Isomorphisms of Diagonalizable Algebras

V. Yu. Shavrukov[*]

Spaarndammerdijk 47–D

1013 ZP Amsterdam

the Netherlands

volodya@phil.ruu.nl

March 13, 1996

For a formal theory T, the *diagonalizable algebra* (a.k.a. Magari algebra) of T, denoted $\mathfrak{D}_T$, is the Lindenbaum sentence algebra of T endowed with the unary operator $\square_T$ arising from the provability predicate of T: (the equivalence class of) a sentence φ is sent by $\square_T$ to (the equivalence class of) the T-sentence expressing that T proves φ. It was shown in Shavrukov [6] that the diagonalizable algebras of PA and ZF, as well as the diagonalizable algebras of similarly related pairs of Σ_1-sound theories, are not isomorphic. Neither are these algebras first-order equivalent (Shavrukov [7, Theorem 2.11]).

In the present paper we establish a sufficient condition, which we name $\mathcal{B}(\square)$ coherence, for the diagonalizable algebras of two theories to be isomorphic. It is then immediately seen that $\mathfrak{D}_{ZF} \cong \mathfrak{D}_{GB}$, which answers a question of Smoryński [11]. We also construct non-identity automorphisms of diagonalizable algebras of all theories under consideration. The techniques we use are a combination of those developed in the context of partially conservative sentences (cf. Lindström [4]), and those of Pour-El & Kripke [5]. A related construction appears in Solovay [12, Theorem 1].

1. CONVENTIONS. *Theories* in this paper are first order r.e. theories over classical predicate logic. We assume that a translation of the arithmetic language into that of each theory is fixed and do not distinguish between arithmetical and translated arithmetical formulas. The arithmetic of any theory is presumed to be at least as strong as $I\Delta_0 + Exp$ (cf. Hájek & Pudlák [3, I.1(b)]).

We also assume that each theory T comes equipped with a *provability predicate* $\square_T$ (we shall be omitting the subscript whenever reasonable). Note that we use the same notation for the provability predicate as for an opeator of $\mathfrak{D}_T$. The formula $\square_T \varphi$ expresses in a natural way that φ has a proof from an axiom set specified by a (Kalmár) elementary formula which describes, in the standard model, a set of sentences axiomatizing T. (Kalmár elementary formulas are essentially the same as $\Sigma_0^{exp}(exp)$ formulas of Hájek & Pudlák [3, I.1.29].) It is then easily verified in $I\Delta_0 + Exp$ that $\square_T$ is closed under

---

classical predicate logic (see [3, I.4(a)]). We shall use the notation $x : \Box_T \varphi$ to denote the elementary formula expressing that the sentence φ has a T-proof whose code is $\leq x$.

2. DEFINITIONS. Let S and T be theories and let Γ be a class of sentences common to the languages of both theories. S and T are called Γ-*coherent* if

(i) S $\vdash \gamma$ iff T $\vdash \gamma$ for any Γ sentence γ, and

(ii) both S and T prove $\forall x\, \exists y\, \forall\gamma{\in}\Gamma\, ((x : \Box_S \gamma \to y : \Box_T \gamma) \land (x : \Box_T \gamma \to y : \Box_S \gamma))$.

If $\mathcal{B}(\Gamma) = \Gamma$, i.e. if Γ is closed under Boolean combinations, and $\vec{\xi} = (\xi_1, \ldots, \xi_n)$ and $\vec{\varphi} = (\varphi_1, \ldots, \varphi_n)$ are tuples of S- and T-sentences respectively, then the pairs $(\text{S}; \vec{\xi})$ and $(\text{T}; \vec{\varphi})$ are Γ-*coherent* if

(i$'$) S $\vdash b(\vec{\gamma}; \vec{\xi})$ iff T $\vdash b(\vec{\gamma}; \vec{\varphi})$ for any tuple $\vec{\gamma}$ of Γ sentences and any Boolean term b, and

