arithmetic, Fund. Math. 76(1972), 131-137.

P. HÁJEK, A. KUČERA [∞] : On recursion theory in fragments of arithmetic (to appear).

G. KREISEL [62]: On weak completeness of intuitionistic predicate logic, J. Symb. Logic 27(1962), 139-158.

G. KREISEL [68]: A survey on proof theory, J. Symb. Logic 33(1968), 321-388.

P. LINDSTRÖM [79]: Some results on interpretability (Jensen, Mayoh, Motler, ed.), Proc. 5th Scandinavian Logic Symp., Aalborg Univ. Press 1979.

P. LINDSTRÖM [84]: On partially conservative sentences and interpretability, Proceedings Amer. Math. Soc. 91(1984), 436-443.

P. LINDSTRÖM [84a]: On faithful interpretability, in: Logic Colloquium 83, Lect. Notes in Math. vol. 1104, Springer-Verlag 1984.

S. OREY [61]: Relative interpretations, Z. Math. Logik und Grund. Math. 7 (1961), 146-153.

J. QUINSEY [81]: Sets of Σ_k-conservative sentences are Π_2-complete, J. Symb Logic 46(1981), 442 (abstract.

P. PUDLÁK [85]: Cuts, consistency statements and interpretations, J. Symb. Logic 50(1985), 423-441.

J.S. SHEPERDSON [60]: Representability of recursively enumerable sets in formal theories, Archiv f. math. Logik 5(1960), 119-127.

C. SMORYŃSKI [81]: Fifty years of self-reference, Notre Dame Journal of Formal Logi 22(1981), 357-374.

C. SMORYŃSKI [81a]: Calculating self-referential sentences: Guaspari sentences of the first kind, J. Symb. Logic 46(1981), 329-344.

C. SMORYŃSKI [85]: Self-reference and modal logic, Springer-Verlag 1985.

V. ŠVEJDAR [81]: A sentence that is difficult to interpret, Comment. Math. Univ. Carolinae 22(1981), 661-666.

V. ŠVEJDAR [83]: Modal analysis of generalized Rosser sentences, J. Symb. Logic 48(1983), 986-999.

A. TARSKI, A. MOSTOWSKI, R.M. ROBINSON [53]: Undecidable theories, North -Holland P.C. 1953.

Mathematical Institute of the Czechoslovak Academy of Sciences, Žitná 25,
115 67 Praha 1, Czechoslovakia

(Oblatum 20.7. 1987)