Categories of Theories and Interpretations

Albert Visser*

April 18, 2004

Abstract

In this paper we study categories of theories and interpretations. In these categories, notions of *sameness of theories*, like synonymy, bi-interpretability and mutual interpretability, take the form of isomorphism.

We study the usual notions like monomorphism and product in the various theories. We provide some examples to separate notions across categories. In contrast, we show that, in some cases, notions in different categories *do* coincide. E.g., we can, under such-and-such conditions, infer synonymity of two theories from their being equivalent in the sense of a coarser equivalence relation.

We illustrate that the categories offer an appropriate framework for conceptual analysis of notions. For example, we provide a 'coordinate free' explication of the notion of axiom scheme. Also we give a closer analysis of the object-language/ meta-language distinction.

Our basic category can be enriched with a form of 2-structure. We use this 2-structure to characterize a salient subclass of interpetations, the direct interpretations, and we use the 2-structure to characterize induction. Using this last characterization, we prove a theorem that has as a consequence that, if two extensions of Peano Arithmetic in the arithmetical language are synonymous, then they are identical.

Finally, we study preservation of properties over certain morphisms.

Contents

*I thank Lev Beklemishev, Harvey Friedman, Spencer Gerhardt, Volker Halbach, Joost Joosten, Benedikt Löwe, Vincent van Oostrom, Darko Sarenac for enlightening conversations. I am grateful to Ieke Moerdijk for giving me some explanation about fibrations. John Corcoran provided historical background in e-mail correspondence, for which I am grateful. I thank Panu Raatikainen for sharing his ideas on interpretations and the definability of truth with me. Panu also provided many pointers to the relevant literature.

1 Introduction

Interpretations are ubiquitous in mathematics and logic. Some of the greatest achievements of mathematics, like the internal models of non-euclidean geometries are, in essence, interpretations.

Given the importance of interpretations, it would seem that there is some room for a systematic study of interpretations and interpretability as objects in their own right. This paper is an attempt to initiate one such line of enquiry. It is devoted to the study of the category of interpretations, or, more precisely the study of a sequence of categories of interpretations.

Below, I will briefly address three issues: motivation & desiderata, comparison to some earlier work, and comparison to boolean morphisms.

1.1 Motivation & Desiderata

The fact that interpretations play an important role in mathematics does not ipso facto mean that we should study them in a systematic way. Perhaps, as a totality, they are too diverse to make systematic study sensible. Perhaps, the only general insights are trite and not very useful. The work in this paper certainly does not bring the subject far enough to exclude such pessimistic expectations. The paper should be viewed in an experimental spirit: it is at least worth the effort to pursue such a study up to some level.

Interpretations have several uses. First, *comparing theories* is a philosophical need of human beings. E.g., we want to explicate the notion of strength of a theory. One possible explication is in terms of degrees of interpretability.[1] Or, we just want to say of certain theories that they are essentially the same. Consider, for example, the theory of partial order involving the weak ordering relation versus a formulation involving the strong ordering. Many people would judge these versions to embody the same theory as a matter of course. But what does sameness mean here? One possible explication is *synonymy* (see [dB65a], [dB65b], [Cor80], [Kan72]), which, as will be shown in this paper, can be best viewed as isomorphism in a certain category of interpretations.

[1] Of course, there are other explications. The most succesful one is Π^0_2-conservativity: a theory is stronger if it proves more Π^0_2-sentences. Note, however, that conservativity implicitly uses interpretations: in many examples the designated class of sentences (e.g. Π^0_2) is not really a class of sentences of the language (e.g. the language of set-theory), but is embedded via an interpretation. Thus, the 'objects' to which we ascribe conservativity w.r.t. a class of sentences Γ are really pairs $\langle T, \tau \rangle$, where τ is an translation of the language containing Γ into the language of T.

Conservativity and interpretability diverge in some cases: GB is conservative w.r.t. the full language of set theory over ZF. However, since GB proves the consistency of ZF on a definable cut, GB is not interpretable in ZF.

Secondly, there are mathematico-logical applications. Interpretations can be used to prove properties of theories. E.g., Tarski proves that group theory is undecidable because true arithmetic can be interpreted in an extension of group theory. (See [TMR53].) Certain interpretations preserve important properties. E.g., consistency is preserved by interpretations in the reverse direction: from interpreting theory to interpreted theory. Decidability is preserved by *faithful* interpretations in the reverse direction.[2]

Thirdly, interpretations are an essential ingredient of other notions. For example, the notion of conservativity (implicitly) employs interpretations. Also notions of axiom scheme and rule employ interpretations, e.g. when we say that set theory enjoys full induction. See Section 5. A third example is the relation of object-theory and meta-theory. See Section 7.

A good theory of interpretations should do some justice to the various uses of interpretations. So, we want to be able to analyse notions of strength and notions of sameness. We also want to be able to analyse notions like axiom scheme and object-theory/meta-theory in our setting. Moreover, we want to develop a theory of preservation of properties along interpretations. In this paper, we will pay special attention to preservation over retractions (in a certain category). (See Section 10.) In connection with questions of preservation, it is important that we are able to distinguish *sorts of interpretation*. We often want to know, not just that there is an interpretation, but that *a certain kind of interpretation* (like an isomorphism or a retraction) holds between two theories. Further information will follow from the fact that the interpretation is of such-and-such a kind.[3]

One further desideratum is simply that (known) salient reasoning about interpretations can be reconstructed within the categorical framework. We did some work in this direction, e.g. in Sections 8 and 9.

We opted for developing *a category of interpretations*. Two clear internal desiderata are the following. (a) As many notions as possible should receive categorical definitions. (b) There should be interesting uses of category theory. With regard to desideratum (a), we followed a middle road. A number of important concepts like *extension of theories in the same language* are treated as *enrichments* of the category, rather than somehow defined. A nice example of a characterization in categorical terms is provided by Theorem 6.4, where we characterize *direct interpretations* in terms of discrete fibrations (defined in a certain 2-category). With regard to (b), certainly more should be done. We only use quite elementary category theory.

[2]Note however, that we do not necessarily want to uncritically maximize preservation of properties. For certain purposes, we might wish to have an equivalence relation over which the properties from a given class of properties $\mathcal{P}$ are preserved, but such that every equivalence class contains an element having a designated desirable property Q. Clearly, we do not wish Q to be preserved. For an example, see [Per97], where Q is *finite axiomatizability* and the equivalence relation is *semantical similarity w.r.t.* $\mathcal{P}$.

[3]An important question in this connection, which is left unresolved in this paper, is formulated in appendix A, item (8).

Some caveats w.r.t. the present project are in order. First there is the question how the enterprise relates to model theory. It is certainly true that one can think of lots of connections. However, it is important to realize that the present approach is too narrow to constitute something like a framework for applications of interpretations in model theory. The reader is referred to the discussion of interpretations in [Hod93] where this point is made very clear. A second caveat is that I am neither a model theorist or a category theorist. Undoubtedly, I will have missed some questions and methods that would have come naturally for a specialist in one of these subjects. For example, since most of my work concerned arithmetical theories, there is more attention for the wild world in this paper than for the tame stuff so dear to model theorists.

1.2 Comparison to Earlier Work

There is a considerable literature on the systematic study of interpretations. There are two main approaches, which will be presented in the next to subsections.

1.2.1 Interpretability Logic

There is the study of modal logics for interpretability. These logics are extensions of provability logic. There are two survey articles [JdJ98] and [Vis98]. Since these survey articles are in part complimentary they can very well be read together.

The main focus of the study of interpretability logics is to characterize schematic principles of interpretability that can be verified in theories that contain a sufficient amount of arithmetic. This line of research has yielded some beautiful results and some great techniques. It is however clear that it only treats a rather restricted range of theories. Moreover, it offers no possibility to talk explicitly about specific interpetations. Also there are some questions that it simply cannot address. E.g., it does not provide tools to make *distinctions* among interpretations.

1.2.2 Degrees of Interpretability

Degree structures of interpretability have been extensively studied. See, for example, [Šve78], [Ben86], [Lin97], [MPS90] and [Vis02].[4] This study yielded many remarkable results. As in the case of interpretability logic, many results were proven for theories with a lot of coding machinery. However, this limitation is an accident of interest of the researchers, not intrinsic to the subject. Apart from orderings of interpretability, also orderings of faithful interpetability and of local interpretability[5] are studied. For local interpretability, see [MPS90].

[4] The paper [MPS90] has an excellent introduction that in some respects supplements our introduction.

[5] A theory U is locally interpretable in a theory V iff every finite subtheory of U is interpretable in V.

Even if, for many purposes, interpretability is a rather crude relation, still there are some very useful preservation properties.

- Interpretability preserves inconsistency (from interpreted theory to interpreting theory), and, hence, preserves consistency in the reverse direction.
- A theory U is *reflexive* if it interprets a suitable weak arithmetical theory, like Buss's S^1_2, plus all statements of the form $\mathsf{con}_n(U)$, where $\mathsf{con}_n(U)$ expresses the consistency of the first n axioms of U for provability involving formulas of complexity below n. Here the measure of complexity is depth of quantifier changes. One can show that reflexive theories are not finitely axiomatizable. We have: mutual interpretability preserves reflexivity.
- A theory is *essentially undecidable* iff every consistent extension of it is undecidable. We have; interpretability preserves essential undecidability from interpreted theory to interpreting theory.
- Interpretability does *not* preserve decidability from interpreting theory to interpreted theory. Not even mutual interpretability preserves decidability. However, faithful interpretability does in the direction from interpreting theory to interpreted theory.

On the other hand, many properties are not preserved, or are at least not obviously preserved. Moreover, degree theory gives us no tool to make distinctions between different interpretations of a theory U in a theory V.

In this paper, we wil consider degree theory as a 'limiting case' of a category of interpretations. It will appear as our category INT_4.

1.3 Recursive Boolean Morphisms

Recursive boolean morphisms are recursive morphisms of the Lindenbaum algebras of theories considered as numerated objects. Recursive boolean morphisms are extensively and deeply studied. See [Han65], [PEK67], [Per97] and [Háj70][6]. We highlight two of the main results. Pour-El and Kripke show that all theories into which Q is interpretable are recursively boolean isomorphic. (See [PEK67].) The result continues to hold when we replace *recursive* by *primitive recursive* or even *elementary*. Peretyat'kin proves a theorem that implies that every recursively axiomatizable theory without finite models is recursively boolean isomorphic with a finitely axiomatized theory. (See [Per97]).

Every relative interpretation gives rise to a recursive boolean morphism, but not vice versa. Isomorpism in each of our categories INT_i, for $i = 0, 1, 2, 3$, will be a refinement of recursive boolean isomorphism. However, mutual relative interpretability and recursive boolean isomorphism are incomparable. E.g., Q and PA are, by the result of Pour-El and Kripke, recursively boolean isomorphic,

[6] Hájek studies a category with morphisms which are a sort of generalized boolean morphisms.

but not mutually relatively interpretable. Conversely, predicate logic with only unary predicate symbols and predicate logic with at least one more-than-unary predicate symbol (distinct from identity) are mutually interpretable, but not recursively boolean isomorphic. (Recursive boolean isomorphisms preserve decidability.) We will give an example of two theories that are mutually faithfully interpretable, but not recursively boolean isomorphic. See Subsubsection 4.8.4.

Peretyat'kin, in his book [Per97], studies an equivalence relation, *semantical similarity* w.r.t. a list of model-theoretic properties, that is finer than recursive boolean isomorphism. Roughly, his notion is recursive boolean isomorphism plus the demand that the model-theoretic properties from the list are preserved. None of our notions of isomorphism can fulfil the desiderata for Peretyat'kin's notion: all our notions are refinements of mutual relative interpretability. However, PA cannot be mutually interpretable with a finitely axiomatized theory.

2 Translations and Interpretations

In this section, we sketch the basic framework.[7] In some respects, the framework introduced here is too limited. We will briefly discuss three natural extensions of the framework in appendix B. We will not pursue these extensions in this paper, even if, in many cases, it is rather obvious how things would go.

2.1 Predicate Logic

Officially, we will consider only relational languages for predicate logic of finite signature. Unofficially, we will also consider languages with terms. These languages can be translated using a standard algorithm to corresponding relational languages.

A signature Σ is a triple $\langle \mathsf{Pred}, \mathsf{ar}, E\rangle$, where Pred is a finite set of predicate symbols, where $\mathsf{ar} : \mathsf{Pred} \to \omega$ is the arity function. and where E is a binary predicate representing the identity. We will often write '$=$' for: E.

We assume we are given a fixed ω-ordered sequence of variables $v_0, v_1, \ldots$. We will use $x, y, x_0, \ldots$ as metavariables ranging over variables. (We will follow the usual convention that, if e.g. "v_3", "x" and "y" are used in one formula, they are supposed to be distinct.) Formulas and sentences based on Σ and $v_0, v_1, \ldots$ are defined in the usual way.

A theory of signature Σ will be given by a set of sentences of the signature, the axioms. Derivability from the theory employs the axioms and the rules of predicate logic including the identity axioms and rules for E. We will assume that the axiom set of our theories is appropriately simple, say p-time decidable. However, most of the results of this paper will hold for more complicated axiom sets too.

[7] A much better treatment of the predicate logical language is possible, using sharing graphs. Under this alternative treatment, a lot of *ad hoc* choices and unpleasant details concerning α-conversion, would simply disappear. However, setting things up in the alternative way would take a lot of space and would detract from the proper subject of this paper.

2.2 Relative Translations

Let Σ and Θ be signatures. A *relative translation* $\tau : \Sigma \to \Theta$ is given by a pair $\langle \delta, F \rangle$. Here δ is a Θ-formula representing the *domain* of the translation. We demand that δ contains at most v_0 free. The mapping F associates to each relation symbol R of Σ with arity n an Θ-formula $F(R)$ with variables among $v_0, \ldots, v_{n-1}$.

We translate Σ-formulas to Θ-formulas as follows:

- $(R(y_0, \cdots, y_{n-1}))^\tau := F(R)(y_0, \cdots, y_{n-1})$;
  here $F(R)(y_0, \cdots, y_{n-1})$ is our sloppy notation for:

  $$F(R)[v_0 := y_0, \cdots, v_{n-1} := y_{n-1}],$$

  the result of substituting the y_i for the v_i; we assume that some mechanism for α-conversion is built into our definition of substitution to avoid variable-clashes;
- $(\cdot)^\tau$ commutes with the propositional connectives;
- $(\forall y\, A)^\tau := \forall y\, (\delta(y) \to A^\tau)$;
- $(\exists y\, A)^\tau := \exists y\, (\delta(y) \wedge A^\tau)$.

Suppose τ is $\langle \delta, F \rangle$. Here are some convenient notations.

- We write δ_τ for δ and F_τ for F.
- We write R_τ for $F_\tau(R)$.
- We write $\vec{x} : \delta$ for: $\delta(x_0) \wedge \ldots \wedge \delta(x_{n-1})$.
- We write $\forall \vec{x} : \delta\ A$ for: $\forall x_0 \ldots \forall x_{n-1}\ (\vec{x} : \delta \to A)$.
- We write $\exists \vec{x} : \delta\ A$ for: $\exists x_0 \ldots \exists x_{n-1}\ (\vec{x} : \delta \wedge A)$.

We can compose relative translations as follows:

- $\delta_{\tau\nu} := (\delta_\nu \wedge (\delta_\tau)^\nu)$,
- $R_{\tau\nu} = (R_\tau)^\nu$.

We write $\nu \circ \tau := \tau; \nu := \tau\nu$. Note that $(A^\tau)^\nu$ is provably equivalent in predicate logic to $A^{\tau\nu}$. The identity translation $\mathsf{id} := \mathsf{id}_\Theta$ is defined by:

- $\delta_{\mathsf{id}} := (v_0 E v_0)$,
- $R_{\mathsf{id}} := R(v_0, \ldots, v_{n-1})$.

Note that translations as defined here only have good properties modulo provable equivalence. E.g., $\delta_{\mathsf{id} \circ \mathsf{id}} = (v_0 E v_0 \wedge v_0 E v_0)$, which is not strictly identical to δ_{id}.

Consider a relative translation $\tau : \Sigma \to \Theta$. Let $\mathcal{M} = \langle M, I \rangle$ be a model of signature Θ. Suppose $\mathcal{M}$ satisfies $\exists v_0\ \delta$ and the τ-translations of the identity axioms in the Σ-language. We can define a model $\mathcal{N} := \tau^{\mathcal{M}}$ as follows.

- Clearly, E^τ_Σ defines an equivalence relation, say $\simeq$, in $\mathcal{M}$ on $N_0 := \{m{\in}M \mid \mathcal{M} \models \delta_\tau(m)\}$. The domain N of $\mathcal{N}$ is $N_0/\simeq$.
- The relation $\simeq$ is a congruence w.r.t. the relations R, given by:

$$R(\vec{m}) :\Leftrightarrow \mathcal{M} \models P^\tau(\vec{m}).$$

Thus, it makes sense to define, for $\vec{n}$ a sequence of elements of N,

$$P^{\mathcal{N}}(\vec{n}) :\Leftrightarrow \exists m_0{\in}n_0 \ldots \exists m_{k-1}{\in}n_{k-1}\ R(m_0, \cdots, m_{k-1}).$$

Suppose $\vec{n}$ is a sequence of elements of N, the domain of $\mathcal{N} := \tau^{\mathcal{M}}$. Let $\vec{m}$ be a sequence of elements of M, the domain of $\mathcal{M}$, such that $m_i \in n_i$. One can show:

$$\mathcal{N} \models A(\vec{n}) \Leftrightarrow \mathcal{M} \models A^\tau(\vec{m}).$$

Further, one can show that ${\tau^\nu}^{\mathcal{M}}$ is isomorphic to $(\tau\nu)^{\mathcal{M}}$ (if defined). We have, e.g., for sentences A of the appropriate signature:

$$\begin{array}{rcl} {\tau^\nu}^{\mathcal{M}} \models A & \Leftrightarrow & \nu^{\mathcal{M}} \models A^\tau \\ & \Leftrightarrow & \mathcal{M} \models (A^\tau)^\nu \\ & \Leftrightarrow & \mathcal{M} \models A^{\tau\nu} \\ & \Leftrightarrow & (\tau\nu)^{\mathcal{M}} \models A. \end{array}$$

2.3 Relative Interpretations

A translation τ supports a *relative interpretation* of a theory U in a theory V, if, for all U-sentences A, $U \vdash A \Rightarrow V \vdash A^\tau$. (Note that this automatically takes care of the theory of identity. Moreover, it follows that $V \vdash \exists v_0\ \delta_\tau$.) We will write $K = \langle U, \tau, V\rangle$ for the interpretation supported by τ.

Suppose T has signature Σ and $K : U \to V$, $M : V \to W$. We define:

- $\mathsf{id}_T : T \to T$ is $\langle T, \mathsf{id}_\Sigma, T\rangle$,
- $M \circ K : U \to W$ is $\langle U, \tau_M \circ \tau_K, W\rangle$.

We identify two interpretations $K, K' : U \to V$ if:

- $V \vdash \delta_K \leftrightarrow \delta_{K'}$,
- $V, \vec{v} : \delta \vdash P^K \leftrightarrow P^M$, where $\mathsf{ar}(P) = n$ and $\vec{v} = v_0, \ldots, v_{n-1}$.

One can show that modulo this identification, the above operations give rise to a category of interpretations that we call INT.

When we speak about K, ... we will mean the objects of the form $\langle U, \tau, V\rangle$ and not the equivalence classes. We will think of these objects modulo the notion of equality defined above, but, e.g. the notation 'P^K' calls for the triple and not the corresponding equivalence class.

If $K = \langle U, \tau, V\rangle$, and $\mathcal{M} \models V$, we will often write $K^{\mathcal{M}}$ for $\tau^{\mathcal{M}}$.

We assign to a theory T its class of models $\mathsf{MOD}(T)$. Consider $K : U \to V$. This interpretation gives rise to the mapping $\mathsf{MOD}(K) : \mathsf{MOD}(V) \to \mathsf{MOD}(U)$ given by $\mathcal{M} \mapsto K^{\mathcal{M}}$. We have:

- $\mathsf{MOD}(\mathsf{id}_T) := \mathsf{id}_{\mathsf{MOD}(T)}$,
- $\mathsf{MOD}(M \circ K) = \mathsf{MOD}(K) \circ \mathsf{MOD}(M)$.

Thus, MOD gives rise to a contravariant functor from INT to the category of sets and classes.

3 Categories of Interpretations

In this section, we introduce various categories of interpretations. Our basic category is INT. We call two interpretations that implement the same morphism in the sense of INT: *equal*. We call theories that are isomorphic in this category: *synonymous* or *definitionally equivalent*.[8]

3.1 Definable Maps between Interpretations

We extend INT with extra structure. In this enriched category, $\mathsf{INT}^{\mathsf{morph}}$, the arrows between two objects play themselves the role of objects in a category as follows. Consider $K, M : U \to V$. An arrow $F : K \Rightarrow M$ is a V-definable, V-provable morphism from K to M considered as 'parametrized internal models'. Specifically, this means that a morphism from K to M is given as a triple $\langle K, F, M\rangle$, where F is a formula with the following properties.

