6. INTERPRETABILITY

Let S and S′ be arbitrary theories. S′ is interpretable in S if, roughly speaking, the primitive concepts and the range of the variables of S′ are definable in S in such a way as to turn every theorem of S′ into a theorem of S. If, in addition every nontheorem of S′ is transformed into a nontheorem of S, then S′ is faithfully interpretable in S.

In this chapter, we assume that PA$\dashv$ T. Thus, T is essentially reflexive.

§1. Interpretability. Let S and S′ be arbitrary theories. By a *translation* (of the language of S′ into the language of S) we understand a function t on the set of formulas (of S′) into the set of formulas (of S) for which there are formulas $\eta_0(x)$, $\eta_S(x,y)$, $\eta_+(x,y,z)$, $\eta_\times(x,y,z)$ and a formula $\mu_t(x)$ such that t satisfies the following conditions for all formulas φ, ψ, $\xi(x)$:

(*) $t(x = y) := x = y,$

$t(x = 0) := \eta_0(x),$

$t(Sx = y) := \eta_S(x,y),$

$t(x + y = z) := \eta_+(x,y,z),$

$t(x \times y = z) := \eta_\times(x,y,z),$

$t(\neg\varphi) := \neg t(\varphi),$

$t(\varphi \wedge \psi) := t(\varphi) \wedge t(\psi),$

$t(\exists x\xi(x)) := \exists x(\mu_t(x) \wedge t(\xi(x))).$

(Here x, y, z are arbitrary variables.) We assume that $\forall$ and the connectives $\vee$, $\rightarrow$, $\leftrightarrow$ are defined in terms of $\exists$, $\neg$, $\wedge$. Note that t, on the formulas for which it is defined by the above conditions, is uniquely determined by its values on atomic formulas together with the formula $\mu_t(x)$.

So far $t(\varphi)$ is only defined provided that φ is written in a certain "normal form". For example, t is not defined on the formula $x + 0 = y$. But this formula is equivalent to $\exists z(z = 0 \wedge x + z = y)$ and t is defined on this formula so we can set $t(x + 0 = y) := t(\exists z(z = 0 \wedge x + z = y))$. Similarly, for any formula φ not already on "normal form", replace φ in some canonical way by φ^* on "normal form" (logically equivalent to φ) and set $t(\varphi) := t(\varphi^*)$. It follows, for example, that $t(\forall x\xi(x))$ is equivalent to $\forall x(\delta(x) \rightarrow t(\xi(x)))$. Clearly t is a primitive recursive function.

The translation t is an *interpretation in* S iff

(**) $S \vdash \exists x\mu_t(x),$

$S \vdash \exists x(\mu_t(x) \wedge \forall y(\mu_t(y) \rightarrow (\eta_0(y) \leftrightarrow y = x))),$

$S \vdash \forall x\big(\mu_t(x) \rightarrow \exists y(\mu_t(y) \wedge \forall z(\mu_t(z) \rightarrow (\eta_S(x,z) \leftrightarrow z = y)))\big),$

$S \vdash \forall xy\big(\mu_t(x) \wedge \mu_t(y) \rightarrow \exists z\big(\mu_t(z) \wedge \forall u(\mu_t(u) \rightarrow (\eta_*(x,y,u) \leftrightarrow u = z))\big)\big),\ * = +, \times.$

Thus, t is an interpretation in S iff $S \vdash t(\varphi)$ for every logically valid sentence φ.

t is an *interpretation of* S′ *in* S, t: S′$\leq$ S, iff $S \vdash t(\varphi)$ for every φ such that $S' \vdash \varphi$. S′

is *interpretable in* S, $S' \leq S$, if there is an interpretation of S' in S. $S' < S$ means that $S' \leq S \nleq S'$.

Trivially, if $S' \dashv S$, then $S' \leq S$. The reader should check that $\leq$ is a transitive relation. Also note that if $S' \leq S$, then every finite subtheory of S' is interpretable in a finite subtheory of S.

If $S' \leq S$ and S is consistent, so is S'. For suppose S' is not consistent. Let φ be any sentence. Then $S' \vdash \varphi \wedge \neg\varphi$. But then $S \vdash t(\varphi \wedge \neg\varphi)$. But $t(\varphi \wedge \neg\varphi) := t(\varphi) \wedge \neg t(\varphi)$, whence $S \vdash t(\varphi) \wedge \neg t(\varphi)$ and so S is inconsistent.

Since every translation t is a primitive recursive function, we may in (extensions of) PA use t as a function symbol. t can always be defined such that the following Fact holds and the argument in the preceding paragraph can be formalized in PA.

Fact 12. Suppose t: $S' \leq S$.

(a) The conditions (*) and (**) are provable in PA.

(b) $PA \vdash Pr_{\emptyset}(x) \rightarrow Pr_S(t(x))$.

This Fact has the following:

Corollary 1. Suppose t: $S' \leq S$ and S' is finite. Then $PA \vdash Pr_{S'}(x) \rightarrow Pr_S(t(x))$ and consequently $PA \vdash Con_S \rightarrow Con_{S'}$.

The assumption that S' is finite in Corollary 1 cannot be omitted: $S' \leq S$ may be true but not provable in PA (see Corollary 5 and Theorem 12, below). But we do have the following weaker result. (Recall that a *numeration* of a set X numerates X in PA.)

Theorem 1. Suppose $S_0 \leq S_1$ and let $\sigma_1(x)$ be a Σ_1 numeration of S_1. There is then a Σ_1 numeration $\sigma_0(x)$ of S_0 such that

$$PA \vdash Con_{\sigma_1} \rightarrow Con_{\sigma_0}.$$

Proof. Suppose t: $S_0 \leq S_1$. Let $\sigma(x)$ be a PR binumeration of S_0 and let $\sigma_0(x) := \sigma(x) \wedge Pr_{\sigma_1}(t(x))$. Then $\sigma_0(x)$ is a Σ_1 numeration of S_0 and

(1) $PA \vdash Pr_{\sigma_0}(x) \rightarrow Pr_{\sigma_1}(t(x))$.

To prove this, we reason (informally) in PA as follows: "Suppose φ is derivable from formulas satsifying $\sigma_0(x)$. Then there are $\psi_0,\dots,\psi_n$ of formulas satisfying $\sigma_0(x)$, such that $\bigwedge\{\psi_k: k \leq n\} \rightarrow \varphi$ is provable in logic. But then, by Fact 12 (this chapter), $t(\bigwedge\{\psi_k: k \leq n\}) \rightarrow t(\varphi)$ is provable from the set defined by $\sigma_1(x)$. But $t(\bigwedge\{\psi_k: k \leq n\}) := \bigwedge\{t(\psi_k): k \leq n\})$. Also, by the definition of $\sigma_0(x)$, each $t(\psi_k)$ is derivable from the set defined by $\sigma_1(x)$. But then so is $\bigwedge\{t(\psi_k): k \leq n\})$. It follows that $t(\varphi)$ is derivable from the set defined by $\sigma_1(x)$." This proves (1).

From (1) we easily get the desired conclusion. ■

Theorem 1 in combination with Gödel's second incompleteness theorem (Theorem 2.4) yields the following strengthening of Gödel's result. For a different improvement of Theorem 2.4, see Theorem 8, below.

Theorem 2. $T + Con_T \not\leq T$.

Proof. Suppose $T + Con_T \leq T$. Then, by Theorem 1, there is a Σ_1 numeration $\tau'(x)$ of $T + Con_T$ such that $T + Con_T \vdash Con_{\tau'}$. By Theorem 2.4 it now follows that $T + Con_T$ is inconsistent. But then, since $T + Con_T \leq T$, T is inconsistent, contrary to Convention 2. ■

Since Con_T is Π_1, Theorem 2 is also a direct consequence of Theorem 2.4 and the following:

Lemma 1. If π is a Π_1 sentence and $Q + \pi \leq T$, then $T \vdash \pi$.

Proof. There is a k such that $Q + \pi \leq T \restriction k$. So, by Corollary 1, $T \vdash Con_{T \restriction k} \to Con_{Q+\pi}$. It follows that $T \vdash Con_{Q+\pi}$. Since $\neg\pi$ is Σ_1, we have, by provable Σ_1–completeness, $T \vdash \neg\pi \to \neg Con_{Q+\pi}$. It follows that $T \vdash \pi$. ■

Note that we have actually proved that $Q + Con_T \not\leq T$.