(ii$'$) both S and T prove

$$\forall x\, \exists y\, \forall\vec{\gamma}{\in}\vec{\Gamma}\, \forall b{\in}\mathcal{B}$$

$$\Big(\big(x : \Box_S b(\vec{\gamma}; \vec{\xi}) \to y : \Box_T b(\vec{\gamma}; \vec{\varphi})\big) \land \big(x : \Box_T b(\vec{\gamma}; \vec{\varphi}) \to y : \Box_S b(\vec{\gamma}; \vec{\xi})\big) \Big),$$

where $\vec{\gamma} \in \vec{\Gamma}$ means that $\vec{\gamma}$ is a finite tuple of Γ sentences, and $b \in \mathcal{B}$ is short for 'b is a Boolean term'. The purpose of assuming the closure of Γ under $\mathcal{B}$ is that this definition agree with the previous one for empty tuples $\vec{\xi}$ and $\vec{\varphi}$.

Let us note that the complicated statement of (ii) will be equivalent to the more familiar condition of formalized mutual Γ conservativity

$$\text{S}, \text{T} \vdash \forall\gamma{\in}\Gamma\, (\Box_S \gamma \leftrightarrow \Box_T \gamma)$$

in case the theories S and T are closed under the Σ_1 *collection rule* (Σ_1–CR):

$$\frac{\forall x\, \exists y\, \varsigma(x,y)}{\forall z\, \exists w\, \forall x{\leq}z\, \exists y{\leq}w\, \varsigma(x,y)} \quad \text{for } \Sigma_1 \text{ formulas } \varsigma(x,y)$$

(see Beklemishev [1, Proposition 5.3]). Clause (ii$'$) can be similarly simplified in the presence of this rule. Σ_1–CR is obviously equivalent over $\text{I}\Delta_0 + \text{Exp}$ to the collection rule for elementary formulas. Theories containing $\text{B}\Sigma_1$ are closed under Σ_1–CR.

For a theory S, we let $\Box_S$ denote the class of sentences of the form $\Box_S \varphi$ and hope this does not lead to confusion. The Friedman–Goldfarb–Harrington Principle, which is readily verifiable in $\text{I}\Delta_0 + \text{Exp}$, states that the class of $\Box_S$ sentences is, modulo ($\text{I}\Delta_0 +$ Exp)-provable equivalence, the same as the class of Σ_1 sentences S-provably implied by $\Box_S \bot$ (see e.g. Smoryński [10, p. 366]). If another theory T is $\mathcal{B}(\Box_S)$-coherent with S, then both S and T prove $\Box_S \bot \leftrightarrow \Box_T \bot$, and since $\Box_T \varphi$ is a Σ_1 sentence implied by $\Box_T \bot$, it is provably equivalent to a $\mathcal{B}(\Box_S)$ sentence. Thus the classes $\Box_S$ and $\Box_T$ essentially coincide, as do $\mathcal{B}(\Box_S)$ and $\mathcal{B}(\Box_T)$. Therefore we can just speak of $\mathcal{B}(\Box)$ coherence in place of $\mathcal{B}(\Box_S)$ or $\mathcal{B}(\Box_T)$ one.

$\mathcal{B}(\Box)$ coherence is the central notion in this paper. Observe that by the Friedman–Goldfarb–Harrington Principle the class $\Box$ is closed under conjunction and disjunction, and therefore, when establishing $\mathcal{B}(\Box)$ coherence between two theories, it is sufficient in

(i) and (ii) to consider only sentences γ of the forms $\Box\,\varphi \to \Box\,\psi$ and $\neg\,\Box\,\varphi$ rather than arbitrary $\mathcal{B}(\Box)$ sentences.