- The free variables of F are among v_0, v_1. We write $F(x, y)$ or xFy, for: $F[v_0 := x, v_1 := y]$.
- $V \vdash xFy \to (x : \delta_K \wedge y : \delta_M)$.
- Writing E for the identity relation:
  $V \vdash (x : \delta_K \wedge y : \delta_M \wedge xE_Kx'Fy'E_My) \to xFy$.
- $V \vdash \forall x : \delta_K \, \exists y : \delta_M \; xFy$.
- $V \vdash (xFy \wedge xFy') \to yE_My'$.
- $V \vdash \vec{x}F\vec{y} \to (P_K\vec{x} \to P_M\vec{y})$.[9]
  Here '$\vec{x}F\vec{y}$' abbreviates $x_0Fy_0 \wedge \ldots \wedge x_{n-1}Fy_{n-1}$, for appropriate n.

[8] Karel de Bouvère uses *synonymy* in his [dB65a] and [dB65b]. Stig Kanger, John Corcoran and Wilfrid Hodges use *definitional equivalence*. See [Kan72], [Cor80] or [Hod93]. Mikhail Peretyat'kin uses *isomorphism*. See [Per97].

[9] Note that if P represents a function in U, then, by elementary reasoning, we have: $V \vdash \vec{x}F\vec{y} \to (P_K\vec{x} \leftrightarrow P_M\vec{y})$.

We will call the arrows between interpretations: *i-maps*. We consider $F, G : K \Rightarrow M$ as *equal* when they are V-provably the same. The identity $\mathsf{ID}_K : K \Rightarrow K$ is given by: $v_0(\mathsf{ID}_K)v_1 :\leftrightarrow v_0, v_1 : \delta_K \wedge v_0 E_K v_1$. If $A : K \Rightarrow L$ and $B : L \Rightarrow M$, then $B \cdot A : K \Rightarrow M$ is the obvious composition of B and A.

An isomorphism of interpretations is easily seen to be a morphism with the following extra properties.

- $V \vdash \forall y : \delta_M \, \exists x : \delta_K \; xFy$,
- $V \vdash (xFy \wedge x'Fy) \rightarrow xE_K x'$,
- $V \vdash \vec{x}F\vec{y} \rightarrow (P_M\vec{y} \rightarrow P_K\vec{x})$.

Suppose we have the constellation depicted in the diagram below.

$$U \;\underset{K}{\overset{L}{\Uparrow F}}\; V \xrightarrow{\;M\;} W$$

We define $M \circ F : M \circ K \Rightarrow M \circ L$ as follows:

- $v_0(M \circ F)v_1 :\leftrightarrow v_0, v_1 : \delta_M \wedge v_0 F^M v_1$.

We may show that, indeed, $M \circ F : M \circ K \Rightarrow M \circ L$.

The corresponding idea of composing F with an interpretation to the right does not generally make sense: an i-map is analogous to a morphism between structures. Such morphisms do not generally induce morphisms between 'substructures with defined operations'. It does only make sense if we put suitable constraints on the interpretations or if we restrict ourselves to i-isomorphisms. We will follow this last strategy in the present paper. Suppose we have the situation depicted in the diagram below. Let G be an i-isomorphism.

$$U \xrightarrow{\;K\;} V \;\underset{L}{\overset{M}{\Uparrow G}}\; W$$

We define $G \circ K : L \circ K \Rightarrow M \circ K$ as follows:

- $v_0(G \circ K)v_1 :\leftrightarrow v_0 : \delta_{L\circ K} \wedge v_1 : \delta_{M\circ K} \wedge \exists x, y \;\; v_0 E_{L\circ K} x G y E_{M\circ K} v_1$.

We can easily show that: $G\circ K : L\circ K \Rightarrow M\circ K$, where $G\circ K$ is an i-isomorphism. Suppose we have the following situation, where G is an i-isomorphism.

$$U \;\underset{K}{\overset{L}{\Uparrow F}}\; V \;\underset{M}{\overset{N}{\Uparrow G}}\; W$$

We define $G \circ F : M \circ K \Rightarrow N \circ L$ as follows:

- $G \circ F := (N \circ F) \cdot (G \circ K)$.

One may show that: $G \circ F = (G \circ L) \cdot (M \circ F)$.

We will call the restriction of $\mathsf{INT}^{\mathsf{morph}}$ to i-isomorphisms: $\mathsf{INT}^{\mathsf{iso}}$. We can show that $\mathsf{INT}^{\mathsf{iso}}$ is 2-category with the defined operations. See [Mac71] and [Bor94] for an explanation.

Here is a pleasant way to look at the properties of our enriched category $\mathsf{INT}^{\mathsf{morph}}$. Let us fix some theory, say $\mho$. We define a 2-functor $[\![\cdot]\!]$ (or, more explicitly $[\![\cdot]\!]_{\mho}$) from $\mathsf{INT}^{\mathsf{iso}}$ to the 2-category of categories, functors and natural transformations as follows.

- $[\![U]\!]$ is the category with as objects the interpretations $M : \mho \to U$ and as arrows the i-maps between the objects.
- Suppose $K : U \to V$, we define $[\![K]\!] : [\![U]\!] \to [\![V]\!]$, by $[\![K]\!](M) := K \circ M$, $[\![K]\!](F) := K \circ F$.
- Suppose $G : K \Rightarrow K'$ is an i-isomorphism. Then, $[\![G]\!] : [\![K]\!] \Rightarrow [\![K']\!]$ is given by: $[\![G]\!](M) := G \circ M$.

The verification that we have indeed defined a 2-functor holds no surprises. Similarly, we can define a contravariant 2-functor $(\![\cdot]\!)$ from $\mathsf{INT}^{\mathsf{iso}}$ to the 2-category of categories, functors and natural transformations, as follows. Again we fix a theory $\mho$.

- $(\![U]\!)$ is the category with as objects the interpretations $M : U \to \mho$ and as arrows the i-isomorphisms between the objects.
- Suppose $K : U \to V$, we define $(\![K]\!] : (\![V]\!] \to (\![U]\!]$, by $(\![K]\!](M) := M \circ K$, $(\![K]\!](F) := F \circ K$.
- Suppose $G : K \Rightarrow K'$ is an i-isomorphism. Then, $(\![G]\!] : (\![K]\!] \Rightarrow (\![K']\!]$ is given by: $(\![G]\!](M) := M \circ G$.

Note that the definition of $(\![\cdot]\!)$ also makes sense on $\mathsf{INT}^{\mathsf{morph}}$. Only we get no 2-functor since $\mathsf{INT}^{\mathsf{morph}}$ is not a 2-category.

3.2 Subcategories

We may wish to restrict our interpretations to certain subclasses. E.g., we restrict ourselves to unrelativized interpretations, obtaining the category $\mathsf{INT}_{\mathsf{unr}}$. Or we restrict ourselves to interpretations that interpret identity by itself, obtaining the category $\mathsf{INT}_{=}$. An important case is the restriction to unrelativized interpretations that preserve identity. Such interpretations (and the corresponding morphisms) we call: *direct*. The associated category is $\mathsf{INT}_{\mathsf{unr},=}$.

3.3 When are Two Interpretations the Same?

The most important variations on INT are obtained by considering cruder notions of equality on interpretations.

1. The finest identification that we will consider is equality of interpretations as defined above. For reasons of systematicity we will also call this equality: equal_0 or $=_0$. Similarly, we sometimes call the category INT: INT_0.

2. The next level of identification is i-isomorphism of interpretations in our extended category $\mathsf{INT}^{\mathsf{morph}}$. We will call i-isomorphism also: equal_1 or $=_1$. Wilfrid Hodges, in his [Hod93], uses *homotopy* for i-isomorphism and calls i-isomorphic interpretations *homotopic*. We will call the category of theories with the morphisms so obtained: INT_1 or hINT.

3. We may also consider two interpretations $K : U \to V$ and $M : U \to V$ as the same iff, for all models $\mathcal{M} \in \mathsf{MOD}(V)$, we have that $\mathsf{MOD}(K)(\mathcal{M})$ is isomorphic to $\mathsf{MOD}(M)(\mathcal{M})$. We call this equality: equal_2 or $=_2$. We will call the category of theories with the morphisms so obtained: INT_2 or whINT.

   There are all kinds of variants of this category which can be obtained by restricting the class of models.

4. We may take two interpretations $K : U \to V$ and $M : U \to V$ as the same iff, for all U-sentences A, we have $V \vdash A^K \leftrightarrow A^M$. We will say that these interpretations are equivalent, equal_3 or $=_3$. We will call the category of theories with the morphisms so obtained: INT_3 or eqINT.

5. We can simply identify all interpretations between U and V. In this case we obtain the preorder associated with the degrees of interpretability. We will call the category of theories with the morphisms so obtained: INT_4 or DEG.

The variables K, L, ... will always range over triples of the form $\langle U, \tau, V\rangle$ and not over the corresponding equivalence classes. We think of these triples modulo the appropriate notion of equality in the appropriate context. We will write things like: 'K is a monomorphism in INT_0, but not in INT_1', thus avoiding explicitly mentioning the embedding functors between our various categories.

We note a coincidence of categories.

Theorem 3.1 *The category INT_3 coincides with $\mathsf{INT}_2^{\mathsf{rs}}$, i.e. the category INT_2 with the class of models restricted to recursively saturated models.*

Proof

The proof is immediate from two well-known facts. Definable internal models of recursively saturated models are isomorphic iff they are elementary equivalent. Moreover, any consistent theory has a recursively saturated model. ❑

Our various categories induce different notions of sameness on theories:

1. two theories are *synonymous* or *definitionally equivalent* if they are isomorphic in INT;
2. two theories are *bi-interpretable*, if they are isomorphic in INT_1;
3. two theories are *weakly bi-interpretable* if they are isomorphic in INT_2;
4. two theories are *sententially equivalent* if they are isomorphic in INT_3;
5. two theories are *mutually interpretable* if they are isomorphic in INT_4.

Note that, due to the simple form of the definition of isomorphism, the more interpretations we identify, the more theories are isomorphic. So, each subsequent notion is sameness of theories is cruder.

3.4 Generalizing the MOD-functor

In Subsection 2.3, we introduced the contravariant MOD-functor on INT. Consider $K =_i M : U \to V$ for $i = 1, 2$. Note that, for any model $\mathcal{M}$ of V, we have that the model $\mathsf{MOD}(K)(\mathcal{M})$ need not be the same as the model $\mathsf{MOD}(M)(\mathcal{M})$, however these models will be isomorphic. Thus, to make sense of the MOD-functor on INT_1 and INT_2, we should consider it as a functor to models modulo isomorphism.

Clearly, there is a cardinality problem here, since isomorphism classes will not be sets. There are all kinds of ways to get around that problem. E.g., we can stipulate that we only consider models in V_κ for some cardinal κ, or we can refrain from dividing out the equivalence relation, or we can demand the elements of the domains of the models are chosen from the cardinal that is the cardinality of the domain, etc. We pretend that we have settled on some such solution.

We will call the new functors MOD_i, for $i = 1, 2$, or, if no confusion is possible, simply MOD. For uniformity, sometimes we will call our original MOD-functor: MOD_0.

In the case of INT_3, we only get, for $K =_3 M : U \to V$, that $\mathsf{MOD}(K)(\mathcal{M})$ is elementarily equivalent to $\mathsf{MOD}(M)(\mathcal{M})$. Thus, in the case of INT_3, we take the MOD_3-functor to give us models modulo elementary equivalence, or alternatively: complete theories.

For INT_4 we will not have a MOD-functor.

3.5 Factorization

A slightly odd aspect of interpretations and the corresponding morphisms in our categories is the fact that there is a certain favouritism towards the target theory, i.e. the interpreting theory. The interpreted theory often plays a minor role. E.g. if you look at the question whether two interpretations between U and V are the same, the answer only depends on the underlying translations

and the target theory. Similarly, i-morphisms are defined in terms of the underlying translations and target theories. In this subsection, we spell out some consequences of this lopsidedness.

We work in $\mathsf{INT}^{\mathsf{morph}}$. Consider an interpretation $K = \langle U, \tau, V\rangle$. We define the theory $\tau^{-1}[V]$ or $K^{-1}[V]$ as the theory in the signature of U given by $\{A{\in}\mathcal{S}_U \mid V \vdash A^\tau\}$. (We may find an efficient axiomatization using Craig's Trick.)

We write $\mathcal{E}_{TZ}$ for the identical interpretation witnessing that T is a subtheory of Z in the same language. Note that restricting the morphisms to the $\mathcal{E}_{TZ}$ will give us a subcategory. In Subsection 4.6, we will see that $\mathcal{E}_{UV}$ is an epimorphism in each of our categories.

Let $\breve{K} := \langle K^{-1}[V], \tau, V\rangle$. Then, we obviously have the following decomposition of K:

$$U \xrightarrow{\mathcal{E}_{U,K^{-1}[V]}} K^{-1}[V] \xrightarrow{\breve{K}} V$$

Clearly, $\breve{K}$ is faithful. In Subsection 4.5, we will see that faithful interpretations are monomorphisms in categories INT_0 and INT_3. Ergo, in INT_0 and INT_3, our factorization is an epi-mono factorization. We have the following obvious theorem.

Theorem 3.2 *Suppose we have $K : U \to V$, $M : W \to V$, $U \subseteq W$ and $K = M \circ \mathcal{E}_{UW}$. Then, $K^{-1}[V] = M^{-1}[V]$ and $\breve{K} = \breve{M}$.*

We omitt the proof.

Corollary 3.3 *Suppose $K : U \to V$ and $M : V \to W$. Then, $K^{-1}[V] \subseteq (M \circ K)^{-1}[W]$.*

Proof

Suppose $K : U \to V$ and $M : V \to W$. We have $(M \circ K) : U \to W$, $M \circ \breve{K} : K^{-1}[V] \to W$ and $U \subseteq K^{-1}[V]$. Hence, by Theorem 3.2,

$$K^{-1}[V] \subseteq (M \circ \breve{K})^{-1}[W] = (M \circ K)^{-1}[W].$$

❑

More generally, we can consider the interaction of an $\mathcal{E}$-arrow and two subsequent arrows. We get the following theorem.

Theorem 3.4 *Suppose the following diagram obtains:*

$$U \xrightarrow{\mathcal{E}_{UV}} V \underset{L}{\overset{K}{\Uparrow F}} W$$

We define $F \circ \mathcal{E}_{UV} : (L \circ \mathcal{E}_{UV}) \Rightarrow (K \circ \mathcal{E}_{UV})$*, where* $F \circ \mathcal{E}_{UV}$ *is given by the same formula as* F*.*[10] *We have:*

- $(\cdot) \circ \mathcal{E}_{UV}$ *is an injective and full functor from the category of arrows from* V *to* W *to the category of arrows from* U *to* W*;*
- $(F \circ \mathcal{E}_{UV}) \circ \mathcal{E}_{TU} = F \circ \mathcal{E}_{TV}$*;*
- *suppose* $M : W \to Z$*, then* $(M \circ F) \circ \mathcal{E}_{UV} = M \circ (F \circ \mathcal{E}_{UV})$*.*

Conversely, suppose we have the following situation.

$$U \;\underset{L \circ \mathcal{E}_{UV}}{\overset{K \circ \mathcal{E}_{UV}}{\Uparrow G}}\; W$$

Then, we can find a unique $\tilde{G} : L \Rightarrow K$ *such that* $G = \tilde{G} \circ \mathcal{E}_{UV}$*. Moreover* $\tilde{(\cdot)}$ *is a functor from the subcategory of arrows of the form* $P \circ \mathcal{E}_{UV} : U \to W$ *to the category of arrows from* V *to* W*.*

4 A Closer Look at Familiar Concepts

In this section, we treat some evident 'homework questions'. If you have a category, you want to know whether there are, e.g., products and, if so, what they are. We do not try to be exhaustive here: we just treat some nice bits.

4.1 Initial Objects

The theory **1** is the theory in the language with just identity, which states that there is precisely one object. It is the initial object in INT_i with $i \in \{1, 2, 3, 4\}$, i.e. it has precisely one interpretation (modulo the chosen notion of equivalence) in any other theory.

The situation is different for INT_0. Consider the theory U with unary predicate symbol P and axiom:

- $\exists x, y\, (P(x) \wedge \neg P(y) \wedge \forall z\, (z = x \vee z = y)\,)$.

We can interpret **1** into U via K_0 given by:

- $\delta_{K_0} :\leftrightarrow P(v_0)$,
- $E_{K_0} :\leftrightarrow v_0 = v_1$.

We can also employ K_1 given by:

- $\delta_{K_1} :\leftrightarrow \neg P(v_0)$,

[10] Remember that, previously, we only admitted $F \circ M$ in case F was an i-isomorphism.

- $E_{K_1} :\leftrightarrow v_0 = v_1$.

Suppose J would be an initial object of INT_0 and $M : J \to \mathbf{1}$. Then, it is easily seen that $K_0 \circ M : J \to U$ cannot be equal to $K_1 \circ M : J \to U$. A contradiction. Of course there are many *weak* initial objects in INT.

When we restrict INT_i with $i \in \{1, 2, 3, 4\}$ to interpretations that send identity to identity. the situation also changes. Consider $\Im := \mathsf{FOL}_{\mathsf{id}}$, predicate logic with just identity. Suppose J is an initial object in one of the $\mathsf{INT}_{i,=}$, for $i = 1, 2, 3, 4$. Then, there is a unique morphism $M : J \to \mathsf{FOL}_{\mathsf{id}}$. Note that, by a simple model theoretical argument, δ_M must be equivalent to $v_0 = v_0$. Now we can consider identity preserving interpretations K_0 and K_1 of $\mathsf{FOL}_{\mathsf{id}}$ into, say, PA such that $\mathsf{PA} \vdash \mathbf{1}^{K_0}$ and $\mathsf{PA} \vdash \mathbf{2}^{K_1}$. Here $\mathbf{2}$ is the theory in the languge of pure identity with as axiom the statement that there are precisely two objects. It is easy to see that $K_0 \circ M : J \to \mathsf{PA}$ cannot be equal to $K_1 \circ M : J \to \mathsf{PA}$.

For any theory T, we define the morphism $\mathcal{I}_T : \Im \to T$ by: $\delta_{\mathcal{I}_T} :\leftrightarrow v_0 = v_0$ and $v_0 E^{\mathcal{I}_T} v_1 :\leftrightarrow v_0 = v_1$. It follows that $\Im$ is a weak initial object in any of our categories. It is easy to see that $\Im$ is the initial object in $\mathsf{INT}_{i,\mathsf{unr},=}$, for $i = 0, 1, 2, 3, 4$.

4.2 End Objects

The inconsistent theory of any signature is the end object in any of our categories. Any theory has precisely one interpretation into the inconsistent theory.

If we restrict our categories to faithful interpretations, the situation changes dramatically: there is not even a *weak* end object.

4.3 The Cartesian Product

Surprisingly, there is a stable notion of product that works for all our categories. The easiest way to introduce the product is as follows. First we consider theories U and V of the same signature such that $U \cup V$ is inconsistent. By a simple compactness argument, we find that there is an A such that $U \vdash A$ and $V \vdash \neg A$. We will call A, a *separating* formula for U and V. (Note: a separating formula is assigned to the *ordered* pair U, V.)

Lemma 4.1 *Suppose U and V have the same signature and suppose $U \cup V$ is inconsistent. Let A be a separating formula for U and V. The theory $W := U \cap V$ can be axiomatized by the following sets of axioms: axioms of the form $A \to B$, where B is an axiom of U, and axioms of the form $\neg A \to C$, where C is an axiom of V.*

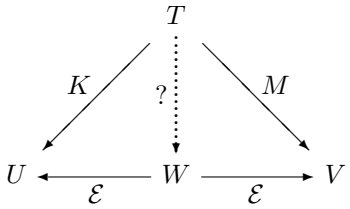
Note that it follows that we can write U as $W + A$ and V as $W + \neg A$.

Proof

It is clear that the axioms $(A \to B)$ and $(\neg A \to C)$ are in the intersection. Conversely, suppose $W \vdash D$. Then some conjuction β of U-axioms B proves D and some conjunction γ of V-axioms C proves D. We claim that $\delta := ((A \to \beta) \wedge (\neg A \to \gamma))$ proves D. This is immediate, since, ex hypothesi, δ implies $(A \to D) \wedge (\neg A \to D))$. Finally, δ can be rewritten to a conjunction of axioms of the form $(A \to B)$ and $(\neg A \to C)$. ❑

We claim that $W := U \cap V$ is the cartesian product of U and V in any of our categories. The projections are the standard embeddings $\mathcal{E}_{WU}$ and $\mathcal{E}_{WV}$.