In Chapter 2 (Corollary 2.1) we proved that PA is essentially infinite (in fact, PA is essentially unbounded; Corollary 4.1). This can now be improved as follows:

Theorem 3. T is not interpretable in any finite subtheory of T.

Proof. Let S be a finite subtheory of T and suppose $T \leq S$. By Theorem 1, there is then a Σ_1 numeration $\tau(x)$ of T such that $PA \vdash Con_S \to Con_\tau$. Since, by Fact 11, T is reflexive, we have $T \vdash Con_S$ and so $T \vdash Con_\tau$, contradicting Theorem 2.4. ■

Most positive results on the existence of interpretations in the sequel are applications of the following fundamental result, the arithmetization of Gödel's completeness theorem.

Theorem 4. Let $\sigma(x)$ be a formula numerating S in T. Then $S \leq T + Con_\sigma$.

Proof (informal outline). A full proof of this result would be quite long and we shall be content to give a fairly detailed sketch. The main idea is to show that (the denumerable case of) the Henkin completeness proof for first order logic can be formalized in PA. (The reader is assumed to be familiar with that proof.)

We begin with an outline of Henkin's proof. Let S be a (countable) set of sentences (theory) assumed to be consistent. Let c_n, $n \in N$, be new individual constants. Let L be the language obtained from L_S by adding the constants c_n. Let $\alpha_n(x_n)$, $n \in N$, be a primitive recursive enumeration of all formulas of L with one free variable. We can then form a primitive recursive set

$$Z = \{\exists x_n \alpha_n(x_n) \to \alpha_n(c_{j_n}) \colon n \in N\}$$

such that

(1) for every sentence θ of S, if $S + Z \vdash \theta$, then $S \vdash \theta$.
It follows that $S + Z$ is consistent.

Now let θ_n, $n \in N$, be a primitive recursive enumeration of all sentences of L. The sentences φ_n are then inductively defined as follows:

(2) $\varphi_n = \theta_n$ if $S + Z \vdash \bigwedge\{\varphi_m\colon m < n\} \to \theta_n$,
$= \neg\theta_n$ otherwise.

(Here $\bigwedge\{\varphi_m\colon m < 0\} := 0 = 0$.) φ_n is not in general a recursive function of n.

Let $X = \{\varphi_n\colon n \in N\}$. Then

(3) $Th(S) \subseteq X$

and, since $S + Z$ is consistent,

(4) X is *Henkin complete*

in the sense that X is complete and consistent and for every formula $\alpha(x)$ of L with the one free variable x, if $\exists x\alpha(x) \in X$, there is a constant c_k such that $\alpha(c_k) \in X$.

We can now define a model

$\mathbf{M} = (M, S^{\mathbf{M}}, +^{\mathbf{M}}, \times^{\mathbf{M}}, 0^{\mathbf{M}})$

of X in the following way. The domain M of the model is the set $\{c_n\colon n \in N\}$. (Here we ignore the minor difficulty that X may contain sentences of the form $c_k = c_m$ with $k \neq m$ and so the members of M cannot in general be the constants themselves but must instead be certain "equivalence classes" of these constants or, in the present context, members of such equivalence classes. If we disregard the trivial case where S has only finite models, this can be avoided by defining Z in a slightly different way.)

$0^{\mathbf{M}} = c_{i_0}$, $c_n^{\mathbf{M}} = c_n$,
$S^{\mathbf{M}} = \{(c_k, c_m)\colon Sc_k = c_m \in X\}$,
$+^{\mathbf{M}} = \{(c_k, c_m, c_n)\colon c_k + c_m = c_n \in X\}$,
$\times^{\mathbf{M}} = \{(c_k, c_m, c_n)\colon c_k \times c_m = c_n \in X\}$,

where c_{i_0} is the (uniquely determined) constant such that $0 = c_{i_0} \in X$.

Finally, it can be shown, by induction and using the fact that X is Henkin complete, that for every sentence φ of L,

(5) φ is true in $\mathbf{M}$ iff $\varphi \in X$.

This is true, by the definition of $\mathbf{M}$, if φ is atomic.

Finally, $Th(S) \subseteq X$ and so $\mathbf{M}$ is a model of Th(S).

We can now transform this into a proof that $S \leq T + Con_\sigma$ in the following way. We first define in PA a primitive recursive function c(x) (= the x^{th} new individual constant). By a c–formula we understand a formula obtained from a formula of L_A by replacing each free variable v by $c(\dot{v})$. (Thus, the c–formulas are the counterparts of the sentences of L.) Let $\zeta(x)$ be a suitably defined PR binumeration of Z, where Z is defined as above except that we now use the function symbol c. Then (the reader will hopefully believe that) for every sentence φ of S,

(6) $PA \vdash Pr_{\sigma\vee\zeta}(\varphi) \to Pr_\sigma(\varphi)$.

(compare (1)). It follows that

(7) $PA \vdash Con_\sigma \to Con_{\sigma\vee\zeta}$.

The inductive definition of φ_n can, using methods available in PA, be turned into an explicit definition. Let $\chi(x,y)$ be a suitable formalization of this explicit definition (cf. Chapter 1, p. 9). Let $\xi(x) := \exists y\chi(x,y)$. (Thus, intuitively, $\xi(x)$ means "x is a member of X".) Then (compare (3))

(8) $\quad PA\vdash Pr_\sigma(x) \to \xi(x)$.

Let Hcm_ξ be the sentence saying that the set defined by $\xi(x)$ is Henkin complete. Thus, for all c–formulas α, β,

(9) $\quad PA + Hcm_\xi\vdash \xi(\neg\alpha) \leftrightarrow \neg\xi(\alpha)$.

(10) $\quad PA + Hcm_\xi\vdash \xi(\alpha) \wedge Pr_\varnothing(\alpha\to\beta) \to \xi(\beta)$.

Moreover, for every formula $\alpha(x)$ such that $\exists x\alpha(x)$ is a c–formula,

(11) $\quad PA + Hcm_\xi\vdash \xi(\exists x\alpha(x)) \to \exists u\xi(\alpha(c(\dot{u})))$.

The (inductive) proof of (4) does not use any means of proof beyond those available in PA. Thus, we get $PA\vdash Con_{\sigma\vee\zeta} \to Hcm_\xi$ and so, by (7),

(12) $\quad PA\vdash Con_\sigma \to Hcm_\xi$.

We can now define a translation t, corresponding to the model **M**, as follows. Let

$$\begin{aligned}
&\mu_t(x) := \exists u(x = c(u)),\\
&t(x = 0) := \exists u(x = c(u) \wedge \xi(0 = c(\dot{u}))),\\
&t(Sx = y) := \exists uv(x = c(u) \wedge y = c(v) \wedge \xi(Sc(\dot{u}) = c(\dot{v}))),\\
&t(x + y = z) := \exists uvw(x = c(u) \wedge y = c(v) \wedge z = c(w) \wedge \xi(c(\dot{u}) + c(\dot{v}) = c(\dot{w}))),\\
&t(x \times y = z) := \exists uvw(x = c(u) \wedge y = c(v) \wedge z = c(w) \wedge \xi(c(\dot{u}) \times c(\dot{v}) = c(\dot{w}))).
\end{aligned}$$

These equations uniquely determine t.

The proof corresponding to the proof of (5) now yields for every formula $\beta(x_0,\ldots,x_{n-1})$ of L_A containing no free variables other than $x_0,\ldots,x_{n-1}$,

(13) $\quad PA + Hcm_\xi\vdash \mu_t(x_0) \wedge\ldots\wedge \mu_t(x_{n-1}) \to (t(\beta(x_0,\ldots,x_{n-1})) \leftrightarrow$
$$\exists u_0,\ldots,u_{n-1}(x_0 = c(u_0) \wedge\ldots\wedge x_{n-1} = c(u_{n-1}) \wedge \xi(\beta(c(\dot{u})_0,\ldots,c(\dot{u})_{n-1})))).$$

By the definition of t, this holds for atomic $\beta(x_0,\ldots,x_{n-1})$. The inductive steps dealing with $\neg$ and $\wedge$ follow easily, by (9) and (10).