Given a class Γ of arithmetical sentences of bounded complexity, e.g. $\Gamma = \Sigma_n$, we recall that there is a formula $\mathrm{True}(\cdot)$, to which we shall refer as *truth definition for* Γ, s.t.

$$\mathrm{I}\Delta_0 + \mathrm{Exp} \vdash \forall\gamma\!\in\!\Gamma\ \Box_{\mathrm{T}}(\gamma \leftrightarrow \mathrm{True}(\gamma))$$

(see e.g. Hájek & Pudlák [3, I.1(d), V.5(b)]). Obviously, for a finite tuple $\vec{\varphi}$ of T-sentences we can (effectively in $\vec{\varphi}$) modify the formula $\mathrm{True}(\cdot)$ in such a way that

$$\mathrm{I}\Delta_0 + \mathrm{Exp} \vdash \forall\chi\!\in\!\mathcal{B}(\Gamma;\vec{\varphi})\ \Box_{\mathrm{T}}(\chi \leftrightarrow \mathrm{True}(\chi)),$$

where $\mathcal{B}(\Gamma;\vec{\varphi})$ is, of course, the class of Boolean combinations of Γ sentences and sentences from $\vec{\varphi}$.

The Small Reflection Principle (Hájek & Pudlák [3, Lemma III.4.40], or Shavrukov [7]) says:

$$\mathrm{I}\Delta_0 + \mathrm{Exp} \vdash \forall x, \varphi\ \Box_{\mathrm{T}}(x : \Box_{\mathrm{T}}\varphi \to \varphi).$$

Combining this with the aforementioned properties of truth definitions we get

$$\mathrm{I}\Delta_0 + \mathrm{Exp} \vdash \forall x\ \Box_{\mathrm{T}}\forall\chi\!\in\!\mathcal{B}(\Gamma;\vec{\varphi})\ (x : \Box_{\mathrm{T}}\chi \to \mathrm{True}(\chi)).$$

By Theorem V.1.4 in [3] it follows that an upper bound on the size of the relevant T-proof is (Kalmár) elementary in x.

3. LEMMA. *Let* $(\mathrm{S};\ \vec{\zeta})$ *and* $(\mathrm{T};\ \vec{\psi})$ *be* $\mathcal{B}(\Box)$-*coherent and let* ξ *be an arbitrary* S-*sentence. Then there is a* T-*sentence* φ, *found effectively in* $(\vec{\zeta},\vec{\psi},\xi)$, *s.t.* $(\mathrm{S};\ \vec{\zeta},\xi)$ *and* $(\mathrm{T};\ \vec{\psi},\varphi)$ *are* $\mathcal{B}(\Box)$-*coherent.*

PROOF. Let $\mathrm{True}(\cdot)$ be a truth definition appropriate for $\mathcal{B}(\Box;\vec{\psi})$ sentences. Consider the following self-referentially defined sentence:

$$\varphi \equiv \forall x \Bigg(\bigg(\exists\chi\!\in\!\mathcal{B}(\Box;\vec{\psi})\,\big(\mathrm{True}(\chi) \wedge x : \Box_{\mathrm{T}}(\chi \to \varphi)\big)$$

$$\vee\ \exists\vec{\sigma}\!\in\!\vec{\Box}\ \exists b\!\in\!\mathcal{B}\,\Big(\neg\mathrm{True}\big(b(\vec{\sigma};\vec{\psi})\big) \wedge x : \Box_{\mathrm{S}}\big(\xi \to b(\vec{\sigma};\vec{\zeta})\big)\Big)\bigg)$$

$$\to \bigg(\exists\theta\!\in\!\mathcal{B}(\Box;\vec{\psi})\,\big(\neg\mathrm{True}(\theta) \wedge x : \Box_{\mathrm{T}}(\varphi \to \theta)\big)$$

$$\vee\ \exists\vec{\tau}\!\in\!\vec{\Box}\ \exists c\!\in\!\mathcal{B}\,\Big(\mathrm{True}\big(c(\vec{\tau};\vec{\psi})\big) \wedge x : \Box_{\mathrm{S}}\big(c(\vec{\tau};\vec{\zeta}) \to \xi\big)\Big)\bigg)\Bigg)$$

We claim that the sentence φ is as required in the statement. Recalling the assumption of $\mathcal{B}(\Box)$ coherence between $(\mathrm{S};\ \vec{\zeta})$ and $(\mathrm{T};\ \vec{\psi})$ and considering conjunctive normal forms of Boolean terms one easily sees that it suffices to show that for all $\vec{\sigma} \in \vec{\Box}$ and all $b \in \mathcal{B}$

(1) $\mathrm{T} \vdash \varphi \to b(\vec{\sigma};\vec{\psi})$ iff $\mathrm{S} \vdash \xi \to b(\vec{\sigma};\vec{\zeta})$, and

3

(2) $T \vdash b(\vec{\sigma}; \vec{\psi}) \to \varphi$ iff $S \vdash b(\vec{\sigma}; \vec{\zeta}) \to \xi$

by a proof that formalizes in both S and T producing bounds on the size of proofs as in clause (ii$'$) of Definition 2.