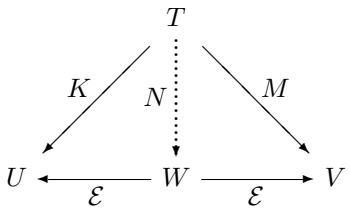
To show that $W = U \times V$ with the stated projections, we have to uniquely provide the dotted arrow that makes the following diagram commute.



Suppose the translations associated with K, M are τ, ρ. Let A be the chosen separating sentence for U and V. We define a new translation $\tau\langle A\rangle\rho$ as follows.

- $\delta_{\tau\langle A\rangle\rho} :\leftrightarrow ((A \wedge \delta_\tau) \vee (\neg A \wedge \delta_\rho))$
- $P_{\tau\langle A\rangle\rho} :\leftrightarrow ((A \wedge P_\tau) \vee (\neg A \wedge P_\rho))$

The interpretation $N := K\langle A\rangle M : T \to W$ is the interpretation corresponding to $\tau\langle A\rangle\rho$. We take:



In each of our categories we may easily verify that $K\langle A\rangle M$ supports the unique morphism that makes our diagram commute. This also shows that equality is a congruence for $(\cdot)\langle A\rangle(\cdot)$. Thus, $(\cdot)\langle A\rangle(\cdot)$ can be viewed as an operation on arrows. Note that at the level of the interpretations as mappings from models to models, the set of models of W is the disjoint union of the set of models of U and the set of models of V. We have: $\mathsf{Mod}(K\langle A\rangle M) = \mathsf{Mod}(K) \cup \mathsf{Mod}(M)$. (The union concerns the functions considered as sets of pairs.)

Here are some alternative notations for $K\langle A\rangle M$.

- My notation in earlier papers: $K[A]M$.
- Pseudo-code: `if` A `then` K `else` M.

- Hoare style: $K \lhd A \rhd M$ or $K \lhd A \rhd M$.

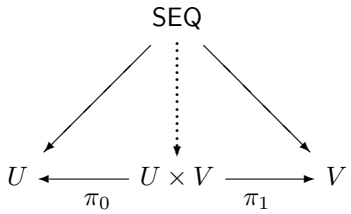
We defined the cartesian product for a special case. We extend the definition to arbitrary theories as follows. Let U and V be arbitrary with signatures Σ and Θ. Let $\Sigma \cup \Theta$ be the union of our signatures. To make sense of taking the union, it is best to consider the arities to be built in into the symbols, so that, say, a binary predicate P and a ternary predicate P are automatically counted as different predicates.

At this point we use Theorem 4.12, which says that a definitional extension of a theory is synonymous with that theory. (This theorem is really a triviality.) Consider definitional extensions U' and V' of U and V in the signature $\Sigma \cup \Theta$. In case $U' \cup V'$ is inconsistent, we take, as $U \times V$, the intersection $W' := U' \cap V'$. Our first projection becomes $\pi_0 := J \circ \mathcal{E}_{W'U'}$, where $J : U' \to U$ is the standard isomorphism associated with definitional extensions. Similarly, $\pi_1 := L \circ \mathcal{E}_{W'V'}$, where $L : V' \to V$ is an isomorphism. It is easy to see that we have defined a product in this way.

If $U' \cup V'$ is consistent, we extend the signature $\Sigma \cup \Theta$ with a fresh 0-ary predicate symbol P, and replace U' be the definitional extension $U' + P$ and V' by the definitional extension $V' + \neg P$. Now we may proceed as before.

Inspecting the construction, we see that it even works in the strict world of $\mathsf{INT}_{\mathsf{unr},=}$.

Remark 4.2 Sequentiality of a theory T is the existence of a morphism in $\mathsf{INT}_{\mathsf{unr},=}$ from a certain theory of sequences SEQ to T. See Section 10. Thus, it follows that sequentiality is preserved by products, as witnessed by the following commutative diagram in $\mathsf{INT}_{\mathsf{unr},=}$.



∎

It is an interesting exercise to compute $T^n := \overbrace{T \times \cdots \times T}^{n\times}$. The result of the exercise is the following theorem.

Theorem 4.3 *We work in any of our categories. Let $\vec{P}$ be a finite set of 0-ary predicates. Suppose $\phi(\vec{P})$ defines a finite boolean algebra with n atoms on generators $\vec{P}$. (There is just one finite boolean algebra with n atoms.) (If $n = 0$, we are in the degenerated case: we take $\phi := \bot$.)*

Consider a theory T with signature Σ and suppose that the symbols $\vec{P}$ do not occur in Σ. Then, we can take as T^n the theory in the signature $\Sigma + \vec{P}$ such that $T^n := T + \phi(\vec{P})$.

Proof

Suppose P_0, P_1, ... is a fixed infinite sequence of 0-ary predicates. We write $\vec{P}_{[n]}$ for $P_0, \ldots, P_{n-1}$. We first prove, by induction on n, that, for each n, we can find a specific ψ_n such that $T + \psi_n(\vec{P}_{[n]})$, in signature $\Sigma + \vec{P}_{[n]}$, is T^n and such that $\psi_n(\vec{P}_{[n]})$ defines the n-atom finite boolean algebra on generators $\vec{P}_{[n]}$. Then, we show that any ϕ defining the n-atom finite boolean algebra on some finite set of generators $\vec{R}$ will do the job.

We take:

- $\psi_0 := \bot$.

Suppose T^n is given as $T + \psi_n(\vec{P}_{[n]})$ in signature $\Sigma + \vec{P}_{[n]}$. We compute $T^{n+1} := T^n \times T$. Clearly, T^{n+1} can be taken to be the theory in the signature $\Sigma + \vec{P}_{[n]}$, which is the intersection of $T + \psi_n(\vec{P}_{[n]}) + P_n$ and $T + \bigwedge \vec{P}_{[n]} + \neg P_n$. So we take:

- $\psi_{n+1}(\vec{P}_{[n+1]}) := (\psi_n(\vec{P}_{[n]}) \wedge P_n) \vee (\bigwedge \vec{P}_{[n]} \wedge \neg P_n)$.

It is easy to see that the algebra defined by ψ_{n+1} on generators $\vec{P}_{[n]}$ has precisely $n + 1$ atoms.

Suppose $\phi(\vec{R})$ is a formula defining the n-atom finite boolean algebra over generators $\vec{R}$. This algebra is isomorphic to the algebra defined by by $\psi_n(\vec{P}_{[n-1]})$. Let σ be the isomorphism these algebras. We define unrelativized interpretations K, M between $T + \phi$, of signature $\Sigma + \vec{R}$, and $T + \psi_n$ of signature $\Sigma + \vec{P}_{[n]}$ and back as follows. K and M are the identity on Σ and $R^K := \sigma(R)$ (to be precise: some $\vec{P}_{[n]}$-formula defining $\sigma(R)$) and $P^M := \sigma^{-1}(P)$. It is easy to see that K and M specify a synonymy of theories. ❑

Of course, we can find far more efficient sequences of propositional formulas that do the job. E.g., we could use the sequence χ_n with:

- $\mathsf{e}(0) := 0$, $\mathsf{e}(2k+1) := \mathsf{e}(2k)$, $\mathsf{e}(2k+2) := \mathsf{e}(k+1) + 1$,
- $\chi_0 := \bot$, in signature Σ,
- $\chi_1 := \top$, in signature Σ,
- $\chi_{2k+2} := \chi_{k+1}$, in signature $\Sigma + \vec{P}_{[\mathsf{e}(2k+2)]}$,
- $\chi_{2k+3} := (\chi_{2k+2} \wedge P_{\mathsf{e}(2k+3)}) \vee (\bigwedge \vec{P}_{[\mathsf{e}(2k+3)]} \wedge \neg P_{\mathsf{e}(2k+3)})$, in signature $\Sigma + \vec{P}_{[\mathsf{e}(2k+2)]}$.

Example 4.4 One may show that PA^2 is not bi-interpretable with PA. See Corollary 9.9.

We provide an example, due to Lev Beklemishev, of a theory T such that T^2 is synonymous to T. Take T to be PA expanded with a unary relation symbol Q.

We interpret T^2 in T via K, taking $P_K := Q(0)$ and $Q_K(v_0) := Q(Sv_0)$. For the rest of the signature K is the identity interpretation. Conversely we interpret T in T^2 via M, taking $Q_M(v_0) := ((v_0 = 0 \wedge P) \vee \exists y\,(v_0 = Sy \wedge Qy))$. For the rest of the signature, M is the identity interpetation. ❑

4.4 The Sum

In the present subsection we provide a partial treatment of the occurrence of sums in our categories.[11] Consider two theories U and V. Say, U has signature Σ_U and V has signature Σ_V. The sum $U \oplus V$ is given as a theory W of signature Σ_W, where Σ_W is given as the disjoint union of Σ_U and Σ_V plus two additional fresh unary predicate symbols Δ_U and Δ_V and a new binary identity symbol E_W.[12] Let τ_U and τ_V be the obvious translations of the languages of U, respectively V into the language of W, where we relativize to Δ_U in the first case and to Δ_V in the second case. We take W to be axiomatized by the following axioms.

- $\vdash P_{\tau_U}\vec{v} \to \vec{v} : \Delta_U$,
- $\vdash P_{\tau_V}\vec{v} \to \vec{v} : \Delta_V$,
- $\vdash A^{\tau_U}$, for A a U-axiom,
- $\vdash A^{\tau_V}$, for A a V-axiom,
- $\vdash \forall x\,(x : \Delta_U \vee x : \Delta_V)$,
- $\vdash xE_Wy \leftrightarrow \forall z\,((xE_Uz \leftrightarrow yE_Uz) \wedge (xE_Vz \leftrightarrow yE_Vz))$.

Note that, in the presence of the other axioms, the last axiom says that E_W is the crudest congruence relation w.r.t. all predicates of W.

It is easy to check that $\oplus$ is a sum in the sense of category theory for the category INT. It follows that $\oplus$ yields a *weak* sum in the categories INT_1, INT_2, INT_3. Moreover, $\oplus$ is again a sum in INT_4. I have not thought about the situation in INT_2 and INT_3, but it is not difficullt to see that the situation w.r.t. the categorical sum is markedly different in INT_1, i.e. hINT.

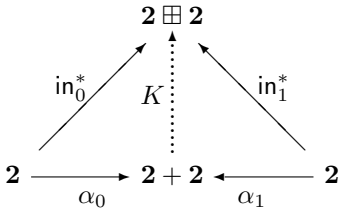
Let's call the categorical sum of hINT , whenever it exists: $+$. First, consider $\mathbf{1}$, the theory in the language with just identity stating that there is precisely on element. The theory $\mathbf{1}$ is an initial object of hINT. Hence we have, for any A, $A + \mathbf{1} = A$. Hence, e.g. $\mathbf{1} + \mathbf{1}$ is not bi-interpretable with $\mathbf{1} \oplus \mathbf{1}$.

Now consider the theory $\mathbf{2}$. This is the theory in the language with just identity stating that there are precisely two elements. We claim that $\mathbf{2} + \mathbf{2}$ does not exist. Suppose, to get a contradiction, that $\mathbf{2} + \mathbf{2}$ does exist. We take α_0 and α_1 to be the in-arrows associated with $\mathbf{2} + \mathbf{2}$. We will call the formulas Δ of $\mathbf{2} \oplus \mathbf{2}$: Δ_0 and Δ_1 Let $\mathbf{2} \boxplus \mathbf{2}$ be $\mathbf{2} \oplus \mathbf{2}$ extended with the axiom stating that

[11] A number of details of the sum definition were studied by Spencer Gerhardt in the context of a small project. Specifically, he formulated the definition of sum in INT.

[12] Of course, if $U = V$, we take the appropriate measures to make the Δ disjoint.

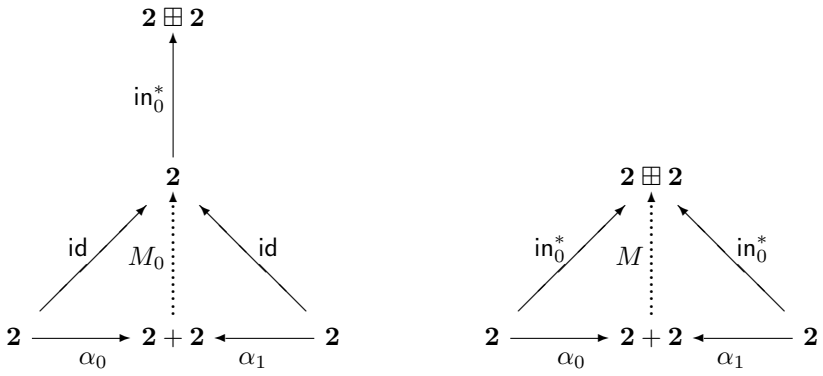
the intersection of Δ_0 and Δ_1 is empty. Let $\mathsf{in}_i^* := \mathcal{E}_{\mathbf{2}\oplus\mathbf{2},\mathbf{2}\boxplus\mathbf{2}} \circ \mathsf{in}_i$. Let K be the unique arrow making the following diagram commutative.



Note that, in $\mathbf{2} \boxplus \mathbf{2}$, we have $E_i := E_U \restriction \Delta_i$. We find that $\mathbf{2} \boxplus \mathbf{2}$ is categorical. Let $\mathcal{A}$ be a model of $\mathbf{2} \boxplus \mathbf{2}$ with domain $\{a, b, c, d\}$ and with $\Delta_0^{\mathcal{A}} = \{a, b\}$ and $\Delta_1^{\mathcal{A}} = \{c, d\}$. Since definable sets are closed under automorphisms, the only definable sets of $\mathcal{A}$ are $\emptyset$, $\{a, b\}$, $\{c, d\}$ and $\{a, b, c, d\}$. Similarly, there are just three interpretations of $\mathbf{2}$ in $\mathcal{A}$. Moreover, there are no definable isomorphisms between any two different interpretations of $\mathbf{2}$. Thus, we may conclude that:

$$\mathbf{2} \boxplus \mathbf{2} \vdash x : \Delta_i \leftrightarrow x : \delta_{K \circ \alpha_i}.$$

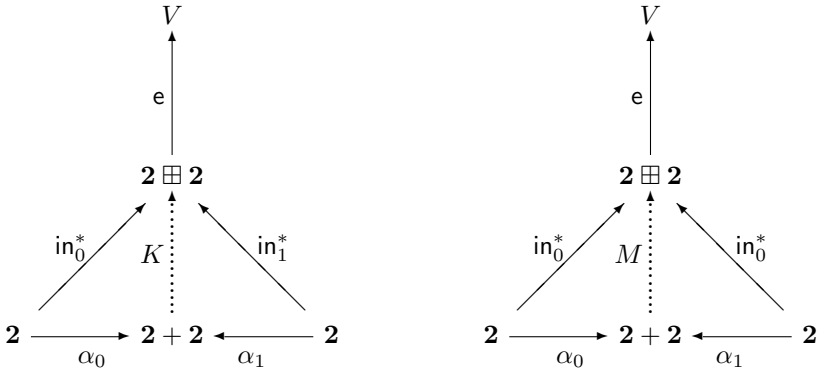
Thus, $\mathbf{2} \boxplus \mathbf{2} \vdash (\forall x \neg (\delta_{\alpha_0}(x) \wedge \delta_{\alpha_1}(x)))^K$. Now compare the following commutative diagrams.



Clearly $M = \mathsf{in}_0^* \circ M_0$. Hence, $\mathbf{2} \boxplus \mathbf{2} \vdash (\forall x\ (\delta_{\alpha_0}(x) \leftrightarrow \delta_{\alpha_1}(x)))^M$.

Let V be $\mathbf{2} \boxplus \mathbf{2}$ extended with a new predicate Φ and an axiom stating that Φ is a bijection between Δ_0 and Δ_1. Let e be the obvious embedding of U in

V. We compare the commutative following diagrams.



Note that $\mathsf{e} \circ \mathsf{in}_0^*$ is equal to $\mathsf{e} \circ \mathsf{in}_1^*$ in hINT. By the uniqueness property of $+$, it follows that $\mathsf{e} \circ K$ is equal to $\mathsf{e} \circ M$. On the other hand,

1. $V \vdash (\forall x \neg (\delta_{\alpha_0}(x) \wedge \delta_{\alpha_1}(x)))^{\mathsf{e} \circ K}$,
2. $V \vdash (\forall x \ (\delta_{\alpha_0}(x) \leftrightarrow \delta_{\alpha_1}(x)))^{\mathsf{e} \circ M}$.

This gives us a contradiction.

4.5 Monomorphisms

The arrow $K : U \to V$ in INT_i is a monomorphism if, for all W and for all $M_0 : W \to U$ and $M_1 : W \to U$, we have $K \circ M_0 =_i K \circ M_1$ implies $M_0 =_i M_1$. So, whenever the diagram

$$W \underset{M_1}{\overset{M_0}{\rightrightarrows}} U \xrightarrow{K} V$$

commutes, then so does:

$$W \underset{M_1}{\overset{M_0}{\rightrightarrows}} U$$

Since equality in INT_i occurs both positively and negatively in the definition of *monomorphism*, the property of being a monomorphism need not be preserved when i increases. We have the following theorem.

Theorem 4.5 *(a) Monomorphisms in* INT_i *for* $i = 0, 1, 2, 3$ *are faithful. This also holds when we relativize* INT_2 *to a class of models in which we have the completeness theorem. (b) The faithful interpretations of* INT_j *for* $j \in \{0, 3\}$ *are monomorphisms. (c) All interpretations of* INT_4 *are monomorphisms,*

Proof

We prove (a). Let $i \in \{0,1,2,3\}$. We work in INT_i. Suppose that $K : U \to V$ is a monomorphism. Suppose that $V \vdash B^K$. Let W be U extended with a 0-ary predicate symbol P. Let $M_0 : W \to U$ be the interpretation that is the identity when restricted to the signature of U and that sends P to $\top$. Let $M_1 : W \to U$ be the interpretation that is the identity when restricted to the signature of U and that sends P to B. Then, clearly, $K \circ M_0$ is equal_0, and, hence, equal_i, to $K \circ M_1$. So, since K is a monomorphism, we have that M_0 is equal_i, and, hence equal_3, to M_1. Ergo $V \vdash P^{M_0} \leftrightarrow P^{M_1}$. I.o.w. $V \vdash B$. Note that this last step fails for $i = 4$.

We prove (b). Let $i = 0, 3$. Suppose that $K : U \to V$ is faithful. We show that K is a monomorphism in INT_i. Suppose that $M_j : W \to U$, for $j = 0, 1$, and that $K \circ M_0$ is equal_i to $K \circ M_1$. We have to show that M_0 is equal_i to M_1.

We first treat $i = 0$. By our assumption V proves that $\tau_K \circ \tau_{M_0}$ and $\tau_K \circ \tau_{M_1}$ are identical. So:

- $V \vdash \forall x \, (\, (x : \delta_K \wedge x : \delta^K_{M_0}) \leftrightarrow (x : \delta_K \wedge x : \delta^K_{M_1}) \,)$.
- $V \vdash \forall \vec{x} \, (\, (\vec{x} : \delta_K \wedge \vec{x} : \delta^K_{M_0}) \to (P^K_{M_0}(\vec{x}) \leftrightarrow P^K_{M_1}(\vec{x})) \,)$.

But this just another way of saying:

- $V \vdash (\forall x \, (x : \delta_{M_0} \leftrightarrow x : \delta_{M_1}))^K$.
- $V \vdash (\forall \vec{x} : \delta_{M_0} \, (P_{M_0}(\vec{x}) \leftrightarrow P_{M_1}(\vec{x})))^K$.

By faithfulness, we find that U proves that τ_{M_0} and τ_{M_1} are identical.

We treat $i = 3$. Consider any W-sentence A. We have $V \vdash A^{M_0 K} \leftrightarrow A^{M_1 K}$. Hence: $V \vdash (A^{M_0} \leftrightarrow A^{M_1})^K$. By the faithfulness of K, we find $U \vdash A^{M_0} \leftrightarrow A^{M_1}$. Ergo M_0 is equal_3 to M_1.

(c) is trivial. ❑

In the next theorem, we give some connections between *being a monomorphism* and the behaviour of the MOD-functor.

Theorem 4.6 *(a) For $i = 0, 2, 3$, if $\mathsf{MOD}_i(K)$ is surjective, then $K : U \to V$ is a monomorphism. (b) If $K : U \to V$ is a monomorphism in INT_3, then $\mathsf{MOD}_3(K)$ is surjective.*

In (a), we can improve the cases (0) and (3): these work also if we have the completeness theorem in the range of $\mathsf{MOD}_i(K)$.

Proof

We prove (a). Let $i \in \{0, 2, 3\}$. We will say that two models are *equal*$_i$ if they are the same if $i = 0$, isomorphic if $i = 2$ and elementarily equivalent if $i = 3$.