Let us consider the step dealing with $\exists$. For simplicity, let $n = 1$ and write x for x_0. Let $\alpha(x,y)$ be such that $\beta(x) := \exists y\alpha(x,y)$. Then $t(\beta(x)) := \exists y(\mu_t(y) \wedge t(\alpha(x,y))$. By the inductive hypothesis,

$$PA + Hcm_\xi\vdash \mu_t(x) \wedge \mu_t(y) \to (t(\alpha(x,y)) \leftrightarrow \exists uv(x = c(u) \wedge y = c(v) \wedge \xi(\alpha(c(\dot{u}),c(\dot{v}))))).$$

By (10) and (11),

$$PA + Hcm_\xi\vdash \exists v\xi(\alpha(c(\dot{u}),c(\dot{v}))) \leftrightarrow \xi(\exists y\alpha(c(\dot{u}),y)).$$

But then it is fairly easy to see that

$$PA + Hcm_\xi\vdash \mu_t(x) \to (\exists y(\mu_t(y) \wedge t(\alpha(x,y)) \leftrightarrow \exists u(x = c(u) \wedge \xi(\exists y\alpha(c(\dot{u}),y)))),$$

as desired. This proves (13).

From (12) and (13), we get for every sentence φ,

(14) $\quad PA + Con_\sigma\vdash t(\varphi) \leftrightarrow \xi(\varphi)$.

Finally, let φ be any sentence provable in S. Then $T\vdash Pr_\sigma(\varphi)$. Hence, by (8), $T\vdash \xi(\varphi)$ and so, by (14), $T + Con_\sigma\vdash t(\varphi)$. It follows that $t: S \le T + Con_\sigma$.

This concludes our sketch of the proof of Theorem 4. ■

If we don't insist on mimicking every detail of Henkin's proof, we can instead use the simpler interpretation t′ defined in the following way:

$$\mu_{t'}(x) := x = x,$$
$$t'(x = 0) := \xi(0 = c(\dot{x})),$$
$$t'(Sx = y) := \xi(Sc(\dot{x}) = c(\dot{y})),$$
$$t'(x + y = z) := \xi(c(\dot{x}) + c(\dot{y}) = c(\dot{z})),$$
$$t'(x \times y = z) := \xi(c(\dot{x}) \times c(\dot{y}) = c(\dot{z})),$$

Thus, $\mu_{t'}(x)$ is trivial and can be omitted. (This is true as long as we are dealing with theories of (elementary) arithmetic; it is *not* true in general.)

It is via the following lemma, the (Feferman–)Orey–Hájek Lemma (and Theorem 6, below) that Theorem 4 becomes such a powerful tool in the theory of interpretability (of arithmetical theories; see also Lemma 8.4).

Lemma 2. $S \leq T$ iff $T \vdash \mathrm{Con}_{S \restriction k}$ for every k.

To prove this we need the following lemma whose proof is essentially the same as that of Theorem 2.7.

Lemma 3. Suppose $T \vdash \mathrm{Con}_{S \restriction k}$ for every k. Let $\sigma(x)$ be any formula binumerating S in T and let

$$\sigma^*(x) := \sigma(x) \wedge \mathrm{Con}_{\sigma|x}.$$

Then (i) $\sigma^*(x)$ binumerates S in T and (ii) $PA \vdash \mathrm{Con}_{\sigma^*}$.

Proof of Lemma 2. Suppose first $S \leq T$. Let k be arbitrary. There is then an m such that $S \restriction k \leq T \restriction m$. By Corollary 1, $PA \vdash \mathrm{Con}_{T \restriction m} \to \mathrm{Con}_{S \restriction k}$. But $T \vdash \mathrm{Con}_{T \restriction m}$ and so $T \vdash \mathrm{Con}_{S \restriction k}$.

Next suppose $T \vdash \mathrm{Con}_{S \restriction k}$ for every k. Let $\sigma(x)$ be a PR binumeration of S and let $\sigma^*(x) := \sigma(x) \wedge \mathrm{Con}_{\sigma|x}$. Then, by Lemma 3, $\sigma^*(x)$ binumerates S in T and $PA \vdash \mathrm{Con}_{\sigma^*}$. Hence, by Theorem 4, $S \leq T$. ■

There are alternative notions of interpretability more general than the one defined here. For example, we may "interpret" the equality symbol = of one theory S as a certain relation definable in another S′ (and having, provably in S′, the required properties) or we may "interpret" the individuals of S as finite sequences of individuals of S′ etc. It turns out, however, that if S is "interpretable" in T in some such more general, and reasonably natural, sense, then, by Lemma 2, $S \leq T$ (and conversely). Thus, in the present context, there is no reason to consider these more general "interpretations".

From Lemmas 2 and 3 and Theorem 4 we get the following:

Corollary 2. $S \leq T$ iff there is a formula $\sigma(x)$ (bi)numerating S in T such that $T \vdash \mathrm{Con}_{\sigma}$.

From Lemma 2 we also obtain the following result known as Orey's compactness theorem.

Theorem 5. $S \leq T$ iff $S \restriction k \leq T$ for every k.

In the following we use A, B, etc. to denote (consistent, primitive recursive) extensions of T. Recall that $A \dashv_{\Gamma} B$ means that every Γ sentence provable in A is provable in B.

Theorem 6. $A \leq B$ iff $A \dashv_{\Pi_1} B$.

Proof. Suppose first $A \dashv_{\Pi_1} B$. Now, $A \vdash Con_{A \restriction k}$ for every k. It follows that $B \vdash Con_{A \restriction k}$ for every k. But then, by Lemma 2, $A \leq B$.

Suppose next $A \leq B$. Let π be any Π_1 sentence such that $A \vdash \pi$. By Lemma 1, $B \vdash \pi$, as desired. ■

By Theorem 6, $A + \varphi \leq A$ iff φ is Π_1–conservative over A.

Theorem 6 has the following immediate:

Corollary 3. If $A \leq B$ and σ is any Σ_1 sentence, then $A + \sigma \leq B + \sigma$.

Combining Theorem 6 and Theorem 4.5 we get:

Corollary 4. $T + Rfn_T \leq PA + Con^{\omega}_T$.

In fact, this follows directly from Lemma 2 and the fact, established in the proof of Theorem 4.5, that $PA + Con^{\omega}_T \vdash Con_{T_n}$ for every n.

Theorem 6 can also be used to prove the following model–theoretic characterization of interpretability:

Theorem 7. $A \leq B$ iff for every model **M** of B, there is a model **M′** of A such that **M** is (isomorphic to) an initial segment of **M′**.

Proof (sketch). "If". Let θ be any Π_1 sentence such that $A \vdash \theta$. We show that θ holds in all models of B. Let **M** be any model of B. By hypothesis, there is a model **M′** of A such that **M** is isomorphic to an initial segment of **M′**. θ holds in **M′**. Since θ is Π_1, it follows that θ holds in **M**. Thus, θ holds in all models of B and so $B \vdash \theta$. We have shown that $A \dashv_{\Pi_1} B$ and so $A \leq B$, by Theorem 6.

"Only if". Let t: $A \leq B$. Let **M** be any model of B and let **M′** be the structure defined by t in **M**. **M′** is a model of A. Since induction holds in **M**, we can in **M** define a function f on M satisfying the following conditions: $f(0^{\mathbf{M}}) = 0^{\mathbf{M}'}$, $f(S^{\mathbf{M}}(a)) = S^{\mathbf{M}'}(f(a))$. f maps **M** isomorphically onto an initial segment of **M′**. ■

Given Theorem 6, we can now derive Theorems 8 – 12 below as corollaries to results from Chapter 5.

Like Theorem 2 the following result is a sharpening of Gödel's second incompleteness theorem.

Theorem 8. $T + \neg Con_T \leq T$.

Proof. This follows from Theorem 5.1 and Theorem 6. ■

A more direct proof of Theorem 8 is as follows. We need the following:

Lemma 4. If $S \leq S' + \varphi_0$ and $S \leq S' + \varphi_1$, then $S \leq S' + \varphi_0 \vee \varphi_1$. Thus, if $S + \varphi \leq S + \neg\varphi$, then $S + \varphi \leq S$.