(1, if): Suppose $S \vdash \xi \to b(\vec{\sigma}; \vec{\zeta})$. Fix an x for which we have $x : \square_S(\xi \to b(\vec{\sigma}; \vec{\zeta}))$. Note that there is then a y s.t. $y : \square_S(\gamma \to b(\vec{\sigma}; \vec{\zeta}))$ whenever $x : \square_S(\gamma \to \xi)$ and γ is any sentence. Also, by $\mathcal{B}(\square)$ coherence, there is a z for which there holds $z : \square_T d(\vec{\tau}; \vec{\psi})$ once $y : \square_S d(\vec{\tau}; \vec{\zeta})$, where $\vec{\tau}$ is any sequence of $\square$ sentences and d is a Boolean term. Reason in T:

> Assume φ and $\neg\mathrm{True}(b(\vec{\sigma}; \vec{\psi}))$. Since $x : \square_S(\xi \to b(\vec{\sigma}; \vec{\zeta}))$, one of the two disjuncts in the succedent of φ holds.
>
> C a s e   1. $x : \square_T(\varphi \to \theta)$ and $\neg\mathrm{True}(\theta)$ for some $\theta \in \mathcal{B}(\square; \vec{\psi})$.
> This clearly contradicts the Small Reflection Principle.
> C a s e   2. $x : \square_S(c(\vec{\tau}; \vec{\zeta}) \to \xi)$ and $\mathrm{True}(c(\vec{\tau}; \vec{\psi}))$.
>
> Since $x : \square_S(\xi \to b(\vec{\sigma}; \vec{\zeta}))$, we have $y : \square_S(c(\vec{\tau}; \vec{\zeta}) \to b(\vec{\sigma}; \vec{\zeta}))$. Hence there holds $z : \square_T(c(\vec{\tau}; \vec{\psi}) \to b(\vec{\sigma}; \vec{\psi}))$. By Small Reflection we get $b(\vec{\sigma}; \vec{\psi})$ which contradicts the assumption.

Thus $T \vdash \varphi \to b(\vec{\sigma}; \vec{\psi})$ by a proof that is clearly elementary in x, y, and z, and therefore both S- and T-provably recursive in x.

(2, only if): Suppose $T \vdash b(\vec{\sigma}; \vec{\psi}) \to \varphi$. There is then an x s.t. $x : \square_T(b(\vec{\sigma}; \vec{\psi}) \to \varphi)$. Reason in T:

> Assume $b(\vec{\sigma}; \vec{\psi})$. Then we have φ. Note that $b(\vec{\sigma}; \vec{\psi}) \in \mathcal{B}(\square; \vec{\psi})$. Thus we must either have $x : \square_T(\varphi \to \theta)$ and $\neg\mathrm{True}(\theta)$ with $\theta \in \mathcal{B}(\square; \vec{\psi})$ which, in view of φ and Small Reflection, is impossible, or there must be a $\mathcal{B}(\square; \vec{\zeta})$ sentence $c(\vec{\tau}; \vec{\zeta})$ s.t. $x : \square_S(c(\vec{\tau}; \vec{\zeta}) \to \xi)$ and $\mathrm{True}(c(\vec{\tau}; \vec{\psi}))$.