Suppose that $\mathsf{Mod}_i(K)$ is surjective. Suppose $K \circ M_0 =_i K \circ M_1$, for $M_j : W \to U$. Consider any model $\mathcal{M} \models U$. By surjectivity, there is a model $\mathcal{N} \models V$ such that $\mathcal{M}$ is equal$_i$ to $K^{\mathcal{N}}$. It follows that $M_j^{\mathcal{M}}$ is equal$_i$ to $M_j^{K^{\mathcal{N}}}$ which is, in its turn, equal$_i$ to $(M_j K)^{\mathcal{N}}$. Since also $(M_0 K)^{\mathcal{N}}$ and $(M_1 K)^{\mathcal{N}}$ are equal$_i$ to each other, we find that $M_0^{\mathcal{M}}$ and $M_1^{\mathcal{M}}$ are equal$_i$. Ergo $M_0 =_i M_1$. (For the cases $i = 0, 3$, we only use, in the last step, that '$\mathcal{M}$' ranges over a class of models in which we have the completeness theorem.)

We prove (b). Let U^* be any consistent complete extension of U. If $V + U^{*K}$, were inconsistent, then for some A in U^*, $V \vdash \neg A^K$, and so $V \vdash (\neg A)^K$. Ergo: $U \vdash \neg A$. Quod non. Take V^* any complete extension of $V + U^{*K}$. Clearly, $\mathsf{Mod}_3(K)(V^*) = U^*$. ❑

We consider a number of examples of morphisms and consider the question whether they are monomorphisms in our various categories.

a. $\mathsf{id} : \mathsf{PA} \to \mathsf{PA}$.

b. $\mathsf{ext}_0 : \mathsf{PA} \to \mathsf{PA}^c + \{c \neq \underline{0}, c \neq \underline{1}, \cdots\}$. Here PA^c is simply PA with its signature extended with constant c. ext_0 is the standard 'identical' interpretation.

c. Let $\mathbf{2} \boxplus \mathbf{2}$ and V be as defined in Subsection 4.4. Let ext_1 be the standard 'identical' interpretation of $\mathbf{2} \boxplus \mathbf{2}$ in V. (This interpretation was called e in Subsection 4.4.)

d. $\mathsf{ext}_2 : \mathsf{PA} \to \mathsf{ACA}_0$. Here ext_2 is the standard 'identical' interpretation.

e. $\mathsf{ext}_3 : \mathsf{PA} \to \mathsf{PA}_{\mathsf{ns}}$. Here $\mathsf{PA}_{\mathsf{ns}}$ is PA plus a non-standard satisfaction predicate and ext_3 is the standard identical interpretation.

f. $\mathcal{E} : \mathsf{PA} \to \mathsf{PA} + \mathsf{con}(\mathsf{PA})$. Here $\mathcal{E}$ witnesses the subtheory relation.

In the diagram below $i \in 0, 1, 2, 3, 4$ indicates INT_i. $2'$ indicates $\mathsf{INT}_2^{\mathsf{count}}$, i.e. INT_2, where we restrict ourselves in the definition of equality of interpretations to countable models. We first sum up our results in a diagram and then run through the proofs.

	0	1	2	$2'$	3	4
a	+	+	+	+	+	+
b	+	+	?	−	+	+
c	+	−	+	+	+	+
d	+	−	+	+	+	+
e	+	−	?	?	+	+
f	−	−	−	−	−	+

The a-row is trivial and so is the 4-column. the 0- and 3-columns are easy because monomorphisms in INT_0 and INT_3 are precisely the faithful interpretations. The f-row is immediate since monomorphisms in INT_i, for $i = 0, 1, 2, 3$ are all faithful. We treat the remaining cases.

case b1 Suppose $\mathsf{ext}_0 \circ M_0 =_1 \mathsf{ext}_0 \circ M_1$. So $\mathsf{PA}^c + \{c \neq \underline{0}, c \neq \underline{1}, \cdots\}$ proves that τ_{M_0} and τ_{M_1} are isomorphic via some isomorphism represented by $J[x := c]$. Here J is a formula of the language of PA having only x, v_0, v_1 free. The fact that $J[x := c]$ presents an isomorphism can be stated in a single sentence. Hence, by compactness there must be some number N, such that PA proves that $J[x := \underline{N}]$ is an isomorphism between τ_{M_0} and τ_{M_1}.

case b2′ Let Z_0, Z_1 be the obvious interpretations of the order theory of $\mathbb{Z}$ and $\mathbb{Z} \times \mathbb{Q}$ in PA. In $\mathbb{N}$ these interpretations give us precisely the orderings of $\mathbb{Z}$ and $\mathbb{Z} \times \mathbb{Q}$. In a countable non-standard model, both will give us the ordering of $\mathbb{Z} \times \mathbb{Q}$.

case c0,2,3 Every model of $\mathbf{2} \boxplus \mathbf{2}$ can be ext_1-extended to a model of V. This makes $\mathsf{Mod}_i(\mathsf{ext}_1)$ surjective, for $i \in \{0, 2, 3\}$. (Note that we already knew c0 and c3.)

case c1 By a simple model theoretical argument, in_0^* and in_1^* are distinct in INT_1. However, $\mathsf{ext}_1 \circ \mathsf{in}_0^* =_1 \mathsf{ext}_1 \circ \mathsf{in}_1^*$.

case d0,2,3 Every model of PA can be ext_2-extended to a model of ACA_0. This makes $\mathsf{Mod}_i(\mathsf{ext}_2)$ surjective, for $i \in \{0, 2, 3\}$. (Note that we already knew d0 and d3.)

case d1 This beautiful example is due to Harvey Friedman (in conversation). Let W be the theory of linear order. We take:

- $\delta_{M_0} :\leftrightarrow \forall y\, (\mathsf{proof}_{\mathsf{PA}}(y, \bot) \to v_0 < y)$,
- $\delta_{M_1} :\leftrightarrow \forall y\, (\mathsf{proof}_{\mathsf{PA}}(y, \bot) \to v_0 \leq y)$.
- $\leq_{M_i}$ will be $\leq$ restricted to δ_{M_i}.

In ACA_0 we may produce an isomorphism between $\mathsf{ext}_2 \circ M_0$ and $\mathsf{ext}_2 \circ M_1$ as follows. There is a definable cut I in ACA_0 such that $\mathsf{ACA}_0 \vdash \mathsf{con}^I(\mathsf{PA})$. Thus $\delta_{M_i} \supseteq I$. Define:

$$\begin{aligned} v_0 G v_1 \quad :\leftrightarrow \quad & v_0 : \delta_{M_0} \wedge v_1 : \delta_{M_1} \wedge \\ & ((v_0 : I \wedge v_1 = v_0) \vee (\neg\, (v_0 : I) \wedge v_1 = \mathsf{S} v_0)). \end{aligned}$$

It is easy to see that, indeed, G provides the desired isomorphism.

Finally, note that the existence of a definable isomorphism in PA would imply, in $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$, the existence of a definable order isomorphism between p and $p - 1$, where p is the smallest proof of inconsistency. Quod impossibile.

case e0,3 Lachlan's theorem tells us that a countable model of PA can be extended to a model of $\mathsf{PA}_{\mathsf{ns}}$ iff it is recursively saturated. See [Kay91]. Since we have the completeness theorem for countable recursively saturated models we find that ext_3 is faithful. A proof-theoretical argument for faithfulness is given in [Hal99].

case e1 The argument is the same as the one for d1.

4.6 Epimorphisms

Just as in the case of monomorphisms, *being an epimorphism* is not necessarily preserved when we move from INT_i to INT_{i+1}.

An important class of epimorphisms (in all our categories) is constituted by the extension morphisms $\mathcal{E}_{UV}$. Suppose we have $M_0, M_1 : V \to W$ and $M_0 \circ \mathcal{E}_{UV} = M_1 \circ \mathcal{E}_{UV}$. Then, we must have, in any of our categories, that $M_0 = M_1$, because equality only depends the theory W and the underlying translations, which are not affected by composing an $\mathcal{E}$-morphism.

An interpretation $K : U \to V$ is *surjective* iff, for all $B \in \mathcal{S}_V$, there is an $A \in \mathcal{S}_U$ such that $V \vdash B \leftrightarrow A^K$. Note that in equality of interpretations in the categories INT_i, for $i = 0, 1, 2, 3$, preserves surjectivity. So in these categories we can sensibly consider surjectivity to be a property of morphisms.

Theorem 4.7 *The epimorphisms of* eqINT *are precisely the surjective morphisms.*

Before giving the proof we define an auxiliary sum-like operation —which we already met in the special case of $\mathbf{2} \boxplus \mathbf{2}$. Consider theories T and Z. We define the theory $T \boxplus Z$ as the extension of $T \oplus Z$ with the axiom stating that the intersection of Δ_T and Δ_Z is empty. We define $\mathsf{in}_i^* := \mathcal{E}_{T\oplus Z, T\boxplus Z} \circ \mathsf{in}_i$.[13]

Similarly, we build a disjoint sum of models $\mathcal{M} \boxplus \mathcal{N}$ with domain the disjoint union of the domains of $\mathcal{M}$ and $\mathcal{N}$.

Proof

Consider a morphism $K : U \to V$. Suppose that, for some V-sentence B, there is no U-sentence A such that $V \vdash B \leftrightarrow A^K$. We construct a complete extension U^* of U such that B is independent over $V + U^{*K}$. The theory U^* will be the union of an increasing sequence of theories U_n, where B is is not equivalent with any A^K over $V + U_n^*$. We fix some enumeration C_n of the U-sentences.

- Let $U_0 := U$.

[13] The operation $\boxplus$ gives the sum in the category of (recursive) boolean morphisms. It does not generally give the sum in any of our categories INT_i. We have the remarkable property that every sentence of $T \boxplus Z$ is a boolean combination of sentences of the form $A^{\mathsf{in}_0^*}$, where A is a T-sentence, and $B^{\mathsf{in}_1^*}$, where B is a Z-sentence.

- Suppose $V + U_n^K + C_n^K \vdash B \leftrightarrow A_0^K$ and $V + U_n^K + \neg\, C_n^K \vdash B \leftrightarrow A_1^K$. Then, $V + U_n^K \vdash B \leftrightarrow ((C_n \wedge A_0) \vee (\neg\, C_n \wedge A_1))^K$. Quod non. In case, for no A_0, $V + U_n^K + C_n^K \vdash B \leftrightarrow A_0^K$, we take $U_{n+1} := U_n + C_n$. Otherwise, we take $U_{n+1} := U_n + \neg\, C_n$.

It is easy to see that this construction does the trick. Now consider the following theory:

$$W := (V \boxplus V) + \{A^{K\mathsf{in}_0} \leftrightarrow A^{K\mathsf{in}_1} \mid A \in \mathcal{S}_U\} + B^{\mathsf{in}_0} + \neg\, B^{\mathsf{in}_1}.$$

(In the definition of W we confuse K and the in_i with their underlying translations.) If we take models $\mathcal{M}_0 \models V + U^{*K} + B$ and $\mathcal{M}_1 \models V + U^{*K} + \neg\, B$, then $\mathcal{M}_0 \boxplus \mathcal{M}_1 \models W$. Hence, W is consistent.

Let $M_i := \mathcal{E}_{V\boxplus V,W} \circ \mathsf{in}_i^*$. Clearly, $M_0 \circ K$ and $M_1 \circ K$ are the same in eqINT. However, $W \vdash B^{M_0}$ and $W \vdash \neg\, B^{M_1}$, so M_0 and M_1 are different. Hence, K is not an epimorphism.

Conversely, suppose that K is surjective. Let $M_0, M_1 : V \to W$ and suppose that $M_0 \circ K$ is equal to $M_1 \circ K$ in eqINT. Consider any V-sentence B. By surjectivity, we can find a U-sentence A such that $V \vdash B \leftrightarrow A^K$. We have:

$$\begin{array}{rlcl} W \vdash & B^{M_0} & \leftrightarrow & A^{KM_0} \\ & & \leftrightarrow & A^{KM_1} \\ & & \leftrightarrow & B^{M_1} \end{array}$$

Ergo, M_0 is equal to M_1. ❑

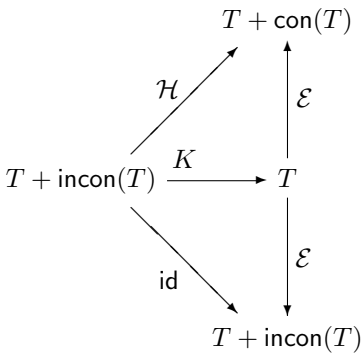
4.7 Split Monomorphisms

Suppose $K : U \to V$, $M : V \to U$ and $M \circ K = \mathsf{id}_U$. In these circumstances, we call K a *split monomorphism* or *co-retraction*. We call M a *split epimorphism* or *retraction*. We say that U *is a retract of* V. Note that if K is a split monomorphism in INT_i and $i < j$, then K is a split monomorphism in INT_i. Note also that the contravariant MOD_i-functor sends split monomorphisms to split epimorphisms and that it sends split epimorphisms to split monomorphisms. It is easily seen that split monomorphisms are monomorphisms and split epimorphisms are epimorphisms.

It seems that many interesting interpretations are split monomorphisms. Here are two examples.

Example 4.8 Gödel's interpretation $\mathsf{gödel} : (\mathsf{ZF} + \mathsf{V} = \mathsf{L}) \to \mathsf{ZF}$ is a split monomorphism in INT, as can be seen by inspecting the construction. The corresponding split epimorphism is simply $\mathcal{E}_{\mathsf{ZF},\mathsf{ZF}+\mathsf{V}=\mathsf{L}}$. Note that it follows that $\mathsf{gödel}$ is faithful. ❑

Example 4.9 Suppose $N : \mathsf{S}^1_2 \to T$. Here S^1_2 is Buss's arithmetic (see [Bus86] or [HP91]). However, any sufficiently rich arithmetical theory would do as well. We will use an arithmetization of metamathematics in T that is implicitly relativized to N. Consider the following commutative diagram in hINT:



Note that $T \vdash \mathsf{con}(T) \to \mathsf{con}(T + \mathsf{incon}(T))$, by the Second Incompletness theorem. The interpretation $\mathcal{H} : (T + \mathsf{incon}(T)) \to (T + \mathsf{con}(T))$ is the Henkin interpretation based on $\mathsf{con}(T + \mathsf{incon}(T))$. See e.g. [Vis91], for an explanation. We take: $K := \mathcal{H}\langle \mathsf{con}(T) \rangle \mathsf{id}$. Clearly, K is a split monomorphism. ❑

In Section 7, we will show that the interpretations involved in the Orey phenomenon, to wit the interpretability of both $T + O$ and $T + \neg O$ in T, for certain T and O, are split monomorphisms. We will prove further several results on (co)retractions in this paper.

4.8 Isomorphisms

In this subsection, we discuss isomorphisms in our various categories.

4.8.1 Bisimulation

Here is a useful insight concerning isomorphisms.

Theorem 4.10 *Being isomorphic in* INT_i *for* $i = 0, 1, 2, 3$ *is a bisimulation w.r.t. theory extension in the same language.*

Proof

Suppose $L : U \to V$ and $M : V \to U$ witness the fact that U and V are isomorphic in one of INT_i, for $i = 0, 1, 2, 3$. Suppose $U \subseteq U'$. Take $V' := \{A \in \mathcal{S}_V \mid U' \vdash A^M\}$. Clearly, M lifts to an interpretation $\tilde{M} : V' \to U'$ with the same underlying translation. Moreover,

$$\begin{array}{rcl} U' \vdash A & \Rightarrow & U' \vdash A^{LM} \\ & \Rightarrow & V' \vdash A^L \end{array}$$

So L lifts to an interpretation of $\tilde{L} : V' \to U'$. The fact that this pair of interpretations witness isomorphisms only depends on the underlying translations and the fact that we have at least U and V available. ❑

Note that our argument even establishes that recursive boolean isomorphism is a bisimulation w.r.t. theory extension. Our theorem does not hold for INT_4 or even for $\mathsf{INT}_{4,\mathsf{faith}}$, which is INT_4 restricted to faithful interpretations. See Subsubsection 4.8.4.

4.8.2 Synonymy

In this subsubsection, we study the classical notion of synonymy.

Theorem 4.11 *If two theories are synonymous, then they are isomorphic in* $\mathsf{INT}_{\mathsf{unr},=}$

This is an immediate consequence of Theorem 6.1, which tells us that, if $M \circ K$ in INT is direct, then M is direct.

Consider a theory T with signature Σ. Let T' be a theory with signature Σ'. We say that T' is a *definitional extension* of T iff Σ' extends Σ and the axioms of T' are the axioms of T plus, for each predicate symbol P of $\Sigma' \setminus \Sigma$, an axiom of the form: $\vdash P\vec{x} \leftrightarrow A\vec{x}$, where A is in the language of T with at most $\vec{x}$ free. We have:

Theorem 4.12 *Any definitional extension of a theory is synonymous to that theory.*

The proof of Theorem 4.12 is easy. One immediate consequence of Theorems 4.11 and 4.12 is that Karel de Bouvère's notion of synonymity coincides with ours. See [dB65a] and [dB65b]. This notion has also been called *definitional equivalence*. See [Kan72] or [Cor80] or [Hod93]. It has also been called *isomorphism*. See [Per97].

Since MOD_i is a contravariant functor, it follows that $\mathsf{MOD}_i(K)$ is an isomorphism if K is an isomorphism in INT_i. The following theorem shows that, if K is direct and $i = 0$, the converse is also true.

Theorem 4.13 *Suppose $K : U \to V$ is direct. We have:*

1. $\mathsf{MOD}(K)$ *is injective iff* $\breve{K} : K^{-1}[V] \to V$ *is an isomorphism in* INT*;*
2. $\mathsf{MOD}(K)$ *is bijective iff K is an isomorphism in* INT*.*

Proof

The proof is an adaptation of the proof of Theorem 2 of [dB65b]. Suppose K is direct.

We first prove (1). Suppose that $\mathsf{MOD}(K)$ is injective. Suppose U has signature Σ and V has signature Θ. Without loss of generality, we may assume that the identity symbol of Σ is the same as the one of Θ and that, for all other predicate symbols, Σ and Θ are disjoint.

Let $\tilde{\Theta}$ be a disjoint copy of Θ. This means that we replace every predicate symbol P of Σ, except the identity symbol E, by a disjoint copy $\tilde{P}$. Again, we may assume that, for all non-identity symbols, Σ and $\tilde{\Theta}$ are disjoint. So, we end up with three signatures that are pairwise disjoint, except for the shared identity symbol.

We take $\tilde{V}$ the obvious copy of V in the signature $\tilde{\Theta}$. Let V^+ be the theory in the signature $\Sigma + \Theta$ obtained by extending the axioms of V with axioms $\forall \vec{v}\ (P(\vec{v}) \leftrightarrow P_K(\vec{v}))$, for each predicate symbol P of Σ. The theory $\tilde{V}^+$ is similarly defined. Let W be the theory in the signature $\Sigma + \Theta + \tilde{\Theta}$ axiomatized by $V^+ + \tilde{V}^+$.

We claim that, for every predicate symbol Q of Θ, $W \vdash \forall \vec{v}\ (Q(\vec{v}) \leftrightarrow \tilde{Q}(\vec{v}))$. Suppose not. By symmetry, we may conclude that, for some Q, there is a model $\mathcal{M}$ of $W + \exists \vec{v}\ (Q(\vec{v}) \wedge \neg\, \tilde{Q}(\vec{v}))$. We define $\mathcal{M}_0$ as the reduct of $\mathcal{M}$ to Θ, $\mathcal{N}$ as the reduct of $\mathcal{M}$ to Σ, and $\mathcal{M}_1$ as the result of first restricting $\mathcal{M}$ to $\tilde{\Theta}$ and, then, replacing the predicates of $\tilde{\Theta}$ by the corresponding predicates of Θ. As is easily seen:

$$\mathsf{MOD}(K)(\mathcal{M}_0) = \mathcal{N} = \mathsf{MOD}(K)(\mathcal{M}_1).$$

Ergo, by injectivity, $\mathcal{M}_0 = \mathcal{M}_1$. A contradiction.

By Beth's Theorem, we find that (†) $V^+ \vdash \forall \vec{v}\ (Q(\vec{v}) \leftrightarrow B_Q(\vec{v}))$, where B_Q is some formula having only $\vec{v}$ free, in the language of Σ. We define a direct translation μ by setting $Q_\mu :\leftrightarrow B_Q$. We have, for $B \in \mathcal{S}_\Theta$,

$$\begin{aligned} K^{-1}[V] \vdash B^\mu &\Leftrightarrow V^+ \vdash B^\mu \\ &\Leftrightarrow V^+ \vdash B \\ &\Leftrightarrow V \vdash B. \end{aligned}$$

So μ lifts to an interpretation $M : V \to K^{-1}[V]$. Also $V^+ \vdash \forall \vec{v}\ (Q(\vec{v}) \leftrightarrow Q^K_M(\vec{v}))$, and so $V \vdash \forall \vec{v}\ (Q(\vec{v}) \leftrightarrow Q^K_M(\vec{v}))$. Moreover, $V^+ \vdash \forall \vec{v}\ (P(\vec{v}) \leftrightarrow P^M_K(\vec{v}))$. Hence, $K^{-1}[V] \vdash \forall \vec{v}\ (P(\vec{v}) \leftrightarrow P^M_K(\vec{v}))$. We may conclude that $\breve{K}$ and M are inverses.