Proof. Suppose t_i: $S \leq S' + \varphi_i$, $i = 0, 1$. Let t be the translation which coincides with t_0 if φ_0 and with t_1 if $\neg\varphi_0 \wedge \varphi_1$. Thus, for example,

$$\mu_t(x) := (\varphi_0 \wedge \mu_{t_0}(x)) \vee (\neg\varphi_0 \wedge \varphi_1 \wedge \mu_{t_1}(x)).$$

It follows that for all φ,

(1) $S' + \varphi_0 \vdash t(\varphi) \leftrightarrow t_0(\varphi)$,

(2) $S' + \neg\varphi_0 \wedge \varphi_1 \vdash t(\varphi) \leftrightarrow t_1(\varphi)$.

Now, suppose $S \vdash \varphi$. Then $S' + \varphi_0 \vdash t_0(\varphi)$ and so, by (1), $S' + \varphi_0 \vdash t(\varphi)$. Also $S' + \varphi_1 \vdash t_1(\varphi)$ and so, by (2), $S' + \neg\varphi_0 \wedge \varphi_1 \vdash t(\varphi)$. It follows that $S' + \varphi_0 \vee \varphi_1 \vdash t(\varphi)$. Thus, t: $S \leq S' + \varphi_0 \vee \varphi_1$, as desired. ■

By Corollary 2.2, $T + Con_T \vdash Con_{T+\neg Con T}$. But then, by Theorem 4, $T + \neg Con_T \leq T + Con_T$ and so, by Lemma 4, $T + \neg Con_T \leq T$, as desired. (In this proof of Theorem 8 it is not necessary to assume that T is (essentially) reflexive.)

Theorem 9. Suppose X is r.e. and monoconsistent with T. There is then a Σ_1 sentence φ such that $T + \varphi \leq T$ and $\varphi \notin X$.

Proof. This follows from Theorem 5.2 and Theorem 6. ■

Corollary 5. Let $\tau(x)$ be a formula numerating T in T such that $T \nvdash \neg Con_\tau$. There is then a (Σ_1) sentence φ such that $T + \varphi \leq T$ and $T \nvdash Con_\tau \rightarrow Con_{\tau+\varphi}$.

Proof. Let $X = \{\psi: T \vdash Con_\tau \rightarrow Con_{\tau+\psi}\}$ and use Theorem 9. ■

Theorem 10. Suppose X is r.e. and monoconsistent with T. There is then a sentence φ such that $T + \varphi \leq T$, $T + \neg\varphi \leq T$, $\varphi \notin X$, $\neg\varphi \notin X$.

Proof. By Theorem 5.3 we can take φ to be, say, a Σ_2 sentence such that φ is Π_2–conservative and $\neg\varphi$ is Σ_2–conservative over T. Now use Theorem 6. ■

A sentence φ such that $T + \varphi \leq T$, $T + \neg\varphi \leq T$ is known as an *Orey sentence* for T. Clearly, any Orey sentence for T is undecidable in T.

The intended applications of Theorems 9 and 10 are as follows. There are consistent finitely axiomatized extensions U of T in languages extending L_A. In fact, U may chosen to be a conservative extension of T in the sense that for every sentence φ of L_A, $U \vdash \varphi$ iff $T \vdash \varphi$. Thus, U and T are equivalent in terms of provability of sentences of L_A. So it is natural to ask if U and T are (ever) equivalent in terms of interpretability of sentences of L_A in the sense that for every sentence φ of L_A, $T + \varphi \leq T$ iff $U + \varphi \leq U$. (We assume the reader can extend the defintion of "interpretation" and "interpretable in" to the case where the theories need not be formalized in L_A.) The answer is a resounding "no" (see also Corollary 8.8). To prove this we need the following essentially trivial lemma whose proof is left to the reader.

Lemma 5. Let V be any r.e. theory, not necessarily in L_A. Then the set $\{\varphi: U + \varphi \leq V\}$ is r.e.

Corollary 6. There is a Σ_1 sentence φ such that $T + \varphi \leq T$ and $U + \varphi \not\leq U$.

Proof. The set $\{\varphi: U + \varphi \leq U\}$ is clearly monoconsistent with T and, by Lemma 5, it is r.e. Now apply Theorem 9. ■

By a similar proof, but using Theorem 10 in place of Theorem 9, we get:

Corollary 7. There is a sentence φ such that $T + \varphi \leq T$, $T + \neg\varphi \leq T$, $U + \varphi \not\leq U$, and $U + \neg\varphi \not\leq U$.

As we saw in Chapter 4, speaking in terms of provability, we have to distinguish between finite, infinite, and unbounded extensions of a given theory T. In terms of interpretability the situation is quite different. We write $S \equiv S'$ to mean that $S \leq S' \leq S$.

Theorem 11. (a) If $A \dashv B$, then there is a sentence φ such that $A + \varphi \equiv B$.

(b) Let X be an r.e. set of Σ_1 sentences. Then there is a Σ_1 sentence σ such that $T + \sigma \equiv T + X$.

Proof. (a) Let $X = Th(B) \cap \Pi_1$. Then, by Theorem 6, $A + X \equiv B$. By Theorem 5.4 (a), there is a sentence φ such that $A + \varphi$ is a Π_1– conservative extension of $A + X$. By Theorem 6, $A + \varphi \equiv A + X$ and so $A + \varphi \equiv B$. ♦

(b) This follows from Theorems 5.4 (a) and 6. ■

Finally, we have a result which proves the claim made earlier that the fact that, for example, $A + \varphi \leq B$ does not imply that this is provable in PA, or in any other preassigned consistent axiomatizable theory.

From the definition of $\leq$ it is clear that the set $\{\varphi: A + \varphi \leq B\}$ is Σ^0_3. From Theorem 6, it follows, however, that $\{\varphi: A + \varphi \leq B\}$ is Π^0_2. That this cannot be improved follows from:

Theorem 12. Suppose $A \leq B$. Then the set $\{\varphi \in \Sigma_1: A + \varphi \leq B\}$ is a complete Π^0_2 set.

Proof. For $A = B$, this follows from Theorem 5.6 and Theorem 6; we leave the proof of the general case to the reader. ■

A translation t is given by a finite amount of information which can certainly be coded by a natural number; thus we may "identify" t with that number. Let $Int_{A,B}$ be the set of interpretations of A in B.

Corollary 8. If $A \leq B$, then $Int_{A,B}$ is Π^0_2 but not Σ^0_2.

Proof. Clearly $Int_{A,B}$ is Π^0_2. Suppose it is Σ^0_2. Evidently

$$A + \varphi \leq B \text{ iff } \exists t \in Int_{A,B}(B \vdash t(\varphi)).$$

It follows that $\{\varphi: A + \varphi \leq B\}$ is Σ^0_2, contradicting Theorem 12. ■

In the next § we are going to prove that $Int_{A,B}$ is, in fact, a complete Π^0_2 set (Corollary 12).

§2. Faithful interpretability. Let $t: S' \leq S$. t is a *faithful* interpretation of S' in S, $t: S' \trianglelefteq S$, if for every sentence φ, if $S \vdash t(\varphi)$, then $S' \vdash \varphi$. S' is *faithfully interpretable* in S, $S' \trianglelefteq S$, if there is a t such that $t: S' \trianglelefteq S$.

Most of the differences between $\leq$ and $\trianglelefteq$ are explained by the following lemma; for example, it is not true in general that if $S \dashv T$, then $S \trianglelefteq T$.

Lemma 6. If $Q \dashv S \trianglelefteq T$, then $T \dashv_{\Sigma_1} S$.

Proof. Suppose $t: S \trianglelefteq T$. Let σ be any Σ_1 sentence such that $T \vdash \sigma$. Clearly $t: Q + \neg\sigma \leq T + \neg t(\sigma)$. But then, by Lemma 1, $T + \neg t(\sigma) \vdash \neg\sigma$, and so $T \vdash t(\sigma)$. Since t is faithful, it follows that $S \vdash \sigma$. ■

Our main aim in this § is to prove the following characterizations of $\trianglelefteq$.