Thus $T \vdash b(\vec{\sigma}; \vec{\psi}) \to \bigvee\{\, c(\vec{\tau}; \vec{\psi}) \in \mathcal{B}(\square; \vec{\psi}) \mid x : \square_S(c(\vec{\tau}; \vec{\zeta}) \to \xi) \,\}$ by a proof elementary in x. Hence by $\mathcal{B}(\square)$ coherence there holds $S \vdash b(\vec{\sigma}; \vec{\zeta}) \to \bigvee\{\, c(\vec{\tau}; \vec{\zeta}) \in \mathcal{B}(\square; \vec{\zeta}) \mid x : \square_S(c(\vec{\tau}; \vec{\zeta}) \to \xi) \,\}$ which implies $S \vdash b(\vec{\sigma}; \vec{\zeta}) \to \xi$ as required. Once again, it is easily seen that the just constructed S-proof of $b(\vec{\sigma}; \vec{\zeta}) \to \xi$ is both S- and T-provably recursive in x.

(1, only if) and (2, if) are symmetric to (2, only if) and (1, if) respectively.   ∎

4. PROPOSITION.   Let $(S; \xi_1, \ldots, \xi_n)$ and $(T; \varphi_1, \ldots, \varphi_n)$ be $\mathcal{B}(\square)$-coherent. Then there is a recursive isomorphism $e : \mathfrak{D}_S \to \mathfrak{D}_T$ s.t. $e(\xi_i) = \varphi_i$ whenever $1 \le i \le n$, and $e(\sigma) = \sigma$ for all $\square$ sentences σ.

PROOF. Iterating Lemma 3 in a back-and-forth way, we extend the given tuples $\vec{\xi}$ and $\vec{\varphi}$ to recursive sequences $(\xi_k)_{k \in \omega}$ and $(\varphi_k)_{k \in \omega}$ exhausting all S- and T-sentences respectively with the property that $(S; \xi_1, \ldots, \xi_k)$ and $(T; \varphi_1, \ldots, \varphi_k)$ are $\mathcal{B}(\square)$-coherent for any $k \in \omega$.

We claim that $e : \xi_k \mapsto \varphi_k$ is the desired isomorphism between $\mathfrak{D}_S$ and $\mathfrak{D}_T$. For assume $S \vdash \xi_k \leftrightarrow \xi_m$. Then $T \vdash \varphi_k \leftrightarrow \varphi_m$ by $\mathcal{B}(\square)$ coherence, so the definition of e is correct. e is one-one for the same reason, and onto because $(\varphi_k)_{k \in \omega}$ exhausts T-sentences. e preserves Boolean operations c since $S \vdash c(\xi_k, \xi_l) \leftrightarrow \xi_m$ implies $T \vdash c(\varphi_k, \varphi_l) \leftrightarrow \varphi_m$.

Finally, if we have $S \vdash \Box_S \xi_k \leftrightarrow \xi_m$ then $T \vdash \Box_S \xi_k \leftrightarrow \varphi_m$ because $\Box_S \xi_k$ is a $\Box$ sentence. Moreover, $T \vdash \Box_S \xi_k \leftrightarrow \Box_T \varphi_k$ by $\mathcal{B}(\Box)$ coherence, whence $T \vdash \Box_T \varphi_k \leftrightarrow \varphi_m$ follows. Thus e is indeed an isomorphism.

If σ is an arbitrary $\Box$ sentence, then $S \vdash \sigma \leftrightarrow \xi_m$ implies $T \vdash \sigma \leftrightarrow \varphi_m$ by $\mathcal{B}(\Box)$ coherence. Therefore $e(\sigma) = \sigma$ as claimed. ∎

5. THEOREM. *Let* S *and* T *be* $\mathcal{B}(\Box)$-*coherent theories. Then* $\mathfrak{D}_S \cong \mathfrak{D}_T$.

PROOF. Follows at once from Proposition 4. ∎

In the following examples concerning individual theories it is understood that the translation of arithmetic into the languages of the latter is the *conventional* (in most cases the identical) one, and that the diagonalizable algebras feature *natural* provability predicates of those theories, so that known conservativity results among the theories formalize for the provability predicates chosen.