Conversely, suppose $\breve{K}$ is an isomorphism in INT between $K^{-1}[V]$ and V. since MOD is a contravariant functor, it follows that $\mathsf{MOD}(\breve{K})$ is a bijection. Clearly, $\mathsf{MOD}(\mathcal{E}_{U,K^{-1}[V]})$ is injective. Hence, since

$$\mathsf{MOD}(K) = \mathsf{MOD}(\breve{K} \circ \mathcal{E}_{U,K^{-1}[V]}) = \mathsf{MOD}(\mathcal{E}_{U,K^{-1}[V]}) \circ \mathsf{MOD}(\breve{K}),$$

we find that $\mathsf{MOD}(K)$ is injective.

We prove (2). Suppose $\mathsf{MOD}(K)$ is a bijection. By Theorem 4.6, it follows that K is a monomorphism in INT. Hence, by Theorem 4.5, K is faithful. We may conclude that $U = K^{-1}[V]$ and $K = \breve{K}$. So, by (1), we are done. The converse is easy. ❑

Remark 4.14 Theorem 4.13 can be considered as a generalization of Beth's Theorem. We sketch how to obtain Beth's Theorem from Theorem 4.13. Suppose Σ is a signature and Q is a predicate symbol, not in Σ. Let $\Sigma(Q)$ be Σ extended with Q. Suppose $W(Q)$ is a theory of signature $\Sigma(Q)$ in which Q is implicitly definable, i.e. $W(Q) + W(Q') \vdash \forall \vec{v}\ (Q(\vec{v}) \leftrightarrow Q'(\vec{v}))$. Let $\hat{W}$ be the theory of the consequences of W in the language of Σ. We have the obvious extension interpretation $\mathsf{ext} : \hat{W} \to W$. Evidently, $\mathsf{MOD}(\mathsf{ext})$ maps a model of $W(Q)$ to its restriction to Σ. Using the implicit definability of Q, we easily see that $\mathsf{MOD}(\mathsf{ext})$ is injective. Clearly, $\mathsf{ext}^{-1}[W] = \hat{W}$. By Theorem 4.13, it follows that ext has a direct inverse K. We find that Q_K is the desired explicit definition of Q. ❑

4.8.3 Bi-interpretability

We prove a modest preservation result for i-limits and i-colimits.

Theorem 4.15 *Suppose that $K : U \to V$ is an isomorphism in hINT. Then $[\![K]\!]$ preserves limits and colimits from $[\![U]\!]$ to $[\![V]\!]$.*

Proof

Clearly, $[\![K]\!]$ is an equivalence of $[\![U]\!]$ and $[\![V]\!]$. Since equivalences are both left and right adjoints, $[\![K]\!]$ preserves limits and colimits. ❑

4.8.4 Some Examples

Consider $\mathsf{FOL}_{\mathsf{arith}}$, predicate logic in the language of arithmetic and $\mathbf{1}$ the theory in the language of pure identity that states that there is precisely one element. $\mathsf{FOL}_{\mathsf{arith}}$ and $\mathbf{1}$ are mutually interpretable. The example shows the following points:

- Mutual interpretability does not preserve decidability, nor does it preserve categoricity.
- Mutual interpretability is not a bisimulation w.r.t. the subtheory relation, since $\mathsf{FOL}_{\mathsf{arith}}$ extends e.g. to Robinson's Arithmetic Q, but there is no matching extension of $\mathbf{1}$.

- Mutual interpretability is not the same as isomorphism in INT_i, for $i = 0, 1, 2, 3$. In fact, our two theories are not even isomorphic in the Boolean sense.

Here is a second example. Predicate Logic with just unary and 0-ary predicate symbols is mutually directly interpretable with Predicate logic with at least one n-ary predicate symbol, for some $n > 1$. However, the first theory is decidable and the second is not. Since, isomorphism in INT, for $i = 0, 1, 2, 3$ preserves decidability, it follows that our theories are not isomorphic in INT_i, for $i = 0, 1, 2, 3$.

In our third example, we separate mutual *faithful* interpretability from isomorphism in INT_i, for $i = 0, 1, 2, 3$. Consider the theory Q^- axiomatized by $(\exists x, y\ x \neq y \to \bigwedge \mathsf{Q})$ in the language of arithmetic. Clearly, the theories Q^- and $\mathsf{FOL}_{\mathsf{arith}}$ are mutually interpretable. (We can interpret the statement 'there is at most one element' in $\mathsf{FOL}_{\mathsf{arith}}$. This implies Q^-.) Since both theories are trustworthy (see [Vis02]), they are mutually faithfully interpretable. We extend $\mathsf{FOL}_{\mathsf{arith}}$ by adding the axiom 'there are precisely two elements'. Say the resulting theory is T. Suppose there is an extension U of Q^-, which is mutually faithfully interpretable with T. Since T is decidable, it follows that U is decidable. Hence the extension of U with Q is inconsistent. Since U contains the Q^--axiom, we may conclude that U proves: *there is at most one element*. But then it is impossible that U interprets T.

It follows that isomorphism in INT_i, for $i = 0, 1, 2, 3$, is not the same as mutual faithful interpretability.[14]

Note that the example also makes clear that mutual faithful interpretability is not a bisimulation w.r.t. extension of theories in the same language.

In example 9.5, we will separate mutual faithful direct interpretability from isomorphism in INT_i, for $i = 0, 1$.

There are several examples of sameness of theories, that prima facie belong in hINT. One such example is as follows. Consider ZF. We expand the signature with a unary predicate symbol U and with a binary relation symbol F. Let T be the theory in the expanded language given by the usual axioms for ZF with ur-elements, where the class of ur-elements is given by U, plus an axiom that says that F is a bijection between U and ω. We can interpret T in ZF by representing ur-elements as pairs $\langle 0, n\rangle$, for $n \in \omega$, and sets as pairs $\langle 1, x\rangle$, where x is a set of representations of sets and ur-elements. To get things to work we need $\in$-induction and $\in$-recursion. We can show that the interpretation so constructed is an isomorphism in hINT. So one might wonder: can we improve this result to show that ZF and T are synonymous? Benedikt Löwe produced an

[14] Our example also works for recursive boolean isomorphism in the place of isomorphism in one of the categories INT_i, for $i = 0, 1, 2, 3$. The argument for non-isomorphism of the theory T and any theory U over $\mathsf{FOL}_{\mathsf{arith}}$ that implies 'there is at most one element', is a simple count of the propositions in the associated Lindenbaum algebra. The theory U has at most 2^{2^3} propositions, where T has strictly more.

argument to show that indeed one can do this. So we are left with the following question.

Open Question 4.16 Give an example of two theories that are bi-interpretable but not synonymous. ❑

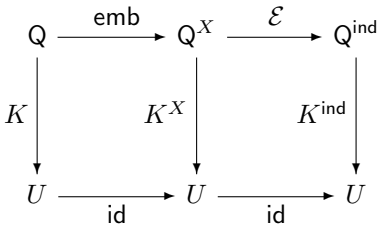
Open Question 4.17 Is there an interesting class of theories on which mutual interpretability and isomorphism in one of INT_i, for $i = 0, 1, 2, 3$, always coincide? What is the situation for the finitely axiomatizable sequential theories? ❑

5 Axiom Schemes

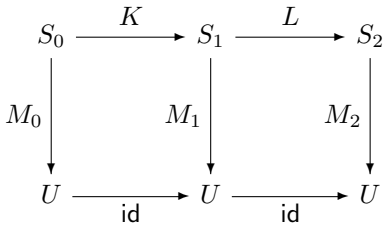
Consider the principle of complete induction over numbers in PA and in ZF. We would like to say that this is *the same scheme* only realized in different languages. Moreover, in the case of ZF, the scheme uses numbers which are not reflected in the signature of the theory. The machinery of relative interpretations provides a simple and convenient way to define such schemes. Consider the following three theories.

- Q, i.e. Robinson's Arithmetic;
- Q^X, i.e. Robinson's Arithmetic with an extra unary predicate symbol X n the signature, but with no further axioms;
- $\mathsf{Q}^{\mathsf{ind}} := \mathsf{Q}^X + \mathsf{IND}(X)$, where $\mathsf{IND}(X)$ is the principle of induction over X.

We have the obvious embeddings $\mathsf{emb} := \mathsf{emb}_{\mathsf{Q},\mathsf{Q}^X}$ of Q in Q^X and $\mathcal{E} := \mathcal{E}_{\mathsf{Q},\mathsf{Q}^{\mathsf{ind}}}$ of Q^X into $\mathsf{Q}^{\mathsf{ind}}$. We can now say that an interpetation $K : \mathsf{Q} \to U$ satisfies full induction iff, for all interpretations $K^X : Q^X \to U$, such that $K^X \circ \mathsf{emb} = K$, there is an interpretation $K^{\mathsf{ind}} : \mathsf{Q}^{\mathsf{ind}} \to U$, such that $K^{\mathsf{ind}} \circ \mathcal{E} = K^X$.



Reflecting on this example leads us to the following definition. An *ae-scheme* is a pair of composable arrows $\langle K, L\rangle$. To emphasize a pair is used in the role of ae-scheme we will use $K \twoheadrightarrow L$. We define, for $K : S_0 \to S_1$, $L : S_1 \to S_2$, $M_0 : S_0 \to U$, that $M_0 \models K \twoheadrightarrow L$ iff, for all interpretations $M_1 : S_1 \to U$, such that $M_1 \circ K = M_0$, there is an interpretation $M_2 : S_2 \to U$, such that $M_2 \circ L = M_1$.



In other words, $M_0 \models (K \twoheadrightarrow L)$ iff $[\![K]\!]_U^{-1}(M_0) \subseteq \mathsf{range}[\![L]\!]_U$. Note that, in our set-up, ae-schemes are not ascribed to theories but to interpretations. Of course, the most standard kind of ae-schemes are ascribed to the identity interpretation. Thus, $\mathsf{id} : \mathsf{PA} \to \mathsf{PA}$ satisfies $\mathsf{emb}_{\mathsf{Q},\mathsf{Q}^X} \twoheadrightarrow \mathcal{E}_{\mathsf{Q}^X,\mathsf{Q}^{\mathsf{emb}}}$, the induction scheme. The definition of satisfaction of an ae-scheme can be used in all our categories. The default is INT. If we employ other categories INT_i we will write: $M \models^i K \twoheadrightarrow L$. Here are some further notations. Suppose $K : U \to V$.

- $[\![\sigma]\!] := \{K \mid K \models \sigma\}$;
  We write $[\![\sigma]\!]_i$ if we want to indicate that we consider the satisfiers in INT_i. No index corresponds with $i = 0$.
- $?K := (\mathsf{id}_U \twoheadrightarrow K)$.
  We will also represent $?K$ as $\langle K \rangle$. We will call a scheme of the form $?K$ an *e-scheme*.
- $\sim K := (K \twoheadrightarrow \bot_V)$.
  Here $\bot$ is the unique arrow from V to the inconsistent theory in the signature with only identity.

The following theorem is obvious.

Theorem 5.1 *Consider any ae-scheme σ. Consider $M : S_0 \to U$, such that $M \models^i \sigma$. Suppose that $N : U \to V$ is an isomorphism of theories in INT_i. Then, $N \circ M \models^i \sigma$. In slogan: ae-schemes are preserved under isomorphism of theories.*

A ae-scheme $(K \twoheadrightarrow L)$ is *direct* iff both K and L are direct. Thus, induction is a direct scheme. We have the following theorem.

Theorem 5.2 *Suppose the ae-scheme σ is direct. Then, $[\![\sigma]\!]_0 = [\![\sigma]\!]_1$.*[15]

The proof uses that direct interpretations are forward looking. See Section 6. It follows that induction is satisfied in hINT whenever it is satisfied in INT. So, induction in the ordinary sense (the 0-sense) is preserved over bi-interpretability.

[15]Strictly speaking we should say something like: M is in $[\![\sigma]\!]_0$ iff its standard embedding into hINT is in $[\![\sigma]\!]_1$.

A severe restriction of our present approach is that we can only treat schemes without parameters. This is due to the fact that we only consider parameter-free interpretations. To get around the restriction, we have to extend our category. Note that, in the case of full induction, the restriction does not matter: parameter-free full induction and full induction with parameters happen to be equivalent.

We provide some extra information about e-schemes. The following theorem is obvious.

Theorem 5.3 *Let $K : U \to V$, $M : U \to W$, $N : W \to Z$. Suppose $M \models ?K$. then $N \circ M \models ?K$.*

In the next theorem, we study how e-schemes are preserved upwards and downwards iff we change categories. Remember our convention that the variables $K, \ldots$ really run over the tuples $\langle U, \tau, V \rangle$.

Theorem 5.4 *1. Suppose $i \leq j$ and $M \models^i ?K$. Then, $M \models^j ?K$.*

2. Suppose $i \leq j$ and $M \models^j ?K$. Suppose further that, for any L, L' with $L =_j L'$, we have $L \models^i ?K$ iff $L' \models^i ?K$. Then, $M \models^i ?K$.

Proof

We prove (2). Suppose $i \leq j$ and (a): $M \models^j ?K$. Suppose further that (b): for any L, L' with $L =_j L'$, we have $L \models^i ?K$ iff $L' \models^i ?K$.

By (a), we have that, for some N, $M =_j N \circ K$. Clearly, $N \circ K \models^i ?K$. Hence, by (b), $M \models^i ?K$. ❑

Let ϕ be a function from theories U to interpretations K with $\mathsf{dom}(K) = U$. We say that $?\phi := \langle \phi \rangle$ is a *uniform* e-scheme if, for any $K : U \to V$, there is a K' such that $K' \circ \phi_U = \phi_V \circ K$. Define:

- $M \models ?\phi :\Leftrightarrow M \models ?\phi_{\mathsf{dom}(M)}$,
- $\mathsf{SAT}_\phi := \{M \mid M \models ?\phi\}$.

It would be nice to define the notion of uniform e-scheme for ϕ as a natural transformation. However, our application in Section 7 asks for our present less restrictive definition. The following theorems are easy.

Theorem 5.5 *Suppose $M \circ N$ exists. Then $M \circ N$ is in SAT_ϕ if M is in SAT_ϕ or N is in SAT_ϕ.*

Theorem 5.6 *If $?\phi$ is a uniform e-scheme in INT_i and $i \leq j$, then $?\phi$ is a uniform e-scheme in INT_j. Moreover, if $K \models^i ?\phi$, then $K \models^j ?\phi$.*

6 i-Isomorphisms

In this section we develop our knowledge of the 2-category $\mathsf{INT}^{\mathsf{iso}}$ a bit further. Subsection 6.1 is devoted to a characterization of direct interpretations. In Subsection 6.2, we provide some sufficient conditions for the transfer of certain properties of morphisms in hINT to corresponding morphisms in INT and vice versa. In Subsection 6.3, we show how in some circumstances one may replace an interpretation by an i-isomorphic one with 'better' properties. Finally, in Subsection 6.4, we consider a specific example, the Ackermann interpretation in some detail.

6.1 Direct Interpretations and Discrete Fibrations

An interpretation K is *direct* iff it is unrelativized and preserves identity. Thus, direct interpretations are the morphisms of $\mathsf{INT}^{\mathsf{unr},=}$. Here is an immediate insight.

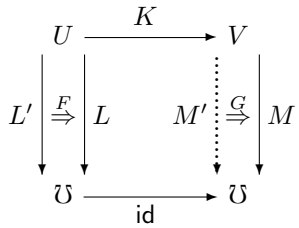
Theorem 6.1 *If $M \circ K$ in INT is direct, then M is direct.*

Here is a first characterization of direct interpretations. Let $\Im$ be predicate logic with just the identity symbol. For any theory T we define the morphism $\mathcal{I}_T : \Im \to T$ by: $\delta_{\mathcal{I}_T} :\leftrightarrow v_0 = v_0$ and $v_0 E_{\mathcal{I}_T} v_1 :\leftrightarrow v_0 = v_1$. Thus, $\Im$ is a weak initial object. The following theorem is obvious.

Theorem 6.2 *$K : U \to V$ is direct iff $K \circ \mathcal{I}_U = \mathcal{I}_V$.*

It turns out that we can characterize direct interpretations fully in terms of $\mathsf{INT}^{\mathsf{iso}}$. We need some preliminary definitions to do this. Consider a functor $\phi : \mathcal{C} \to \mathcal{D}$. The functor ϕ is *a discrete fibration* if, for every c in $\mathcal{C}$ and for every $g : d \to \phi(c)$, there is a unique f in $\mathcal{C}$ with $\phi(f) = g$ and $\mathsf{cod}(f) = c$. We will write $\overline{g}(c) : g^*(c) \to c$ for the unique f corresponding to g and c. If we want to make the role of ϕ explicit, we will write, e.g., $\overline{g}_\phi(c)$.[16]

A morphism K in INT is *forward looking* iff, for every $\mathfrak{V}$, the contravariant functor $[\![K]\!]_{\mathfrak{V}}$ is a discrete fibration. Setting $L := M \circ K$, $L' := M' \circ K$, $M' := F^*_{[\![K]\!]_{\mathfrak{V}}}(M)$ and $G := \overline{F}_{[\![K]\!]_{\mathfrak{V}}}(M)$, we have the following diagram.



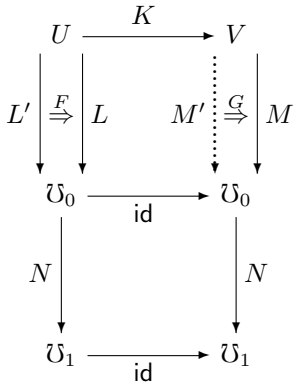
Here F and G are i-isomorphisms. The next theorem shows what happens if we move via a morphism to another 'base theory'.

[16] For an extensive presentation of the theory on fibrations, see [Jac99]. In this paper, we will only use a few trivial facts.

Theorem 6.3 *Suppose K is forward looking and that $N : \mathfrak{V}_0 \to \mathfrak{V}_1$. We have:*

- $(N \circ F)^*_{[\![K]\!]_{\mathfrak{V}_1}}(N \circ M) = N \circ F^*_{[\![K]\!]_{\mathfrak{V}_0}}(M)$,
- $\overline{(N \circ F)}_{[\![K]\!]_{\mathfrak{V}_1}}(N \circ M) = N \circ \overline{F}_{[\![K]\!]_{\mathfrak{V}_0}}(M)$.

The proof of the theorem is easily seen by contemplating the following diagram and using the uniqueness clause in the definition of discrete fibration.

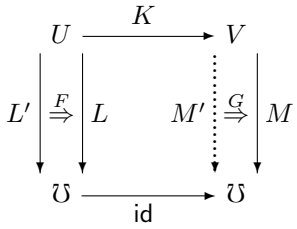


We prove the promised characterization of direct interpretations.

Theorem 6.4 *An interpretation is direct iff it is forward looking.*

Proof

We first prove the left-to-right direction. Suppose we are given a theory $\mathfrak{V}$, a direct interpretation $K : U \to V$, interpretations $M : V \to \mathfrak{V}$, $L' : U \to \mathfrak{V}$, and an i-isomorphism $F : L' \Rightarrow L$, where $L := M \circ K$. We have to show that there is a unique pair $M' : V \to \mathfrak{V}$, $G : M' \Rightarrow M$ such that $L' = M' \circ K$, $G = F \circ K$.



Note that $\mathfrak{V}$ will prove that δ_L is δ_M and E_L is E_M. We define M' and G as follows.

- $\delta_{M'} :\leftrightarrow \delta_{L'}$,
- $P_{M'}(\vec{v}) :\leftrightarrow \exists \vec{w}\, (\vec{v} F \vec{w} \wedge P_M(\vec{w}))$.

- $G :\leftrightarrow F$.

The verification that M' and G indeed uniquely have the desired property holds no surprises.