Theorem 13. $S \trianglelefteq T$ iff $S \leq T$ and for every φ, if $T \vdash Pr_{\varnothing}(\varphi)$, then $S \vdash \varphi$.

Theorem 14. $A \trianglelefteq B$ iff $A \dashv_{\Pi_1} B \dashv_{\Sigma_1} A$.

Corollary 9. (a) $S \trianglelefteq T$ iff for every k, $T \vdash Con_{S \restriction k}$ and for every φ, if $T \vdash Pr_{\varnothing}(\varphi)$, then $S \vdash \varphi$.

(b) If T is Σ_1–sound, then $S \trianglelefteq T$ iff $S \leq T$.

(c) If $S \leq T \dashv S$, then $S \trianglelefteq T$.

Proof. (a) and (b) follow at once from Theorem 13 and Lemma 2. ♦

(c) Suppose $T \vdash Pr_{\varnothing}(\varphi)$. Then, since T is essentially reflexive (Fact 11), $T \vdash \varphi$ and so, by assumption, $S \vdash \varphi$. Now use Theorem 13. ■

By Corollary 9 (c), Theorems 8, 9, 10 remain true when $\leq$ is replaced by $\trianglelefteq$.

Theorem 13 will be derived from the following two lemmas:

Lemma 7. Let $\sigma'(x)$ be a (Σ_1) formula binumerating S in T. There is then a (Σ_1) formula $\sigma(x)$ binumerating S in T and such that

(i) $\vdash \sigma(x) \rightarrow \sigma'(x)$, whence $\vdash Con_{\sigma'} \rightarrow Con_{\sigma}$,

(ii) for every sentence φ, if $T\vdash Pr_{\sigma}(\varphi)$, then there is a q such that $T\vdash Pr_{S\restriction q}(\varphi)$.

Lemma 8. Suppose $\sigma(x)$ numerates S in T and $T\vdash Con_{\sigma}$. There is then an interpretation t: $S \leq T$ such that for every φ, if $T\vdash t(\varphi)$, then $T\vdash Pr_{\sigma}(\varphi)$.

Proof of Lemma 7. For simplicity we assume, as we clearly may, that if p is a proof of φ in T, then $\varphi \leq p$. Let $\sigma(x)$ be such that

$$PA\vdash \sigma(x) \leftrightarrow \sigma'(x) \wedge \forall yz{\leq}x(Prf_T(Pr_{\sigma}(\dot{y}),z) \rightarrow Pr_{\sigma'\restriction z}(y)).$$

Then (i) is trivial.

We now show that

(1) if p is a proof of $Pr_{\sigma}(\varphi)$ in T, then $T\vdash Pr_{\sigma'\restriction p}(\varphi)$.

Let p and φ be as assumed. Then, since $\varphi \leq p$,

$$T\vdash \neg Pr_{\sigma'\restriction p}(\varphi) \rightarrow (\sigma(x) \rightarrow \sigma'(x) \wedge x \leq p).$$

It follows that

$$T\vdash \neg Pr_{\sigma'\restriction p}(\varphi) \rightarrow (Pr_{\sigma}(\varphi) \rightarrow Pr_{\sigma'\restriction p}(\varphi)).$$

But then, since $T\vdash Pr_{\sigma}(\varphi)$, we get $T\vdash Pr_{\sigma'\restriction p}(\varphi)$, as desired.

Since $\sigma'(x)$ binumerates S in T, it follows from (1) that (ii) holds.

To show that $\sigma(x)$ binumerates S in T it suffices to show that for all φ and p,

$$T\vdash Prf_T(Pr_{\sigma}(\varphi),p) \rightarrow Pr_{\sigma'\restriction p}(\varphi).$$

But this, too, follows at once from (1). ∎

Proof of Lemma 8. The following proof is a modification of the proof of Theorem 4. The interpretation t constructed in that proof does not necessarily have the additional property that

(1) $T\vdash t(\varphi)$ implies $T\vdash Pr_{\sigma}(\varphi)$.

To achieve this we proceed as follows. The function c, the set Z, and the formula $\zeta(x)$ are the same as before, but the definition of φ_n is different. Here we put

(2) $\varphi_n := \theta_n$ if $S + Z\vdash \bigwedge\{\varphi_m: m < n\}\rightarrow\theta_n$ or

$$(S + Z\nvdash \bigwedge\{\varphi_m: m < n\}\rightarrow \neg\theta_n \;\&\; n\in Y),$$

$:= \neg\theta_n$ otherwise,

where Y is any set of natural numbers.

As before let $X = \{\varphi_n: n\in N\}$. Either θ_n or $\neg\theta_n$ is put in X. We put θ_n in X if putting $\neg\theta_n$ in X would make X inconsistent, and similarly for $\neg\theta_n$. Otherwise we put θ_n in X iff $n\in Y$. The idea is to achieve (1) by letting Y be formally represented by a sufficiently independent formula $\eta(x)$.

Let $\gamma(x) := \sigma(x) \vee \zeta(x)$. Let $\eta(x)$ be as in Theorem 2.10 with $\delta(x) := Pr_{\gamma}(x)$. Next, as in the proof of Theorem 4, let $\chi(x,y)$ be the formalization of the result of turning the inductive definition of φ_n into an explicit definition using $\eta(x)$ to represent Y. Let Let $\xi(x) := \exists y\chi(x,y)$.

As in the proof of Theorem 4 we can now define an interpretation t of S in T such that

(3) $T \vdash t(\varphi) \leftrightarrow \xi(\varphi)$.

It remains to be shown that (1) holds.

Suppose $T \nvdash Pr_\sigma(\varphi)$. We must then show that $T \nvdash t(\varphi)$. We have $T \nvdash Pr_\gamma(\varphi)$ (see (6) in the proof of Theorem 4). For any $f \in 2^N$, let $Y_f = \{Pr_\gamma(n)^{f(n)}: n \in N\}$. Now let f(n) be such that $f(\varphi) = 1$ and

(4) $T + Y_f$ is consistent.

Next we define ψ_n as follows (compare (2)).

$$\psi_n := \theta_n \text{ if } Pr_\gamma(\bigwedge\{\psi_m:m<n\} \to \theta_n) \in Y_f \text{ or}$$
$$\big(Pr_\gamma(\bigwedge\{\psi_m:m<n\} \to \neg\theta_n) \notin Y_f \ \&\ Pr_\gamma(\lambda_n) \notin Y_f\big),$$
$$:= \neg\theta_n \text{ otherwise,}$$

where $\lambda_n := \bigwedge\{\psi_m: m < n\} \wedge \theta_n \to \varphi$. Let $g \in 2^N$ be such that

$$g(n) = 0 \text{ iff } Pr_\gamma(\lambda_n) \notin Y_f$$

and set

$$Y_{f,g} = Y_f + \{\eta(n)^{g(n)}: n \in N\}.$$

Then, by (4) and the choice of $\eta(x)$,

(5) $T + Y_{f,g}$ is consistent.

Recalling the definition of $\chi(x,y)$, we can now show, by induction, that for every n, $T + Y_{f,g} \vdash \chi(\psi_n, n)$ and so

(6) $T + Y_{f,g} \vdash \xi(\psi_n)$.

Next we show, by induction, that for every n,

(7) $Pr_\gamma(\bigwedge\{\psi_m:m<n\} \to \varphi) \notin Y_f$.

Note that, by (4), $\{\psi: Pr_\gamma(\psi) \in Y_f\}$ is closed under logical deduction. Since $Pr_\gamma(\varphi) \notin Y_f$, (7) holds for n = 0. Suppose (7) holds for n = k.

Case 1. $\psi_k := \theta_k$. Then either $Pr_\gamma(\bigwedge\{\psi_m:m<k\} \to \theta_k) \in Y_f$ or $Pr_\gamma(\bigwedge\{\psi_m:m<k+1\} \to \varphi) \notin Y_f$. In the latter case (7) holds for n = k+1. In the former case we have $Pr_\gamma(\bigwedge\{\psi_m:m<k\} \to \psi_k) \in Y_f$ and so (7) for n = k+1 follows from the inductive assumption.

Case 2. $\psi_k := \neg\theta_k$. Then

(8) $Pr_\gamma(\lambda_k) \in Y_f$.