6. EXAMPLES. We have $\mathfrak{D}_S \cong \mathfrak{D}_T$ for the following pairs (S, T):

(a) (ZF, GB), (PA, ACA_0);

(b) (ZFC, ZFC $\pm$ CH);

(c) (PRA, $I\Sigma_1$);

(d) ($I\Sigma_n$, $B\Sigma_{n+1}$), all $n > 0$.

COMMENTS. All these pairs fall under the scope of Theorem 5:

(a). Shoenfield [8] shows that these pairs prove the same theorems in the language of the first-coordinate theory. His proof clearly formalizes in PA.

(b). ZFC knows that ω, $+$, and $\times$ are absolute for L and forcing extensions. Hence these pairs are coherent for first-order arithmetic sentences.

(c). By PRA here we mean the first order theory called (QF($\mathfrak{PR}$)–IA) in Sieg [9]. Among the available proofs of Π_2 conservativity of $I\Sigma_1$ over PRA, the one in Sieg [9, Lemma 2.1.2] probably presents the more convenient opportunity to see that this conservativity formalizes in $I\Sigma_1$ (or even in $I\Delta_0 + \text{Superexp}$) and hence, by that same conservativity, in PRA. On top of that, Π_2 conservativity also allows us to conclude that PRA is closed under Σ_1–CR because $I\Sigma_1$ is, for the latter theory contains $B\Sigma_1$. Thus PRA and $I\Sigma_1$ are Π_2- and hence $\mathcal{B}(\Box)$-coherent.

I hope this example justifies our having mentioned Σ_1–CR in Definition 2, although one can, of course, verify clause (ii) of Definition 2 for PRA and $I\Sigma_1$ directly.

(d). Clote & al. [2, Theorem 14] formalize in $I\Sigma_1$ Paris' theorem that $B\Sigma_{n+1}$ is Π_{n+2}-conservative over $I\Sigma_n$. ∎

We turn next to the subject of automorphisms.

7. THEOREM. *Let* T *be a theory and* ξ *a sentence which is not provably equivalent to any* $\mathcal{B}(\square)$ *sentence. Then there exists a (recursive) automorphism of* $\mathfrak{D}_T$ *that moves* ξ.

PROOF. We construct a sentence φ s.t. $(T; \xi)$ and $(T; \varphi)$ are $\mathcal{B}(\square)$-coherent and $T \nvdash \varphi \leftrightarrow \xi$. To this end we employ essentially the same construction as in Lemma 3:

$$\varphi \equiv \forall x \Bigg(\bigg(\exists \chi \in \mathcal{B}(\square) \cup \{\xi\} \, \big(\mathrm{True}(\chi) \wedge x : \square(\chi \to \varphi) \big)$$

$$\vee \, \exists \rho \in \mathcal{B}(\square) \cup \{\xi\} \, \exists \beta \in \mathcal{B}(\square) \, \Big(\neg\mathrm{True}(\rho) \wedge x : \square \big((\xi \to \beta) \wedge (\beta \to \rho) \big) \Big) \bigg)$$

$$\to \Big(\exists \theta \in \mathcal{B}(\square) \cup \{\xi\} \, \big(\neg\mathrm{True}(\theta) \wedge x : \square(\varphi \to \theta) \big)$$

$$\vee \, \exists \mu \in \mathcal{B}(\square) \cup \{\xi\} \, \exists \delta \in \mathcal{B}(\square) \, \Big(\mathrm{True}(\mu) \wedge x : \square \big((\mu \to \delta) \wedge (\delta \to \xi) \big) \Big) \Big) \Bigg),$$

where $\mathrm{True}(\cdot)$ is an appropriate truth predicate. In a fashion similar to the proof of Lemma 3 we show that for each $\theta \in \mathcal{B}(\square) \cup \{\xi\}$

(1) $T \vdash \varphi \to \theta$ iff for some $\mathcal{B}(\square)$ sentence β, $T \vdash \xi \to \beta$ and $T \vdash \beta \to \theta$, and

(2) $T \vdash \theta \to \varphi$ iff for some $\mathcal{B}(\square)$ sentence β, $T \vdash \theta \to \beta$ and $T \vdash \beta \to \xi$.