We turn to the right-to-left direction. Suppose $K : U \to V$ is forward looking. We define a theory W as follows. The signature of W is the signature of V plus a fresh unary predicate symbol Δ and a fresh binary predicate symbol G. Par abus de langage we will use Δ for the translation of V-formulas by their relativizations to Δ in the language of W and also for the interpretation from V to W carried by the translation Δ. The theory W will be axiomatized by:

- A^Δ, for all axioms A of V,
- an axiom expressing that G is a permutation of the domain of W,
- two axioms expressing that G 'leaves $\Delta \circ K$ fixed', i.e.:
  - $\vdash (xGy \wedge x : \Delta \wedge x : \delta_K^\Delta) \to (y : \Delta \wedge y : \delta_K^\Delta \wedge xE_K^\Delta y)$,
  - $\vdash (xGy \wedge y : \Delta \wedge y : \delta_K^\Delta) \to (x : \Delta \wedge x : \delta_K^\Delta \wedge xE_K^\Delta y)$.

We define an interpretation $M : V \to W$ as follows.

- $x : \delta_M :\leftrightarrow \exists y\, (xGy \wedge y : \Delta)$,
- $P_M(\vec{x}) \leftrightarrow \exists \vec{y}\, (\vec{x}G\vec{y} \wedge P(\vec{y}))$.

As is easily seen, we indeed have: $M : V \to W$ and $G : M \Rightarrow \Delta$. Let $L := \Delta \circ K$ and $F := G \circ K$. Since G 'leaves L fixed', we find that F is ID_L. Note that also $\mathsf{ID}_\Delta \circ K = \mathsf{ID}_L$. Ergo, since K is forward looking, we find that $G = \mathsf{ID}_\Delta$ and $M = \Delta$.

To arrive at a contradiction, suppose K were not direct. Then, there is a model $\mathcal{M}$ of V in which either $\delta_K^\mathcal{M}$ is not the full domain of $\mathcal{M}$, or where $E_K^\mathcal{M}$ has a non-trivial equivalence class. In the first case, we can extend $\mathcal{M}$ to a model $\mathcal{N}$ of W as follows.

- The domain of $\mathcal{N}$ is the domain of $\mathcal{M}$ plus one new element a.
- $\Delta^\mathcal{N}$ is the domain of $\mathcal{M}$.
- The $R^\mathcal{N}$ restricted to $\Delta^\mathcal{N}$ are the $R^\mathcal{M}$; the other choices are don't care, except in the case of identity.
- $G^\mathcal{N}$ is the transposition of one element of $\Delta^\mathcal{N} \setminus \delta_K^{\Delta\mathcal{N}}$ and the new element a.

Clearly, $G^\mathcal{N}$ is not $\mathsf{ID}_\Delta^\mathcal{M}$. In the second case, the construction is similar. We need not extend the domain and take $G^\mathcal{N}$ a transposition of two elements of a non-trivial equivalence class of $E_K^\mathcal{N}$.

We arrive at a contradiction. Hence, K must be direct. ❑

Remark 6.5 Note that, in the proof of the right-to-left direction of Theorem 6.4, we only use the uniqueness clause in the definition of discrete fibration. Thus, in reality, we prove a stronger theorem. ❑

6.2 hINT meets INT

In this subsection, we illustrate that, for direct arrows, we can often transfer properties of morphisms from hINT to INT or from INT to hINT.

Theorem 6.6 *Direct epimorphisms in* INT *are epimorphisms in* hINT.

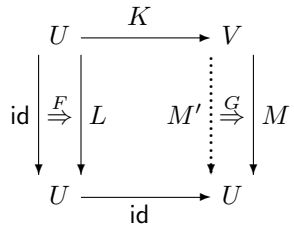
Proof

Let $K : U \to V$ and $M, M' : V \to W$. Suppose K is a direct epimorphism in INT. Suppose further that we have an isomorphism $F : M' \circ K \Rightarrow M \circ K$. Since K is forward looking, we can find an $M'' : V \to W$ and a $G : M'' \Rightarrow M$ such that $M' \circ K = M'' \circ K$ and $G \circ K = F$. Since K is an epimorphism in INT, it follows that $M' = M''$. Hence M' is i-isomorphic to M. ❑

Theorem 6.7 *Consider $K : U \to V$. Suppose K is direct. Then, K is a split monomorphism in* INT *iff K is a split monomorphism in* hINT.

Proof

Suppose K is direct. It is clear that, if K is a split monomorphism in INT, then K is a split monomorphism in hINT. We prove the converse. Suppose K is a split monomorphism in hINT. Let M be the corresponding split epimorphism in hINT. So, we have an isomorphism $F : \mathsf{id}_U \Rightarrow M \circ K$. Let $L := M \circ K$. We have:



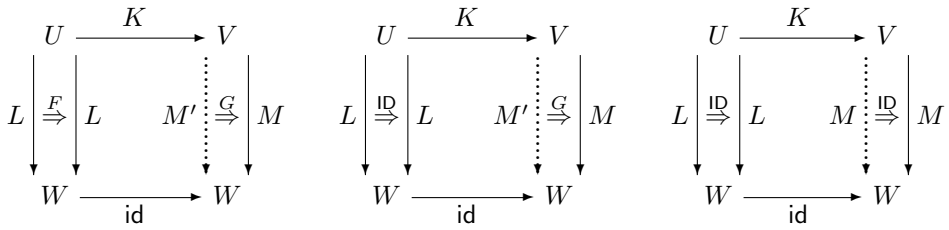
So we can take $M' := F^{\star}_{[\![K]\!]_U}(M)$, as the split epimorphism corresponding to K in INT. ❑

A morphism in $\mathsf{INT}^{\mathsf{iso}}$ is *rigid* iff it has no non-trivial i-automorphisms.

Lemma 6.8 *Consider $K : U \to V$ and $M, M' : V \to W$. Suppose $G : M' \Rightarrow M$ is an isomorphism. Suppose further that $L := M \circ K = M' \circ K$. Suppose L is rigid and K is direct. Then, $M = M'$, and $G = \mathsf{ID}_M$. In other words, for direct K, the $[\![K]\!]_W$-fiber over a rigid L consists of a set of disconnected rigid M's.*

Proof

Let $F := G \circ K$. The corresponding i-arrows in the following three diagrams must be the same.



By rigidity, $F = \mathsf{ID}_U$. By the discreteness of $[\![K]\!]_W$, $G = \mathsf{ID}_V$. ❑

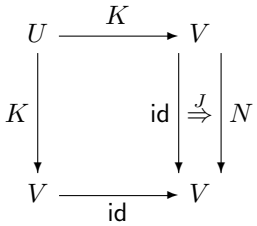
Theorem 6.9 *Suppose K is rigid and direct. Then K is an isomorphism in* INT *iff K is an isomorphism in* hINT.

Proof

Suppose K is rigid and direct. It is clear that, if K is an isomorphism in INT, then K is an isomorphism in hINT. For the converse, suppose that K is an isomorphism in hINT. Suppose $M : V \to U$ and F and H are isomorphisms such that $F : \mathsf{id}_U \Rightarrow M \circ K$ and $H : \mathsf{id}_V \Rightarrow K \circ M$. As in the proof of Theorem 6.7, we can find $M' : V \to U$ and an isomorphism $G : M' \Rightarrow M$ such that $\mathsf{id}_U = M' \circ K$ and $F = G \circ K$. So we have:

$$\mathsf{id}_V \overset{H}{\Longrightarrow} K \circ M \overset{K \circ G^{-1}}{\Longrightarrow} K \circ M'$$

Thus, $J := (K \circ G^{-1}) \cdot H$ is an isomorphism between id_V and $N := K \circ M'$. Thus, we have:



Moreover, $N \circ K = K \circ M' \circ K = K \circ \mathsf{id}_U = K$. So we may apply Lemma 6.8, to obtain $M' \circ K = N = \mathsf{id}_V$. ❑

In Subsection 6.4, we apply Theorem 6.9 to the Ackermann Interpretation.

6.3 Improving Interpretations

In this subsection, we will treat some well-known constructions to replace interpretations by i-isomorphic counterparts having some extra desired properties like directness.

Consider $K : U \to V$. Let U^c be U in the language of U extended with a constant c. Let $\mathsf{emb}_{U,U^c} : U \to U^c$ be the standard embedding of U in U^c. We say that K admits a constant iff $K \models ?\mathsf{emb}_{U,U^c}$.

Theorem 6.10 *Suppose K admits a constant. Then, there is an unrelativized interpretation $K^\star$ i-isomorphic to K.*

Proof

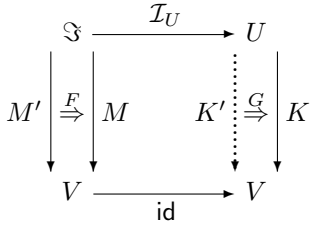
Suppose A defines the promised constant. We define the equivalence relation $E^\star$ in V as follows:

$$\begin{aligned} v_0 E^\star v_1 \quad :\leftrightarrow \quad & (\neg\,(v_0 : \delta_K) \wedge \neg\,(v_1 : \delta_K)) \vee \\ & (\neg\,(v_0 : \delta_K) \wedge v_1 : \delta_K \wedge A v_1)) \vee \\ & (v_0 : \delta_K \wedge \neg\,(v_1 : \delta_K) \wedge A v_0) \vee \\ & (v_0 : \delta_K \wedge v_1 : \delta_K \wedge v_0 E_K v_1) \end{aligned}$$

We define $M' : \Im \to V$ and $F : M' \Rightarrow K \circ \mathcal{I}_U$ by:

- $\delta_{M'} :\leftrightarrow v_0 = v_0$,
- $E_{M'} :\leftrightarrow E^\star$,
- $v_0 F v_1 :\leftrightarrow \delta_K(v_1) \wedge v_0 E^\star v_1$.

Let $M := K \circ \mathcal{I}_U$. Now consider the following diagram, noting that $\mathcal{I}_U$ is direct and, hence, forward looking.



We can take $K' := F^*_\phi(K)$, where $\phi = [\![\mathcal{I}_U]\!]_V$. Clearly, K' is unrelativized. Moreover, it is i-isomorphic to K via $G := \overline{F}_\phi(K)$. ❑

Example 6.11 Consider the theory $T = \mathbf{3} \boxplus \mathbf{1}$.[17] We have: $\mathsf{in}_0^* : \mathbf{3} \to T$. However, by a simple modeltheoretic argument, there is no unrelativized interpretation of $\mathbf{3}$ in T. ❑

[17] The operation $\boxplus$ is defined in Subsection 4.6. The theory $\mathbf{3}$ is the obvious theory in the language of pure identity stating that there are precisely three objects.

Remark 6.12 If we would consider interpretations *with parameters* (see Subsection B.3), we do not need the assumption of a 'constant' in Theorem 6.10. We can always 'unrelativize' using a parameter. See [MPS90]. ❑

Theorem 6.13 *Suppose V is an extension of* PA *in the language of* PA. *Consider $K : U \to V$. Suppose V proves that the domain of K modulo E_K is infinite. Then, K is i-isomorphic to a direct interpretation $K^\star$.*

Proof

Let V be an extension of PA in the language of PA. Consider $K : U \to V$. Suppose V proves that the domain of K modulo E^K is infinite. Define:

- $fx := \mu y : \delta_K \cdot \forall x' {<} x \; fx' \neq^K y$.
- $v_0 F v_1 :\leftrightarrow v_1 : \delta_K \wedge f v_0 E_K v_1$.

From this point on the proof proceeds like the proof of Theorem 6.10. ❑

Open Question 6.14 Is there a direct interpretation of $T + \mathsf{incon}(T)$ into T, where $T \in \{\mathsf{Q}, \mathsf{S}^1_2, \mathsf{EA}, I\Sigma_1, \mathsf{ACA}_0, \mathsf{GB}\}$? Here $\mathsf{incon}(T)$ is given some standard arithmetization and, in ACA_0 and GB we employ the usual interpretation of arithmetic. ❑

6.4 The Ackermann Interpretation

Let ZF^- be set theory without the axiom of infinity. We show that PA is a retract (in INT) of ZF^-.

First we interpret ZF^- in arithmetic. This employs an interpretation $\mathsf{ackermann}$, or, in short, A, first found by Ackermann. A number in binary is read, from right to left, as the characteristic function of a finite set of numbers which in their turn again are taken to code sets. Thus, A is given as follows.

- $\delta_{\mathsf{A}} :\leftrightarrow v_0 = v_0$.
- $v_0 E_{\mathsf{A}} v_1 :\leftrightarrow v_0 = v_1$.
- $v_0 \in_{\mathsf{A}} v_1 :\leftrightarrow \exists x, y \; (v_1 = (2x+1)2^{v_0} + y \wedge y < 2^{v_0})$.
  When no confusion is possible, we will simply write $\in$ for $\in_{\mathsf{A}}$.

The interpretation A is not faithful, since it also interprets the negation of Inf, the axiom of infinity. Clearly, A is direct. Also, A is rigid because, PA-verifiably, if an i-automorphism F of A is the identity on the $\in_{\mathsf{A}}$-elements of x, then $F(x)$ must be x by extensionality. In the converse direction, we proceed as follows. We work in ZF^-, using some well-established abbreviations. Define:

- $0 := \emptyset$,
- $\mathsf{S}v_0 := v_0 \cup \{v_0\}$,

- $\mathsf{prenum}^{v_0}(v_1) :\leftrightarrow v_1 = 0 \vee \exists x{\in}v_0\ v_1 = \mathsf{S}x$,
- $v_0 : \omega :\leftrightarrow \forall y{\in}\mathsf{S}v_0\ \mathsf{prenum}^{v_0}(y)$.

We run through a series of lemmas.

Lemma 6.15 $[\mathsf{ZF}^-]$ *0 is in ω, 0 is not a successor.*

Lemma 6.16 $[\mathsf{ZF}^-]$ *The successor function is injective.*

Lemma 6.17 $[\mathsf{ZF}^-]$ *ω is closed under predecessor, i.o.w. if $\mathsf{S}u : \omega$, then $u : \omega$.*

Proof

Suppose $\mathsf{S}u : \omega$ and $v \in \mathsf{S}u$. It is sufficient to show: $\mathsf{prenum}^u(v)$.

We certainly have $v \in \mathsf{SS}u$. Hence, since $\mathsf{S}u : \omega$, we have $\mathsf{prenum}^{\mathsf{S}u}(v)$. This means that $v = 0$ or, for some $w \in \mathsf{S}u$, $v = \mathsf{S}w$. In the first case, we are immediately done. So suppose for some $w \in \mathsf{S}u$, $v = \mathsf{S}w$. We have either $w \in u$ or $w = u$. In the first case we are done. In the second case, we have $v = \mathsf{S}u \in \mathsf{S}u$. A contradiction. ❑

Lemma 6.18 $[\mathsf{ZF}^-]$ $z : \omega \leftrightarrow (z = 0 \vee \exists u : \omega\ z = \mathsf{S}u)$.

Proof

Suppose $z : \omega$. Then, since $z \in \mathsf{S}z$, we have $\mathsf{prenum}^z(z)$. Hence, $z = 0$ or $z = \mathsf{S}(u)$, for some u. In the first case, we are done by Lemma 6.15. In the second case we are done by Lemma 6.4.

For the converse direction, suppose $z = 0$ or $z = \mathsf{S}u$, for some $u \in \omega$. In the first case we are done by Lemma 6.15. In the second case we have to show that, for all $v \in \mathsf{S}z$, we have $\mathsf{prenum}^z(v)$. Suppose $v \in \mathsf{S}(z)$. In case $v \in z$, we have $v \in \mathsf{S}u$, and hence $\mathsf{prenum}^u(v)$. hence, a fortiori, $\mathsf{prenum}^z(v)$. In case $v = z$, we clearly have $\mathsf{prenum}^z(z)$, since $z = \mathsf{S}u$. ❑

Lemma 6.19 $[\mathsf{ZF}^-]$ *We have induction on ω.*

Proof

We prove induction on ω. Reason in ZF^-. Suppose:

$$A0 \text{ and } \forall x{\in}\omega\ (Ax \to A\mathsf{S}x).$$

Suppose further that, for some z, we have in $z \in \omega$ and $\neg\, Az$. Let z^* be $\in$-minimal with the property. Since $z^* \in \omega$ and $z^* \neq 0$, we have $z^* = \mathsf{S}u$ for some $u \in z^*$. Since, by Lemma 6.4, u is in ω it follows that Au. But, then, Az. ❑

Now we can develop the theory of plus and times in the usual way, defining sequences as functions from elements of ω to sets. Thus we have an interpretation, $\mathsf{neumann}$, or, in short, N, of PA in ZF^-.

It is now easy to see that $\mathsf{N} \circ \mathsf{A} : \mathsf{PA} \to \mathsf{PA}$ is i-isomorphic to the identity interpretation on PA. The mapping, on which the i-isomorphism, say $\mathcal{I}$, is based, is given by the following recursion:

- $\mathcal{I}0 := 0$,
- $\mathcal{I}\mathsf{S}x := 2^{\mathcal{I}x} + \mathcal{I}x$.

Thus, we have shown that $\mathsf{A} : \mathsf{ZF}^- \to \mathsf{PA}$ is a split monomorphism in hINT. Thus, this interpretation is faithful, since, by Theorem 4.5, monomorphisms in hINT are faithful. It follows, by Theorem 6.7, that A is a split monomorphism in INT and, hence, by Theorem 6.1, A is a split monomorphism in $\mathsf{INT}_{\mathsf{unr},=}$.

Now, let HF be $\mathsf{ZF}^- + \neg\,\mathsf{Inf}$. Our two translations also support morphisms $\mathsf{A}^+ : \mathsf{HF} \to \mathsf{PA}$ and $\mathsf{N}^+ : \mathsf{PA} \to \mathsf{HF}$. By our preceding result $\mathsf{A}^+ \circ \mathsf{N}^+$ is i-isomorphic to $\mathsf{id}_{\mathsf{PA}}$. We show that $\mathsf{N}^+ \circ \mathsf{A}^+$ is i-isomorphic to $\mathsf{id}_{\mathsf{HF}}$. Consider the mapping $\mathcal{J}$ from ω to sets given by:

- $\mathcal{J}x := \{\mathcal{J}y \mid y \in^{\mathsf{A}^+\mathsf{N}^+} x\}$.

Note that $\mathcal{J}x$ is indeed a set by Replacement, since $y \in^{\mathsf{A}^+\mathsf{N}^+} n$ implies $y <^{\mathsf{N}^+} x$, which in its turn implies $y \in x$. We can show that $\mathcal{J}$ is injective by $\in$-induction. Now we want to show that the image of $\mathcal{J}$ is all sets. Suppose not. Let z^* be an $\in$-minimal element not in the image of $\mathcal{J}$. If $\mathcal{J}^{-1}[z^*]$ is bounded by a number x, we can easily construct a number y with $\mathcal{J}y = z^*$. So $\mathcal{J}^{-1}[z^*]$ is unbounded in ω. We find that $\mathcal{J}^{-1}[z^*]$ is a set, since $\mathcal{J}^{-1}$ is a function. The union of this set is ω, which will also be a set. Quod non. The further verification that $\mathcal{J} : (\mathsf{N}^+ \circ \mathsf{A}^+) \Rightarrow \mathsf{id}_{\mathsf{HF}}$ is an i-isomorphism holds no surprises.

We may conclude that HF and PA are bi-interpretable and, hence, by Theorem 6.9, synonymous.

Open Question 6.20 It is clear that the Ackermann translation makes sense in Elementary Arithmetic EA aka $I\Delta_0 + \mathsf{Exp}$. So what precise set theory corresponds with EA? ❑

Remark 6.21 The Ackermann translation is for many purposes a good translation, but it has as disadvantage that e.g. the singleton function is exponential: $\{x\}$ is coded as 2^x. This makes it unsuitable for working in weak theories. There are other translations, lacking many of the good properties of the Ackermann translation, for which the code $\{x\}$ is of order x^2 and for which union is of the order of multiplication. ❑

7 Restricted Interpretations

In this section we explore what Tarski's theorem means modulo interpretation.[18] A first question is: what precisely is meant by *object-language* and *meta-language*. This question becomes more interesting if we do not demand that the object-language is part of the meta-language, but just that the object-language is translatable into the meta-language. A first point is that the talk about 'language' is misleading: these 'languages' are really *theories*: object-theory and meta-theory. Secondly, I submit, we do not want to existentially abstract away from the translation. Thus, the right explication of *candidate object-language/meta-language pair* is *morphism in* INT_i, for a chosen $i = 0, 1, 2, 3$. The next step is to define *(succesful) object-language/meta-language pair*. This will be explained as the satisfaction of a certain uniform e-scheme.

Let $\mathsf{Q}^{\mathsf{true}}$, be Robinson's arithmetic extended with a unary predicate true (and no further axioms concerning true). Consider any theory U. We extend U to U^{meta} which is $U \oplus \mathsf{Q}^{\mathsf{true}}$ plus the axioms $\mathsf{true}^{\mathsf{in}_1}(\underline{\#A}) \leftrightarrow A^{\mathsf{in}_0}$, for all U-sentences A. We will notationally suppress the superscript in_i, writing the T-scheme simply as $\mathsf{true}(\underline{\#A}) \leftrightarrow A$. Let $\mathsf{om}_U := \mathcal{E} \circ \mathsf{in}_0$. Here '$\mathsf{om}$' stands for the object-meta embedding. Suppose $K : U \to V$. We can represent the mapping $A \mapsto A^K$ in arithmetic. We will call this arithmetization $(\cdot)^\kappa$.