For suppose $Pr_\gamma(\lambda_k) \notin Y_f$. If $Pr_\gamma(\bigwedge\{\psi_m:m<k\} \to \neg\theta_k) \in Y_f$, then $Pr_\gamma(\bigwedge\{\psi_m:m<k\} \wedge \theta_k \to \theta) \in Y_f$ for every θ and so, in particular, $Pr_\gamma(\lambda_k) \in Y_f$, contrary to assumption. So $Pr_\gamma(\bigwedge\{\psi_m:m<k\} \to \neg\theta_k) \notin Y_f$. But then $\psi_k := \theta_k$, a contradiction. This proves (8) and completes the proof of (7).

From (7) it follows that for some k, $\varphi := \neg\psi_k$. Hence, by (6), $T + Y_{f,g} \vdash \xi(\neg\varphi)$. But then, by (3) and (5), $T \nvdash t(\varphi)$. Thus, (1) holds and the proof is complete. ■

Proof of Theorem 13. "If". By Corollary 2, there is a formula $\sigma'(x)$ binumerating S in T such that $T \vdash Con_{\sigma'}$. But then, by Lemma 7, there is a formula $\sigma(x)$ numerating S in T and such that $T \vdash Con_\sigma$ and Lemma 7 (ii) holds. Now let t be as in Lemma 8.

Then t: $S \leq T$. Let φ be any sentence of S such that $T \vdash t(\varphi)$. Then, by Lemma 8, $T \vdash Pr_{\sigma}(\varphi)$ and so there is a q such that $T \vdash Pr_{S\restriction q}(\varphi)$. It follows that $T \vdash Pr_{\emptyset}(\bigwedge S\restriction q \to \varphi)$ and so, by hypothesis, $S \vdash \bigwedge S\restriction q \to \varphi$, whence $S \vdash \varphi$. Thus, t is faithful.

"Only if". Suppose $S \trianglelefteq T$. Then $S \leq T$. Let φ be such $T \vdash Pr_{\emptyset}(\varphi)$. Suppose t: $S \leq T$ is faithful. Let κ be the sentence saying that t is an interpretation of $\emptyset$ (logic) in T. Then, by Fact 12 (b),

$$PA \vdash Pr_{\emptyset}(\varphi) \to Pr_{\emptyset}(\kappa \to t(\varphi)).$$

But then $T \vdash Pr_{\emptyset}(\kappa \to t(\varphi))$. Since T is essentially reflexive, it follows that $T \vdash \kappa \to t(\varphi)$. But $T \vdash \kappa$ and so $T \vdash t(\varphi)$. But then, t being faithful, $S \vdash \varphi$, as desired. ■

Proof of Theorem 14. Suppose first $A \dashv_{\Pi_1} B \dashv_{\Sigma_1} A$. Then, by Theorem 6, $A \leq B$. Suppose $B \vdash Pr_{\emptyset}(\varphi)$. Then, $Pr_{\emptyset}(\varphi)$ being Σ_1, it follows that $A \vdash Pr_{\emptyset}(\varphi)$. Since A is essentially reflexive, this implies that $A \vdash \varphi$. Hence, by Theorem 13, $A \trianglelefteq B$.

Next suppose $A \trianglelefteq B$. By Theorem 6, $A \dashv_{\Pi_1} B$, and, by Lemma 6, $B \dashv_{\Sigma_1} A$. ■

The analogue of Theorem 11 (a) for $\trianglelefteq$ now follows at once from Theorem 14 and Theorem 5.4 (a) with, say, $\Gamma = \Pi_2$. We write $A \simeq B$ to mean that $A \trianglelefteq B \trianglelefteq A$.

Corollary 10. If $A \dashv B$, there is a sentence φ such that $A + \varphi \simeq B$.

The analogue of Theorem 11 (b), on the other hand, is clearly false. (Let σ_k be Σ_1 sentences such that $T + \{\sigma_k : k < n\} \nvdash \sigma_n$ for every n and let $X = \{\sigma_k : k \in N\}$. Let σ be any Σ_1 sentence such that $T + \sigma \trianglelefteq T + X$. Then, by Lemma 6, $T + \sigma \vdash X$, whence $T + X \nvdash \sigma$ and so, again by Lemma 6, $T + X \ntrianglelefteq T + \sigma$.)

If S is finite, then $\{\varphi : S \leq T + \varphi\}$ is r.e., but if $\leq$ is replaced by $\trianglelefteq$ this is no longer true:

Corollary 11. Suppose $Q \dashv S \trianglelefteq T$. Then $X = \{\varphi : S \trianglelefteq T + \varphi\}$ is a complete Π^0_2 set.

Proof. By Theorem 13, X is Π^0_2. Let Y be any Π^0_2 set. By the proof of Theorem 5.6 (a), for $\Gamma = \Sigma_1$, there is a formula $\xi(x)$ such that

(1) if $k \in Y$, then $\xi(k)$ is Σ_1–conservative over T,

(2) if $k \notin Y$, then there is a Σ_1 sentence σ such that $T + \xi(k) \vdash \sigma$ and $S \nvdash \sigma$.

It is now sufficient to show that

(3) $Y = \{k : \xi(k) \in X\}$.

Suppose first $k \in Y$. Let ψ be any sentence such that $T + \xi(k) \vdash Pr_{\emptyset}(\psi)$. Then, by (1), $T \vdash Pr_{\emptyset}(\psi)$. But then, by Theorem 13, $S \vdash \psi$. Using Theorem 13 once again, we get $S \trianglelefteq T + \xi(k)$, i.e. $\xi(k) \in X$.

Next suppose $k \notin Y$. Let σ be as in (2). Since σ is Σ_1, $PA + \sigma \vdash Pr_Q(\sigma)$ and so $PA + \sigma \vdash Pr_{\emptyset}(\bigwedge Q \to \sigma)$. It follows that $T + \xi(k) \vdash Pr_{\emptyset}(\bigwedge Q \to \sigma)$. On the other hand $S \nvdash \bigwedge Q \to \sigma$. Hence, by Theorem 13, $\xi(k) \notin X$. ■

Finally, we improve Corollary 8 as follows.

Corollary 12. If $A \leq B$, then $\mathrm{Int}_{A,B}$ is a complete Π_2^0 set.

Proof. Let $X = \{k: \forall m R(k,m)\}$, where $R(k,m)$ is r.e., be any Π_2^0 set. By Theorem 3.1, there is a formula $\rho(x,y)$ numerating $R(k,m)$ in B. Let $\alpha(x)$ be a formula binumerating A in B. Let $\sigma(x,y) :=$

$$\alpha(x) \wedge \mathrm{Con}_{\alpha \restriction x} \wedge \forall z{\leq}x\rho(y,z).$$

Then, by Lemma 3, for every k,

(1) $\quad PA \vdash \mathrm{Con}_{\sigma(x,k)}$.

By Lemma 2, for every n, $B \vdash \mathrm{Con}_{A \restriction n}$. It follows that

(2) $\quad$ if $k \in X$, then $\sigma(x,k)$ binumerates A in B.

Also, clearly,

(3) $\quad$ if $k \notin X$, there is an m such that $B \nvdash \exists x(m \leq x \wedge \sigma(x,k))$.

By (1) and the proof of Lemma 8, we can for each k, effectively find a translation t_k such that

(4) $\quad t_k$: $\{\varphi: B \vdash \sigma(\varphi,k)\} \leq B$,

(5) $\quad$ if $B \vdash t_k(\varphi)$, then $B \vdash \mathrm{Pr}_{\sigma(x,k)}(\varphi)$.

To complete the proof it suffices to show that

(6) $\quad X = \{k: t_k \in \mathrm{Int}_{A,B}\}$.

If $k \in X$, then, by (2) and (4), $t_k \in \mathrm{Int}_{A,B}$. Suppose $k \notin X$. Let m be as in (3). Let θ be such that

$$PA \vdash \theta \leftrightarrow \neg \mathrm{Pr}_{A \restriction m}(\theta).$$

Then, since A is essentially reflexive,

(7) $\quad A \vdash \theta$.