By way of example, we only do

(1, only if): Let $x : \square(\varphi \to \theta)$. Step into T:
  Suppose $\neg\theta$. Then $\neg\varphi$. Since there cannot exist χ with $\mathrm{True}(\chi)$ and $x : \square(\chi \to \varphi)$, there are $\rho \in \mathcal{B}(\square) \cup \{\xi\}$ and $\beta \in \mathcal{B}(\square)$ s.t. $\neg\mathrm{True}(\rho)$ and $x : \square((\xi \to \beta) \wedge (\beta \to \rho))$.

So, $T \vdash \bigwedge\{ \rho \in \mathcal{B}(\square) \cup \{\xi\} \mid \exists \beta \in \mathcal{B}(\square) \, x : \square((\xi \to \beta) \wedge (\beta \to \rho)) \} \to \theta$. Let, for each ρ from the conjunction above, $\beta_\rho \in \mathcal{B}(\square)$ be s.t. $x : \square((\xi \to \beta_\rho) \wedge (\beta_\rho \to \rho))$. We have $T \vdash \xi \to \bigwedge_\rho \beta_\rho$, and $T \vdash \beta_\rho \to \rho$ which entails $T \vdash \bigwedge_\rho \beta_\rho \to \theta$. Finally, observe that $\bigwedge_\rho \beta_\rho$ is a $\mathcal{B}(\square)$ sentence.

Now, (1) and (2) imply that for any $\mathcal{B}(\square)$ sentence β

$T \vdash \varphi \to \beta$ iff $T \vdash \xi \to \beta$, and

$T \vdash \beta \to \varphi$ iff $T \vdash \beta \to \xi$,

which, when properly formalized, means that $(T; \varphi)$ and $(T; \xi)$ are $\mathcal{B}(\square)$-coherent. Next, (either one of) (1) and (2) also imply that

$T \vdash \varphi \leftrightarrow \xi$ iff for some $\mathcal{B}(\square)$ sentence β, $T \vdash \xi \leftrightarrow \beta$.

We have assumed the r.h.s. not to be the case. Therefore, $T \nvdash \varphi \leftrightarrow \xi$. By Proposition 4 we conclude the existence of an automorphism as described in the statement. ∎

8. QUESTION. Can automorphisms move $\square$ sentences?

References

[1] L. D. Beklemishev. Induction rules, reflection principles, and provably recursive functions. *Logic Group Preprint Series No. 146.* Department of Philosophy, Utrecht University 1995.

[2] P. Clote, P. Hájek & J. Paris. On some formalized conservation results in arithmetic. *Archive for Mathematical Logic* 30 (1990) 201–218.

[3] P. Hájek & P. Pudlák. *Metamathematics of First-Order Arithmetic.* Springer-Verlag, Berlin 1993.

[4] P. Lindström. On partially conservative sentences and interpretability. *Proceedings of the American Mathematical Society* 91 (1984) 436–443.

[5] M. B. Pour-El & S. Kripke. Deduction-preserving "recursive isomorphisms" between theories. *Fundamenta Mathematicae* 61 (1967) 141–163.

[6] V. Yu. Shavrukov. A note on the diagonalizable algebras of PA and ZF. *Annals of Pure and Applied Logic* 61 (1993) 161–173.

[7] V. Yu. Shavrukov. Undecidability in diagonalizable algebras. *The Journal of Symbolic Logic*, to appear.

[8] J. R. Shoenfield. A relative consistency proof. *The Journal of Symbolic Logic* 19 (1954) 21–28.

[9] W. Sieg. Herbrand analyses. *Archive for Mathematical Logic* 30 (1991) 409–441.

[10] C. Smoryński. Fifty years of self-reference in arithmetic. *Notre Dame Journal of Formal Logic* 22 (1981) 357–374.

[11] C. Smoryński. Fixed point algebras. *Bulletin (New Series) of the American Mathematical Society* 6 (1982) 317–356.

[12] R. M. Solovay. Infinite fixed-point algebras. *Recursion Theory* (A. Nerode & R. A. Shore, eds.). American Mathematical Society, Providence 1985, 473–486.