We may extend K to K^{meta} by interpreting the U-predicates of U^{meta} via K, by interpreting the Q-predicates via the identity, and by interpreting true of U^{meta} by $\mathsf{true}(v_0^\kappa)$ in V^{meta}. We will have, for any U-sentence A,

$$\begin{array}{rcl} V^{\mathsf{meta}} \vdash \mathsf{true}^{K^{\mathsf{meta}}}(\underline{\#A}) & \leftrightarrow & \mathsf{true}((\underline{\#A})^\kappa) \\ & \leftrightarrow & \mathsf{true}(\underline{\#(A^K)}) \\ & \leftrightarrow & A^K \\ & \leftrightarrow & A^{K^{\mathsf{meta}}} \end{array}$$

So, $K^{\mathsf{meta}} : U^{\mathsf{meta}} \to V^{\mathsf{meta}}$. Moreover, we will clearly have $\mathsf{om}_V \circ K = K^{\mathsf{meta}} \circ \mathsf{om}_U$. Ergo, ?$\mathsf{om}$ will be a uniform e-scheme (in each of our categories).

Remark 7.1 One might hope that $U \mapsto U^{\mathsf{meta}}$, $K \mapsto K^{\mathsf{meta}}$ would be a functor and that om would be a natural transformation from ID to $(\cdot)^{\mathsf{meta}}$. However, regrettably, this fails. The mapping $K \mapsto K^{\mathsf{meta}}$ does not necessarily preserve equality of arrows in our categories. So, on the level of interpretations, it is not necessarily a mapping. ❑

An interpretation $K : U \to V$ is *restricted* if $K \models ?\mathsf{om}$.

Lemma 7.2 *Suppose* $L =_3 L'$ *and* $L \models^0 ?\mathsf{om}$. *Then* $L' \models^0 ?\mathsf{om}$.

[18] The need to connect the theory of the definability of truth with the study of interpretations was clearly seen by Panu Raatikainen. An abstract containing some of his ideas can be found at ⟨`http://www.math.helsinki.fi/logic/LC2003/abstracts/`⟩ Alternatively, see: Panu Raatikainen: 'Translation and the definability of truth (Abstract)' in: Logic Colloquium 2003, Helsinki Finland, August 14-20, Abstracts. Yliopistopaino, Helsinki, 2003.

Proof

Consider $L, L' : U \to V$. Suppose $L =_3 L'$ and $L \models^0 ?\mathsf{om}$. Say M witnesses $L \models^0 ?\mathsf{om}$, i.e. $M \circ \mathsf{om}_U = L$. Consider $M_0' : U \oplus \mathsf{Q}^{\mathsf{true}} \to V$ defined by $M_0' := L' \oplus (M \circ \mathsf{in}_1)$. We get:

$$\begin{array}{rcl} V \vdash \mathsf{true}^{M_0'}(\underline{\#A}) & \leftrightarrow & \mathsf{true}^{M}(\underline{\#A}) \\ & \leftrightarrow & A^L \\ & \leftrightarrow & A^{L'} \\ & \leftrightarrow & A^{M_0'} \end{array}$$

❑

Thus M_0' 'lifts' to an interpretation $M' : U^{\mathsf{meta}} \to V^{\mathsf{meta}}$. It is easy to see that M' witnesses the fact that $L' \models^0 ?\mathsf{om}$. Thus, by Theorem 5.4, we find that K is restricted in INT_i iff it is restricted in INT_j, for any $i, j \in \{0, 1, 2, 3\}$.

Theorem 7.3 *We have in* INT_i*, where* $i = 0, 1, 2, 3$*, that* id_U *is not restricted.*

Proof

Suppose M witnesses the fact that id_U were restricted. By the Gödel Fixed Point Lemma, we can find a sentence L such that $U \vdash L \leftrightarrow \neg\, \mathsf{true}^M(\underline{\#L})$. This leads immediately to a contradiction with the T-scheme. ❑

Theorem 7.4 *Suppose* $K : U \to V$*,* $M : V \to W$ *in* INT_i*, for* $i = 0, 1, 2, 3$*. We have:*

1. *If* K *is restricted, then so is* $M \circ K$*.*
2. *If* M *is restricted, then so is* $M \circ K$*.*
3. *If* $M \circ K$ *is restricted and* K *is surjective, then* M *is restricted.*

Proof

Suppose $K : U \to V$, $M : V \to W$. Items (1) and (2) are special cases of Theorem 5.5. We prove (3). Suppose that P witnesses that $M \circ K$ is restricted. Suppose further that K is surjective. Clearly, there is a recursive function f such that $f(\#B)$ is some A such that $V \vdash B \leftrightarrow A^K$. Let ν be a bi-representation of f in Q. We construct Q witnessing the fact that M is restricted as follows.

- We interpret Q according to P.
- $\mathsf{true}^Q(v_0) :\leftrightarrow \mathsf{true}^P(\nu(v_0))$.
- We interpret V according to M.

Suppose $f(\#B) = \#A$. We have:

$$\begin{array}{lcl} W \vdash \ \mathsf{true}^Q(\underline{\#B}) & \leftrightarrow & \mathsf{true}^P(\nu(\underline{\#B})) \\ & \leftrightarrow & \mathsf{true}^P(\underline{\#A}) \\ & \leftrightarrow & A^{KM} \\ & \leftrightarrow & B^M \\ & \leftrightarrow & B^Q \end{array}$$

So $Q : V^{\mathsf{meta}} \to W$. It is easy to see that $Q \circ \mathsf{om}_V = M$. ❑

Corollary 7.5 *In each of our categories* INT_i*, for* $i = 0, 1, 2, 3$*, we have: no split monomorphism is restricted. Similarly, no split epimorphism is restricted.*

From the philosophical point of view, we think, that the statement that no split monomorphism is restricted, is a good statement of Tarski's Theorem of the undefinability of truth, where we take into account the fact that the object-language is translated into the metalanguage. The reasonable condition on this translation is that it is a split monomorphism or co-retraction.

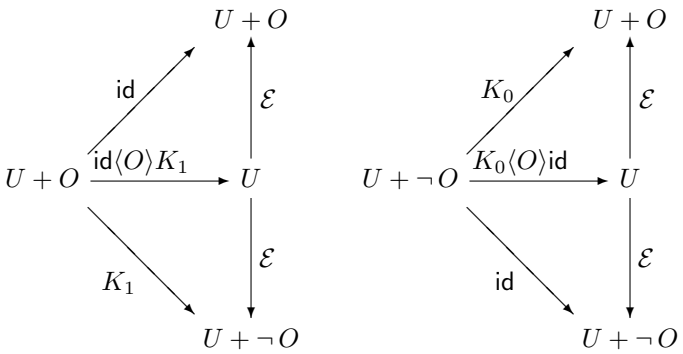
Corollary 7.6 *No surjective morphism is restricted (in* INT_i*, for* $i = 0, 1, 2, 3$*).*

Proof

Suppose $K : U \to V$ is surjective and restricted. It follows, by Theorem 7.4, that id_V is restricted. Quod non, by Theorem 7.3. ❑

Restricted interpretations $K : U \to U$ give rise to the Orey phenomenon. We have restricted $K : U \to U$ e.g., for reflexive theories, like PRA, PA and ZF, and, for finitely axiomatized sequential theories, like Q, S^1_2, EA, $I\Sigma_1$, ACA_0 and GB. (All these examples are constructed via the Henkin construction.)

Suppose $K : U \to U$ is restricted. Let O satisfy: $U \vdash O \leftrightarrow \neg\, \mathsf{true}^{K^{\mathsf{meta}}}(\underline{\#O})$. We find: $U \vdash O \leftrightarrow \neg O^K$. So there are interpretations K_0 and K_1 with the same underlying translation as K, such that $K_0 : U + \neg O \to U + O$ and $K_1 : U + O \to U + \neg O$. Thus, in each of our categories the following diagrams commute:



So, there are split monomorphisms both from $U + O$ and from $U + \neg O$ to U. Note that it follows that $U + O$ and $U + \neg O$ are both faithfully interpretable in U. I feel that it is remarkable that the Orey phenomenon occurs even for such a strict notion as split monomorphism.

Open Question 7.7 Do we have a restricted $K : T \to T$, for all sequential T? ❑

8 i-Initial Arrows

In this section, we study the meaning of the existence of i-initial arrows in $\mathsf{INT}^{\mathsf{morph}}$. An arrow $K : U \to V$ is *i-initial* iff it is initial in the category of arrows $M : U \to V$ with the i-morphisms $F : M \Rightarrow M'$.

We fix a weak arithmetic F. We could choose e.g. Robinson's Arithmetic Q or Buss's Arithmetic S^1_2 (or, rather, an appropriate variant of S^1_2 in the arithmetical language) or $I\Delta_0 + \Omega_1$ (aka S_2). Consider the following theories.

- F^X, i.e. Robinson's Arithmetic with an extra unary predicate symbol X n the signature, but with no further axioms;
- $\mathsf{F}^{\mathsf{ind}} := \mathsf{Q}^X + \mathsf{IND}(X)$, where $\mathsf{IND}(X)$ is the principle of induction over X.

We have the obvious embeddings $\mathsf{emb} := \mathsf{emb}_{\mathsf{F},\mathsf{F}^X}$ of F in F^X and $\mathcal{E} := \mathcal{E}_{\mathsf{F},\mathsf{F}^{\mathsf{ind}}}$ of F^X into $\mathsf{F}^{\mathsf{ind}}$. We can now say that an interpetation $K : \mathsf{F} \to U$ satisfies full induction iff $K \models (\mathsf{emb} \twoheadrightarrow \mathcal{E})$.[19]

Theorem 8.1 *Suppose $\iota : \mathsf{F} \to U$ is i-initial. Then, ι satisfies induction.*

Proof

Suppose $\iota^X : \mathsf{F}^X \to U$ and $\iota = \iota^X \circ \mathsf{emb}$. We write:

- $\Omega :\leftrightarrow \delta_\iota$.
- $\mathsf{PROG}(X) :\leftrightarrow (X0 \wedge \forall x\, (Xx \to X\mathsf{S}x))$.

Let:

- $B(v_0) :\leftrightarrow (v_0 : \Omega \wedge (\mathsf{PROG}(X) \to Xv_0)^{\iota^X})$.

Let $K : \mathsf{F}^X \to U$ be the extension of ι that interprets X as B. Clearly, $U \vdash (\mathsf{PROG}(X))^K$. Now —and here we use the fact that F is a weak theory— we apply Solovay's method of shortening cuts (see e.g. [HP91]) to obtain a formula $C(v_0)$ such that $U \vdash C \to B$ and such that we relativizing to C yields an interpretation of F. Specifically, C is such that we have $\mathsf{R}_C : \mathsf{F} \to U$, where R_C is defined as follows.

[19] Strictly speaking this defines only induction without parameters, but, as is well-known, *full* induction without parameters implies full induction with parameters.

- $\delta_{\mathsf{R}_C} :\leftrightarrow C(v_0)$,
- $P_{\mathsf{R}_C}(v_0, \ldots, v_{n-1}) :\leftrightarrow P_\iota(v_0, \ldots, v_{n-1})$,
  for any arithmetical predicate P of F.

We have an i-morphism $\mathsf{e}_C : \mathsf{R}_C \to \iota$ given by:

- $v_0(\mathsf{e}_C)v_1 :\leftrightarrow (C(v_0) \wedge v_0 E_\iota v_1)$.

By i-initiality, there is an arrow $F : \iota \Rightarrow \mathsf{R}_C$. We find: $\mathsf{e}_C \circ F : \iota \Rightarrow \iota$. By the uniqueness clause of i-initiality, we must have: $(\mathsf{e}_C \circ F) = \mathsf{ID}_\iota$. So, we have in U:

$$\begin{array}{rcl} v_0 E_\iota v_1 & \leftrightarrow & v_0(\mathsf{e}_C \circ F)v_1 \\ & \leftrightarrow & \exists y\ (v_0 F y \wedge y(\mathsf{e}_C)v_1) \\ & \leftrightarrow & \exists y\ (v_0 F y \wedge C(y) \wedge y E_\iota v_1) \\ & \leftrightarrow & v_0 F v_1 \end{array}$$

It follows that $U \vdash \forall v_0 : \Omega\ C(v_0)$, and hence $U \vdash \forall v_0 : \Omega\ B(v_0)$, which is equivalent to the induction principle for ι^X. ❑

Open Question 8.2 Is there an example of an arrow $\iota : \mathsf{F} \to U$ that is *weakly i-initial*, but that does not satisfy full induction? ❑

In case U is sequential, we have a converse of Theorem 8.1. The notion of sequentiality is due to Pavel Pudlák. See, e.g., [HP91] or our Section 10. The idea is that U 'has sequences of all objects of the domain (including the sequences)'. These sequences are not extensional: two sequences may be different even if they have the same projections. The numbers w.r.t. which we project are given by some interpretation, say N, of Q in U. We will need three important properties of sequences.

- There is an empty sequence.
- Given a sequence σ and an object a, we may form $\sigma * \langle a \rangle$.
- Given a sequence σ of length n, and a number $k \leq n$, there is a sequence τ of length k, such that $(\sigma)_i = (\tau)_i$, for all $i < k$.[20]

Theorem 8.3 *Suppose $\iota : \mathsf{F} \to U$ satisfies full induction and $K : \mathsf{F} \to U$.*

1. *Suppose $F, G : \iota \Rightarrow K$. then $F = G$.*
2. *Suppose U is sequential, then there is an $F : \iota \Rightarrow K$.*

It follows that, for sequential U, ι is i-initial.

[20] This property is not among the properties stipulated in Section 10. However, we can obtain it by switching to a 'better' set of numbers for the projections.

Proof

The proof of (1) is easy. We sketch the proof of (2). Suppose U is sequential and that $\iota : \mathsf{F} \to U$ satisfies full induction. Consider $K : \mathsf{F} \to U$. We have to produce $F : \iota \Rightarrow K$. We work in U.

Note that we have to work with three number systems. There are the N-numbers that are used in taking projections from sequences. There are the ι-numbers that satisfy full induction. And there are the K-numbers about which we do not know much. We wil write Ω for δ_ι.

We define approximations of the desired F as follows. An approximation is a sequence of pairs where the first components are from Ω and the second components are from δ_K. The first elements of approximations are pairs of zeros. Successor elements of approximations are pairs of successors (in the respective number systems) of the preceding pair. We take $v_0 F v_1$ iff there is an approximation ending in the pair $\langle v_0, v_1 \rangle$. Now we may prove, by induction, that F is a total relation. We briefly look at the argument for functionality. We prove by induction on x in Ω that:

$$\forall x' : \Omega \; \forall y, y' : \delta_K \; ((xFy \wedge x'Fy' \wedge xE_\iota x') \to yE_K y').$$

Suppose first $\mathsf{Z}_\iota(x)$, xFy, $x'Fy'$, $xE_\iota x'$. Consider an approximating sequence σ for xFy. The N-length of this sequence is either 1 or bigger than 1. In the first case, we find $\mathsf{Z}_K(y)$. In the second case, we would have that x is a ι-successor. Quod non. Since $xE_\iota x'$, we find $\mathsf{Z}_\iota(x')$. By copying the above reasoning, it follows that $\mathsf{Z}_K(y')$. We may conclude that $yE_K y'$.

Next suppose x is a ι-successor and xFy, $x'Fy'$, $xE_\iota x'$. The Induction Hypothesis, tells us that we have the desired property for and ι-predecessor of x. Consider an approximating sequence σ for xFy. The N-length n of this sequence is either 1 or bigger than 1. In the first case, we find $\mathsf{Z}_\iota(x)$. Quod non. In the second case, we find $(\sigma)_{n-2} = \langle u, w \rangle$ and $u\mathsf{S}_\iota x$ and $w\mathsf{S}_K y$. By restricting σ to the first $(n-1)$ elements we obtain a witness for uFw. Since $xE_\iota x'$, we find that x' is a ι-successor. Thus, reasoning as before, we find u', w' with $u'\mathsf{S}_\iota x'$, $w'\mathsf{S}_K y'$ and $u'Fw'$. Since, $u\mathsf{S}x$, $u'\mathsf{S}_\iota x'$, $xE_\iota x'$, we find $uE_\iota u'$. Applying the induction hypothesis, we get $wE_K w'$, and, hence, $yE_K y'$.

We leave the proof that F commutes with successor, plus and times to the reader. ❑

We end this section with a theorem that will be useful in Section 9

Theorem 8.4 *Suppose $\iota : \mathsf{F} \to U$ satisfies full induction. Suppose $K : \mathsf{F} \to U$. Then, there is* at most *one $F : K \Rightarrow \iota$. Moreover, if such an F exists, it is an i-isomorphism.*

We omitt the easy proof.

9 On Comparing Arithmetic and Set Theory

In this section, we prove some results on arithmetics and set theories. We first prove that retractions in hINT preserve weakly i-initial arrows.

Theorem 9.1 *Let $\mho$ be any theory. Let $[\![\cdot]\!] := [\![\cdot]\!]_\mho$. Suppose that $L : V \to U$ is a retraction in hINT. Then $[\![L]\!]$ preserves weakly i-initial arrows in $[\![V]\!]$. (In the last statement, we consider L as a morphism in INT.)*

Proof

We reason in $\mathsf{INT}^{\mathsf{morph}}$. We assume the conditions of the theorem. Let $N : \mho \to V$ be weakly i-initial in $[\![V]\!]$. We want to show that $L \circ N$ is weakly i-initial in $[\![U]\!]$.

Let $K : U \to V$ be the split monomorphism in hINT corresponding to L. Let M be any arrow from $\mho$ to U. By the weak i-initiality of N, we have, for some G:

$$\mho \;\underset{N}{\overset{K \circ M}{\underset{\Uparrow G}{\rightarrow}}}\; V \xrightarrow{\;L\;} U$$

By the defining property of retractions, we have, for some i-isomorphism F:

$$\mho \xrightarrow{\;M\;} U \;\underset{L \circ K}{\overset{\mathsf{id}}{\underset{\Uparrow F}{\rightarrow}}}\; U$$

Ergo:

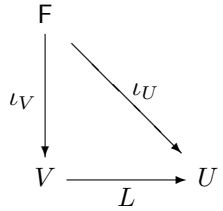
1. $L \circ G : L \circ N \Rightarrow L \circ K \circ M$,
2. $F \circ M : L \circ K \circ M \Rightarrow M$.

($F \circ M$ exists, since F is an i-isomorphism.) So we have:

$$H := (F \circ M) \cdot (L \circ G) : L \circ N \Rightarrow M.$$

Since M was arbitrary, it follows that $L \circ N$ is weakly i-initial. ❑

Theorem 9.2 *Suppose $\iota_U : \mathsf{F} \to U$ and $\iota_V : \mathsf{F} \to V$ are i-initial arrows. Suppose also that U is a retract of V in hINT. Let $L : V \to U$ be the retraction (split epimorphism). Then the following diagram commutes in hINT.*



It follows that $\iota_V^{-1}[V] \subseteq \iota_U^{-1}[U]$.

Proof

We assume the conditions of the theorem. Since ι_V is, a fortiori, weakly i-initial, we may conclude, by Theorem 9.1, that $L \circ \iota_V$ is weakly i-initial. So we have an arrow $H : L \circ \iota_V \Rightarrow \iota_U$. Since ι_U is i-initial, it satisfies full induction. Thus, we may conclude, by Theorem 8.4, that H is an i-isomorphism.

It follows, by Corollary 3.3, that $\iota_V^{-1}[V] \subseteq \iota_U^{-1}[U]$. ❑

Open Question 9.3 Prove or refute the analogue of Theorem 9.2 for eqINT. ❑

Corollary 9.4 *Suppose U and V are extensions of* PA *in the language of* PA. *Suppose U is a retract of V in* hINT. *Then $V \subseteq U$.*

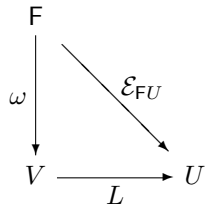
In Subsection 4.7, we showed that $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$ is a retract of PA. This illustrates the fact that, in Corollary 9.4, we cannot replace the subset relation by identity.