Since $A \leq B$, it follows, by Theorem 6, that $B \vdash \neg \mathrm{Pr}_{A \restriction m}(\theta)$. By the definition of $\sigma(x,y)$, this implies that

$$B \vdash \mathrm{Pr}_{\sigma(x,k)}(\theta) \rightarrow \exists x(m \leq x \wedge \sigma(x,k)).$$

But then, by (3), $B \nvdash \mathrm{Pr}_{\sigma(x,k)}(\theta)$ and so, by (5) and (7), $t_k \notin \mathrm{Int}_{A,B}$. This proves (6) and so the proof is complete. ■

Exercises for Chapter 6.

In the following exercises we assume that $PA \dashv T$ and that A, B, C are extensions of T.

1. Show that there is a Π_1 sentence φ such that $Q + \varphi \not\leq S$ and $Q + \neg\varphi \not\leq S$ (compare Theorem 8.2).

2. (a) Suppose $A \dashv B \not\leq A$. Show that there is a C such that $A \dashv C \dashv B$ and $B \not\leq C \not\leq A$. [Hint: There is a sentence θ such that $B \vdash \theta$ and $\{\theta\} \not\leq A$. The sets $\{\varphi: \{\theta\} \leq A + \neg\varphi\}$ and $\{\varphi: Q + \theta \vee \varphi \leq A\}$ are r.e. and monconsistent with Q.]

(b) Suppose $A < B$. Show that there is a C such that $A < C < B$.

3. The proof of Theorem 4 actually yields the following stronger result: There is a finite subtheory PA_σ of PA such that $S \leq PA_\sigma + \{\sigma(\varphi): \varphi \in S\} + Con_\sigma$. Use this to prove the following:

(a) If $\tau(x)$ numerates T in a finite subtheory of T, then $T + Con_\tau \not\leq T$ (compare Theorem 2).

(b) T is interpretable in a bounded subtheory of T (compare Corollary 4.1 (a) and Theorem 3).

4. (a) Suppose σ_0, σ_1 are Σ_1 sentences such that $T + \sigma_i \leq T$, $i = 0, 1$. Show that $T + \sigma_0 \wedge \sigma_1 \leq T$.

(b) Show that there is a Π_1^0 set X of Σ_1 sentences such that $T + Y \leq T$ for every finite (and so for every r.e.) subset Y of X and $T + X \not\leq T$ (compare Theorem 5). [Hint: Let $\tau(x)$ be a PR binumeration of T. Let $\rho(x,y)$ be a PR formula such that $\{k: \exists m PA \vdash \rho(k,m)\}$ is not recursive. Let $\gamma(x,y) := \tau(x) \wedge \forall z{\leq}x \neg\rho(y,z)$. Let

$$X = \{\neg Con_{\gamma(x,k)}: T + \neg Con_{\gamma(x,k)} \leq T\}.]$$

5. Improve Corollary 3 by showing that $Int_{A,B} \subseteq Int_{A+\sigma', B+\sigma}$.

6. (a) Use Exercise 2.15 (b) to give an alternative proof of Theorem 8.

(b) Use Exercise 2.16 and Theorem 8 to give another proof of Theorem 9.

7. Suppose $PA \dashv S_1$. Prove the converse of Theorem 1: If for every Σ_1 numeration $\sigma_1(x)$ of S_1, there is a Σ_1 numeration $\sigma_0(x)$ of S_0 such that $PA \vdash Con_{\sigma_1} \to Con_{\sigma_0}$, then $S_0 \leq S_1$. [Hint: Use Theorem 5 and Exercise 2.16.]

8. Show that there is a (Π_1, Σ_1) sentence θ such that $T \vdash Con_T \to Con_{T+\theta}$ and $T + \theta \not\leq T$.

9. Let θ be a Π_1 Rosser sentence for T and let $\psi :=$

$$\forall u(Prf_T(\neg\theta,u) \to \exists z{\leq}u Prf_T(\theta,z)).$$

Show that $T + \theta \equiv T + \neg\psi$, $T + \psi \equiv T + \neg\theta$, $T + \theta < T + Con_T$, $T + \psi < T + Con_T$.

10. Suppose X is r.e. and monoconsistent with T. Let $\rho(x,y)$ be a PR formula such that $X = \{k: \exists m PA \vdash \rho(k,m)\}$.

(a) Let φ be such that

$$PA \vdash \varphi \leftrightarrow \forall z(Con_{T\restriction z+\varphi} \to \neg\rho(\varphi,z)).$$

Show that $T + \varphi \leq T$ and $\varphi \notin X$.

(b) The sentence φ in (a) is Π_2. This can be improved. Let χ be such that

$$PA \vdash \chi \leftrightarrow \exists z(\neg Con_{T\restriction z+\chi} \wedge \forall u{\leq}z \neg\rho(\chi,u)).$$

Then χ is Σ_1. Show that $T + \chi \leq T$ and $\chi \notin X$ (compare Theorem 9).

11. (a) Let φ be such that

$$PA\vdash \varphi \leftrightarrow \forall z(Con_{T\restriction z+\varphi} \to Con_{T\restriction z+\neg\varphi}).$$

Show that φ is an Orey sentence for T (compare Theorem 10).

(b) Suppose $A \dashv B$. Let φ be such that

$$PA\vdash \varphi \leftrightarrow \forall z(Con_{A\restriction z+\varphi} \to Con_{B\restriction z}).$$

Show that $A + \varphi \equiv B$ (compare Theorem 11 (a)).

(c) Let φ be such that

$$PA\vdash \varphi \leftrightarrow \forall z(Con_{A\restriction z+\varphi} \to Con_{B\restriction z+\neg\varphi}).$$

Show that $A + \varphi \equiv B + \neg\varphi$.

12. (a) Show that

$$PA\vdash \forall x(Con_{S_0\restriction x} \to Con_{S_1\restriction x}) \leftrightarrow (Con_{S_1} \vee \exists x(\neg Con_{S_0\restriction x} \wedge \forall y{<}x Con_{S_1\restriction y})).$$

Conclude that the sentences φ of Exercise 11 are Δ_2 (compare Exercise 5.9 (a) and Theorem 7.8). In particular, there is a Δ_2 Orey sentence for T.

(b) Show that no Orey sentence for T is B_1.

13. Let $\tau^*(x)$ be as in Theorem 2.7. In Theorem 4 let $\sigma(x) := \tau^*(x)$ and $S = T$. Next let $\xi(x)$ be as in (14) of the proof of Theorem 4. Let φ be such that $PA\vdash \varphi \leftrightarrow \neg\xi(\varphi)$. Show that φ is an Orey sentence for T.

14. Let $\tau(x)$ be any formula binumerating T in T. Let φ be such that

$$PA\vdash \varphi \leftrightarrow \neg Pr_\tau(\varphi).$$

Show that

(i) $T + \neg\varphi \le T$,

(ii) for every n, $T + \varphi \le T + Pr_{T\restriction n}(Con_\tau)$.

Let $\tau(x)$ be the formula $\tau^*(x)$ mentioned in Theorem 2.7. Conclude that φ is then an Orey sentence for T.

15. Suppose $T \le S$. Show that there is a Σ_1 formula $\xi(x)$ such that

$$\{k: T + \xi(k) \le S\}$$

is a complete Π^0_2 set (compare Theorem 12). [Hint: Let R(k,m) be an r.e. relation such that {k: $\forall m R(k,m)$} is a complete Π^0_2 set. There is a Σ_1 formula $\rho(x,y)$ such that

if R(k,m), then $Q\vdash \rho(k,m)$,

if not R(k,m), then $Q + \rho(k,m) \not\le S$

(Lemma 3.1). Let $\xi(x)$ be such that

$$PA\vdash \xi(k) \leftrightarrow \exists z(\neg Con_{T\restriction z+\xi(k)} \wedge \forall u{\le}z\rho(k,u)),$$

compare Exercise 10 (b).]

16. (a) By Orey's compactness theorem (Theorem 5), there is a function f(k) such that for every sentence φ, if $T + \varphi \not\le T$, then $T\restriction f(\varphi) + \varphi \not\le T$. Show that f(k) cannot be recursive.

(b) By Theorem 6, there is a function g(k) such that for every sentence φ, if

$T + \varphi \not\trianglelefteq) T$, then $g(\varphi)$ is a Π_1 sentence such that $T + \varphi \vdash g(\varphi)$ and $T \nvdash g(\varphi)$. Show that $g(k)$ cannot be recursive.

17. Show that there are sentences φ_0, φ_1 such that $T + \varphi_i \leq T$, $T + \varphi_0 \wedge \varphi_1 \not\leq T$, $T + \neg\varphi_i \not\leq T$, $T + \neg\varphi_0 \vee \neg\varphi_1 \leq T$, $i = 0, 1$. [Hint: Use Exercise 5.8 (b).]

18. Show that, even if T is not Σ_1–sound, there is a Σ_1 formula $\tau(x)$ binumerating T in T such that $Pr_\tau(x)$ numerates Th(T) in T (by Exercise 2.22 (iv), $\tau(x)$ cannot be PR).

19. Show that if $A \leq B$, then $\{\varphi \in \Sigma_1 : A + \varphi \trianglelefteq B\}$ is a complete Π^0_2 set.

20. (a) Show that if S is finite and $Q \dashv S \trianglelefteq T$, then the set of faithful interpretations of S in T is a complete Π^0_2 set. [Hint: First show that there is a sentence θ such that $S \nvdash \theta$ and $S + \theta \leq T$.]

(b) Suppose $A \trianglelefteq B$. Show that the set of faithful interpretations of A in B is a complete Π^0_2 set.

21. S is *X–faithfully interpretable* in S′, $S \trianglelefteq_X S'$, if there is an interpretation $t: S \leq S'$ which is *X–faithful* in the sense that for every $\varphi \in X$, if $S' \vdash t(\varphi)$, then $S \vdash \varphi$. Show that
(i) $S \trianglelefteq_X T$ iff $S \leq T$ and for every $\varphi \in X$, if there is an m such that $T \vdash Pr_{S \restriction m}(\varphi)$, then $S \vdash \varphi$,
(ii) if $S \leq T$, then $S \trianglelefteq_X T$, where $X = \{\varphi : S \trianglelefteq_{\{\varphi\}} T\}$,
(iii) if $S \trianglelefteq_X T$ and $S \leq T' \dashv T$, then $S \trianglelefteq_X T'$,
(iv) if $S \trianglelefteq T$ and $S \dashv S' \leq T' \dashv T$, then $S' \trianglelefteq T'$,
(v) $\trianglelefteq$ cannot be replaced by $\trianglelefteq_X$ in (iv),
(vi) $A \trianglelefteq_X B$ iff $A + (Th(B) \cap \Sigma_1) \dashv_X B \dashv_{\Pi_1} A$,
(vii) $A \trianglelefteq B$ iff $A \trianglelefteq_{\Sigma_1} B$ iff $A \trianglelefteq_{\{\varphi\}} B$ for every (Σ_1) sentence φ.

22. Show that $A \dashv_{\Pi_n} B$ iff there is a $t: A \leq B$ such that for every Π_n sentence ψ, $B \vdash t(\psi) \to \psi$ (compare Theorem 6; note that for every $t: A \leq B$ and every Π_1 sentence ψ, $B \vdash t(\psi) \to \psi$, by Lemma 1). [Hint: "Only if". For every k and every Π_n sentence φ, $B \vdash Pr_{A \restriction k}(\varphi) \to \varphi$. Use this to construct a formula $\alpha(x)$ binumerating A in B and such that $PA \vdash Con_\alpha$ and $B \vdash \chi \to \alpha(\chi)$ for every Σ_n sentence χ.]

Notes for Chapter 6.

The general concept (*relative*) *interpretation* due to Tarski (cf. Tarski, Mostowski, Robinson (1953); in keeping with recent usage we omit "relative"); it is an important tool in proofs of (relative) consistency and (un)decidability. The investigation of interpretability for its own sake was initiated by Feferman (1960). Theorems 1, 2, 3 are due to Feferman (1960); concerning the (im)possibility of improving Theorem 3, see Exercise 3 (b). Theorem 4 is due to Feferman (1960) building on work of Bernays (Hilbert and Bernays (1939)) and Wang (1951); for a strengthening of Theorem 4, see Exercise 3. Lemma 2 is implicit in Feferman (1960), all but explicit in Orey (1961), and fully explicit in Hájek (1971). Corollary 2 is due to Orey (1961). Theorem 5 is due to Orey (1961) (cf. also Feferman (1960)). Theorem 6 was first stated by Guaspari (1979) and Lindström (1979); for a more general result, due to Guaspari (1979), see Exercise 22. Corollary 4 is due to Goryachev (1986). Theorem 8 is due to Feferman (1960) (with a different proof; see Exercise 6 (a)). Lemma 4 is due to Švejdar (1978). Theorem 9 and Corollary 6 are essentially due to Hájek (1971) (cf. also Hájková and Hájek (1972)) (with a different proof; see Exercise 10 (a)); for yet another proof, see Exercise 6 (b). Theorem 10 less the references to the set X is due to Orey (1961); the full result is proved in Lindström (1979), (1984a); related results, for certain nonreflexive theories, requiring methods not explained here, can be found in Hájek and Pudlák (1993). For more information on Orey sentences, see Exercises 11 (a), 12 (b), 13, 14. Corollary 5 has also been pointed out by Guaspari (1979); for a related result, see Exercise 8. The result on finite conservative extensions mentioned just before Lemma 5 is due to Kleene (1952b) (cf. also Kaye (1991)). Theorem 11 is due to Lindström (1979) (see Exercise 11 (b)) and (1984a); by Exercises 11 (b) and 12, the sentence φ in Theorem 11 (a) can be taken to be Δ_2 (cf. also Theorem 7.8). Theorem 12 is essentially due to Solovay (cf. Hájek (1979)) (with a different proof); the present proof is from Lindström (1984a) (see also Exercise 15).

The concept *faithful interpretation* was introduced in Feferman, Kreisel, Orey (1960). They observed that if $Q \dashv S \trianglelefteq S'$ and S is Σ_1–sound, so is S′ (see Lemma 6). Theorems 13 and 14 are due to Lindström (1984c); see also Exercise 21. Corollary 9 (b) is due to Feferman, Kreisel, Orey (1960). Lemma 7 is due to Lindström (1984c); the present proof is an instance of a general argument described in Lindström (1988). Lemma 8 is due to Lindström (1984c), but the main idea of the proof, to introduce the set Y and represent Y by a sufficiently independent formula, is taken from Feferman, Kreisel, Orey (1960). Corollaries 10, 11, 12 are due to Lindström (1984c); for related results, see Exercises 19 and 20; Exercise 7.8, below, is an improvement of Corollary 10.

An alternative notion of interpretability, *feasible interpretability*, has been studied by Verbrugge (1992), (1994). For any formal entity q, formula, proof, etc., let $|q|$ be the length of q, i.e. the number of (instances of) symbols occurring in q. S is *feasibly interpretable* in T, $S \leq_f T$, if there is an interpretation t: $S \leq T$ which is *feasible* in the sense that there is a polynomial P(n) such that for every $\varphi \in S$, there is a proof p of $t(\varphi)$ in T such that $|p| \leq P(|\varphi|)$. Clearly, $\{\varphi\colon S + \varphi \leq_f T\}$ is Σ^0_2. Thus, by Theorem 12, $S \leq T$ does not imply $S \leq_f T$ (cf. Verbrugge (1992)).

Exercise 1 is due to Montague (1957), (1962). Exercise 2 (a) is due to Jeroslow (1971a); Exercise 2 (b) is due to Švejdar (1978). Exercise 3 is due to Feferman (1960). Exercise 4 (b) (with a different proof) is essentially due to Orey (1961). Exercise 9 is due to Švejdar (1978). Exercise 10 (a) is essentially due to Hájek (1971). Exercise

11 (a) is due to Lindström (1979) and Švejdar (1978). Exercise 12 was pointed out to me by Franco Montagna (compare Theorem 7.8). Exercise 13 is due to Orey (1961). Exercise 16 (a) is due to Jeroslow (1971b). Exercise 17 can be substantially improved using results on the modal logic of (provability and) interpretability, due to Berarducci (1990), Shavrukov (1988), and Strannegård (1997). Exercise 22 is due to Guaspari (1979).