Example 9.5 It is well-known that there is a restricted interpretation from PA to PA. We have seen in Section 7, that it follows that there is an arithmetical sentence O such that both $\mathsf{PA} + O$ and $\mathsf{PA} + \neg O$ are retracts of PA. Clearly one of O, $\neg O$ must be true. Say it is O. Then, PA is interpretable in $\mathsf{PA} + O$, and, since $\mathsf{PA} + O$ is Σ^0_1-sound, PA is faithfully interpretable in $\mathsf{PA} + O$, by the results of Lindström, see e.g. [Lin94]. (Alternatively, see [Vis02].) It follows that PA and $\mathsf{PA} + O$ are mutually faithfully interpretable. By Theorem 6.13, it follows that PA and $\mathsf{PA} + O$ are mutually faithfully directly interpretable. On the other hand, by Corollary 9.4, they are not bi-interpretable. ❑

Corollary 9.6 *Suppose U is an extension of* PA *in the language of* PA *and V is an extension of* ZF *in the language of* ZF. *Then U is not a retract of V in* hINT.

Proof

Suppose U is an extension of PA in the language of PA and V is an extension of ZF in the language of ZF. Moreover, suppose that $K : U \to V$ and $L : V \to U$ witness that U is a retract of V. By Theorem 9.2, the following diagram commutes in hINT.



Here $\omega := \mathcal{E}_{\mathsf{ZF},V} \circ \mathsf{neumann}$, where $\mathsf{neumann}$ is the von Neumann interpretation of the natural numbers in ZF. Since $\mathsf{neumann}$ is restricted, we find, by Theorem 7.4, that $L \circ \omega$ is restricted. Hence, $\mathcal{E}_{\mathsf{F},U}$ is restricted. This gives a contradiction with Corollary 7.6, since $\mathcal{E}_{\mathsf{F},U}$ is surjective. ❑

Corollary 9.7 ZF^-*, i.e.* ZF *minus the axiom of infinity, is not isomorphic in* hINT *to any extension of* PA *in the language of* PA*.*

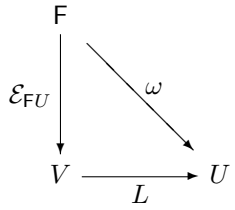
Proof

Let U be an extension of PA in the language of PA. Suppose that ZF^- is isomorphic in hINT to U. It follows, by the bisimulation property of isomorphisms, that ZF is isomorphic in hINT to some extension W of U in the language of PA. But this contradicts Corollary 9.6. ❑

Corollary 9.8 *Suppose* U *is an extension of* ZF *in the language of* ZF *and* V *is an extension of* PA *in the language of* PA*. Then* U *is not a retract of* V *in* hINT*.*

Proof

Suppose U is an extension of ZF in the language of ZF and V is an extension of PA in the language of PA. Moreover, suppose that $K : U \to V$ and $L : V \to U$ witness that U is a retract of V. By Theorem 9.2, the following diagram commutes in hINT.



Here $\omega := \mathcal{E}_{\mathsf{ZF},V} \circ \mathsf{neumann}$, where $\mathsf{neumann}$ is the von Neumann interpretation of the natural numbers in ZF. Note that $\mathcal{E}_{\mathsf{F}U}$ is surjective. Moreover, L is a split epimorphism in hINT. Since split epimorphisms are preserved by functors, L also yields a split epimorphism in eqINT. Hence, L is surjective. It follows that ω is surjective. But ω is also restricted, contradicting Corollary 7.6.

(Alternatively, we can reason as follows. Since $\mathcal{E}_{\mathsf{F}U}$ is surjective and $L \circ \mathcal{E}_{\mathsf{F}U} = \omega$ is restricted, it follows that L is restricted. Quid impossibile, since L is a retraction.) ❑

Corollary 9.9 PA^2 *is not a retract of* PA *in* hINT*.*

Proof

It is easily seen that PA^2 can be taken to be PA in the language expanded with a 0-ary predicate symbol P. We reason analogously to the first part of the proof of Corollary 9.8. This gives us that the standard embedding of F into PA^2 is surjective. But then P would be provably equivalent to an arithmetical sentence, quod non. ❑

10 Preservation over Retractions

We write $U \sqsubseteq_i V$ iff U is a retract of V in INT_i. Clearly $\sqsubseteq_i$ is a pre-order. The induced equivalence relation of $\sqsubseteq_i$ will be $\equiv_i$.

A property $\mathcal{P}$ of theories is *preserved over retractions* in INT_i, or *preserved to retracts in* INT_i iff whenever $\mathcal{P}(V)$ and $U \sqsubseteq_i V$, then $\mathcal{P}(U)$. Note that iff $\mathcal{P}$ is preserved under retractions in INT_i and $j \leq i$, then $\mathcal{P}$ is preserved under retractions in INT_j. We treat some examples of preservation.

Theorem 10.1 *κ-categoricity is preserved under retractions in* whINT.

Proof

Suppose $K : U \to V$ is a co-retraction and $M : V \to U$ is the corresponding retraction, both in whINT. Suppose V is κ-categorical and that $\mathcal{M} \models U$ and $|\mathcal{M}| = \kappa$. We have $M^{\mathcal{M}} \models V$ and $K^{M^{\mathcal{M}}} \models U$. Moreover, $K^{M^{\mathcal{M}}}$ is isomorphic to $\mathcal{M}$. Hence, $|K^{M^{\mathcal{M}}}| = \kappa$. Since, the cardinality of $M^{\mathcal{M}}$ must be between the cardinalities of $K^{M^{\mathcal{M}}}$ and $\mathcal{M}$, we find $|M^{\mathcal{M}}| = \kappa$. Consider any other model with $\mathcal{N} \models U$ and $|\mathcal{N}| = \kappa$. Again we find $|M^{\mathcal{N}}| = \kappa$. Since, $M^{\mathcal{M}}$ and $M^{\mathcal{N}}$ are models of V, we see that $M^{\mathcal{M}}$ and $M^{\mathcal{N}}$ are isomorphic. Hence, $K^{M^{\mathcal{M}}}$ and $K^{M^{\mathcal{N}}}$ are isomorphic and, so, $\mathcal{M}$ and $\mathcal{N}$ are isomorphic. We may conclude that U is κ-categorical. ❑

Theorem 10.2 *Finite axiomatizability is preserved under retractions in* hINT

Proof

Suppose $K : U \to V$ is a co-retraction and $M : V \to U$ is the corresponding retraction, both in hINT. Let $F : \mathsf{id} \Rightarrow M \circ K$ be an i-isomorphism.

Suppose that V is finitely axiomatized and that A is the conjunction of a finite set of axioms of V. We claim that the following axioms form an axiomatization of U.

- The statement witnessing that $F : \mathsf{id} \Rightarrow M \circ K$ is an i-isomorphism.
- A^M

Call the theory given by these axioms: $U^\star$. Clearly, $U^\star$ is a subtheory of U. Conversely, we have:

$$\begin{array}{rcl} U \vdash B & \Rightarrow & A \vdash B^K \\ & \Rightarrow & \exists x\ \delta_M(x), A^M \vdash B^{KM} \\ & \Rightarrow & U^\star \vdash B^{KM} \\ & \Rightarrow & U^\star \vdash B \end{array}$$

The last step is proved by verifying, by induction on formulas $C(\vec{x})$, that

$$U^\star \vdash \vec{x} F \vec{y} \rightarrow (C(\vec{x}) \leftrightarrow C^{KM}(\vec{y})).$$

❑

Open Question 10.3 It is a great open problem whether S_2, aka $I\Delta_0 + \Omega_1$, is finitely axiomatizable. We do know that S_2 is interpretable in Q. By Theorem 10.2, we know that, if S_2 were an hINT-retraction of Q, then S_2 would be finitely axiomatizable. Hence, our question: prove or refute that S_2 is an hINT-retraction of Q. ❑

Some properties of theories are associated with other theories. Consider any theory W. The theory U *has the property* $\mathcal{D}_W$ iff there is a direct interpretation from W to U.

Theorem 10.4 *$\mathcal{D}_U$ is extensionally the same as $\mathcal{D}_V$ iff U and V are mutually directly interpretable.*

Proof

Suppose $\mathcal{D}_U$ is extensionally the same as $\mathcal{D}_V$. Clearly, we have $\mathcal{D}_U(U)$. Hence, we have $\mathcal{D}_U(V)$ and, thus, there is a direct interpretation of U in V. Similarly, there is a direct interpretation of U in V. Hence, U and V are mutually directly interpretable.

Conversely, suppose U and V are mutually directly interpretable. Suppose further that $\mathcal{P}_U(W)$. Say, $L : U \rightarrow W$ is a direct interpretation witnessing this fact. Also, there is a direct interpretation $P : V \rightarrow U$. Hence, $L \circ P : V \rightarrow W$ is direct and, so, $\mathcal{P}(V)$. Similarly, $\mathcal{P}(U)$ follows from $\mathcal{P}(V)$, ❑

An important example of a property associated with a theory that is preserved over retractions in hINT is *sequentiality*.[21]

The notion of sequentiality is due to Pavel Pudlák. See, e.g., [HP91]. A theory is sequential iff it satisfies $\mathcal{D}_{\mathsf{SEQ}}$, where SEQ is the following theory.

[21] After this paper was finished, I realized that there is an alternative elegant treatment of sequentiality. It can be treated as a uniform e-scheme. In this way, sequentiality becomes primarily a property of interpretations. A theory U is sequential iff id_U is sequential. We will explore this line of thought in a later paper.

- We have predicates num, E, Z, S, A, M and axioms to the effect that Q^N, where N is the obvious translation of arithmetic to these predicates. The axioms include axioms to the effect that E is an equivalence relation on num, etc.
- We have predicates seq and proj and the following axioms:
  1. $\vdash \mathsf{seq}(s,u) \to (\mathsf{num}(u) \wedge \forall v{<^N}u\ \exists! x\ \mathsf{proj}(x,v,s))$,
  2. $\vdash \exists e,z\ (\mathsf{Z}(z) \wedge \mathsf{seq}(e,z))$,
  3. $\vdash \mathsf{seq}(s,u) \to \forall y\ \exists s',u'\ (\mathsf{S}(u,u') \wedge \mathsf{seq}(s',u') \wedge$
     $\forall v{<^N}u\ \forall x\ (\mathsf{proj}(x,v,s) \leftrightarrow \mathsf{proj}(x,v,s')) \wedge \mathsf{proj}(y,u,s'))$.

Sequentiality is a very robust notion. We can work with much stronger variants of SEQ. E.g., we could take, instead of Q, the theory $I\Delta_0{+}\Omega_1$. Stronger versions are more pleasant for applications. For the purpose of verifying that a theory is sequential, (seemingly) weaker definitions are better. Here is one such weaker version, due to Pudlák. (See [MPS90].) Let WSET be the following theory.

- The language of WSET has one binary relation symbol $\in$ (in addition to the identity symbol),
- We have the following axioms:
  1. $\exists y\, \forall x\ x \notin y$,
  2. $\forall x,y\, \exists z\, \forall u\ (u \in z \leftrightarrow (u \in y \vee u = x))$.

Theorem 10.5 (Pudlák) WSET *is mutually directly interpretable in* SEQ. *Hence, a theory is sequential iff it directly interprets* WSET.

We provide a slight variant of WSET, that is even more convenient, let's call it WSET'.

- The language of WSET' has one binary relation symbol $\in$ and one unary symbol set (in addition to the identity symbol),
- We have the following axioms:
  1. $\exists y : \mathsf{set}\ \forall x\ x \notin y$,
  2. $\forall x\ \forall y : \mathsf{set}\ \exists z : \mathsf{set}\, \forall u\ (u \in z \leftrightarrow (u \in y \vee u = x))$.

Theorem 10.6 WSET' *is mutually directly interpretable in* WSET. *Hence, a theory is sequential iff it directly interprets* WSET'.

Proof

The direct interpretation of WSET' in WSET sends $\mathsf{set}(x)$ to $x = x$ and leaves the rest unchanged. The direct interpretation of WSET in WSET' sends $x \in y$ to $x \in y \wedge \mathsf{set}(y)$ and leaves the rest unchanged. For the verification of the second WSET-axiom we distinguish the cases where $y : \mathsf{set}$ and where not $y : \mathsf{set}$. In the first case, we are immediately done. In the second case, y functions as an empty set. We are guaranteed an empty set $y^\star$ in set. We use $y^\star$ to find the desired z. ❑

We show that sequentiality is preserved over retractions in hINT.

Theorem 10.7 *Suppose V is sequential and U is a retract of V in hINT. Then, U is sequential.*

Proof

Suppose $K : U \to V$ and $M : V \to U$ and suppose that the isomorphism $F : \mathsf{id}_U \Rightarrow (M \circ K)$ witness the fact that U is a retract of V in hINT. Suppose V is sequential. Let $L : \mathsf{WSET}' \to V$ be a direct interpretation witnessing the sequentiality of V. We define a direct interpretation P of WSET' in U. We take:

- $\delta_P :\leftrightarrow v_0 = v_0$,
- $E_P :\leftrightarrow v_0 = v_1$,
- $\mathsf{set}_P :\leftrightarrow \delta_M(v_0) \wedge \mathsf{set}_L^M(v_0)$,
- $\in_P :\leftrightarrow \mathsf{set}_P(v_1) \wedge \exists x\ (v_0 F x \wedge x \in_L^M v_1)$.

We verify the axioms of WSET' under P. Reason in U.

First we verify the empty-set axiom. We have:

$$(\exists y : \mathsf{set}\ \forall x\ x \notin y)^{LM}. \tag{1}$$

So, we find:

$$\exists y' : \mathsf{set}_P\ \forall x' : \delta_M\ x' \notin_L^M y'. \tag{2}$$

Pick $y := y'$ and suppose $x \in_P y$. We have, for some x', xFx' and $x' \in_L^M y$. Also, from xFx', we get: $x' : \delta_M$. A contradiction with equation (2). So y witnesses the empty-set axiom for $\in_P$.

Next we verify the addition-of-an-element-axiom. We have:

$$(\forall x\ \forall y : \mathsf{set}\ \exists z : \mathsf{set}\ \forall u\ (u \in z \leftrightarrow (u \in y \vee u = x)))^{LM} \tag{3}$$

Hence,

$$\forall x' : \delta_M\ \forall y' : \mathsf{set}_P\ \exists z' : \mathsf{set}_P\ \forall u' : \delta_M\ (u' \in_L^M z' \leftrightarrow (u' \in_L^M y' \vee u' E_M x')). \tag{4}$$

Now consider any x and $y : \mathsf{set}_P$. Pick x' such that xFx', By equation (4), we can find $z' : \mathsf{set}_P$ such that (†) $\forall u' : \delta_M\ (u' \in_L^M z' \leftrightarrow (u' \in_L^M y \vee u' E_M x'))$. We take $z := z'$. Consider any u.

First suppose $u \in_P z$. It follows that, for some u', uFu' and $u' \in_L^M z$. From uFu', it follows that $u' : \delta_M$. Hence, by (†), $u' \in_L^M y$ or $u' E_M x'$.

From the first case, to wit $u' \in_L^M y$, we get $u \in_P y$. From the second case, $u' E_M x'$, we have $u' E_{M\circ K} x'$, since $u', x' : \delta_{M\circ K}$ and, on $\delta_{M\circ K}$, the relation $E_{M\circ K}$ is a coarser equivalence relation than E_M. Since xFx' and uFu', we may conclude that $u = x$.

Next suppose $u \in_P y$ or $u = x$. In the first case, there is an u', such that uFu' and $u' \in_L^M y$. We find $u' : \delta_M$ and, hence, by (†), $u' \in_L^M z$. Ergo $u \in_P z$. In the second case, we have, from (†), $x' \in_L^M z$, and, hence $u = x \in_P z$. ❑

References

[Ben86] C. Bennet. *On some orderings of extensions of arithmetic*. Department of Philosophy, University of Göteborg, 1986.

[Bor94] F. Borceux. *Handbook of Categorical Algebra 1, Basic Category Theory*. Encyclopedia of mathematics and its applications. Cambridge University Press, Cambridge, 1994.

[Bus86] S. Buss. *Bounded Arithmetic*. Bibliopolis, Napoli, 1986.

[Cor80] J. Corcoran. Notes and queries. *History and Philosophy of Logic*, 1:231–234, 1980.

[dB65a] K. L. de Bouvère. Logical synonymy. *Indagationes Mathematicae*, 27:622–629, 1965.

[dB65b] K. L. de Bouvère. Synonymous Theories. In J.W. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, pages 402–406. North Holland, Amsterdam, 1965.

[Háj70] P. Hájek. Logische Kategorien. *Archiv für Mathematische Logik und Grundlagenforschung*, 13:168–193, 1970.

[Hal99] Volker Halbach. Conservative theories of classical truth. *Studia Logica*, 62:353–370, 1999.

[Han65] W. Hanf. Model-theoretic methods in the study of elementary logic. In J.W. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, pages 132–145. North Holland, Amsterdam, 1965.

[Hod93] W. Hodges. *Model theory*. Encyclopedia of Mathematics and its Applications, vol. 42. Cambridge University Press, Cambridge, 1993.

[HP91] P. Hájek and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1991.

[Jac99] B. Jacobs. *Categorical Logic and Type Theory*. Number 141 in Studies in Logic and the Foundations of Mathematics. North Holland, Amsterdam, 1999.

[JdJ98] G. Japaridze and D. de Jongh. The logic of provability. In S. Buss, editor, *Handbook of proof theory*, pages 475–546. North-Holland Publishing Co., amsterdam edition, 1998.

[Kan72] S. Kanger. Equivalent theories. *Theoria*, 38:1–6, 1972.

[Kay91] Richard Kaye. *Models of Peano Arithmetic*. Oxford Logic Guides. Oxford University Press, 1991.

[Lin94] P. Lindström. *The Arithmetization of Metamathematics*, volume 15. Filosofiska meddelanden, blå serien, Institutionen för filosofi, Göteborgs universitet, Göteborg, 1994.

[Lin97] P. Lindström. *Aspects of Incompleteness*, volume Lecture Notes in Logic 10. Springer, Berlin, 1997.

[Mac71] S. MacLane. *Categories for the Working Mathematician*. Number 5 in Graduate Texts in Mathematics. Springer, New York, 1971.

[MPS90] J. Mycielski, P. Pudlák, and A.S. Stern. *A lattice of chapters of mathematics (interpretations between theorems)*, volume 426 of *Memoirs of the American Mathematical Society*. AMS, Providence, Rhode Island, 1990.

[PEK67] M.B. Pour-El and S. Kripke. Deduction-preserving "recursive isomorphisms" between theories. *Fundamenta Mathematicae*, 61:141–163, 1967.

[Per97] M. G. Peretyat'kin. *Finitely axiomatizable theories*. Consultants Bureau, New York, 1997.

[Pud85] P. Pudlák. Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, 50:423–441, 1985.

[Pud86] P. Pudlák. On the length of proofs of finitistic consistency statements in finitistic theories. In J.B. et al Paris, editor, *Logic Colloquium '84*, pages 165–196. North–Holland, 1986.

[Šve78] V. Švejdar. Degrees of interpretability. *Commentationes Mathematicae Universitatis Carolinae*, 19:789–813, 1978.

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

[Vis91] A. Visser. The formalization of interpretability. *Studia Logica*, 51:81–105, 1991.

[Vis98] A. Visser. An Overview of Interpretability Logic. In M. Kracht, M. de Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic, vol 1*, CSLI Lecture Notes, no. 87, pages 307–359. Center for the Study of Language and Information, Stanford, 1998.

[Vis02] A. Visser. Faith & Falsity: a study of faithful interpretations and false Σ^0_1-sentences. Logic Group Preprint Series 216, Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, October 2002.

A Questions

1. Provide separating examples concerning various notions, like monomorphism and isomorphism, across our categories.

2. Treat the notion of sum in INT_2 and INT_3.

3. Prove that equality of interpretations is complete for its prima facie complexity class, for INT_i, $i = 0, 1, 3, 4$. What is the complexity in the case of INT_2?

4. Which important properties of theories are preserved or antipreserved by morphisms? monomorphisms? isomorphisms? etc.

5. Which properties of theories and interpretations have natural formulations in terms of our categories (including DEG)? E.g., given two specific theories (e.g., PA and ZF) are there functorial automorphisms that interchange them?

6. Give an example of two theories that are bi-interpretable but not synonymous. (This is Question 4.16.)

7. Is there an interesting class of theories on which mutual interpretability and isomorphism in one of INT_i, for $i = 0, 1, 2, 3$, always coincide? What is the situation for the finitely axiomatizable sequential theories? (This is Question 4.17.)

8. There are many intimately connected pairs like PRA and $I\Sigma_1$, PA and ACA_0, PA and PA plus a non-standard satisfaction predicate, ZF and GB. For each pair we have strong conservativity results. On the other hand each the second theory of the pair proves the consistency of the first component on a definable cut. One consequence is a superexponential speed-up in the second theory. (See e.g. [Pud85], [Pud86].)

   Is there anything illuminating to say about the nature of the interpretation which embeds the first of the pair into the second?

9. Is there a direct interpretation of $T + \mathsf{incon}(T)$ into T, where

$$T \in \{\mathsf{Q}, \mathsf{S}^1_2, \mathsf{EA}, I\Sigma_1, \mathsf{ACA}_0, \mathsf{GB}\}?$$

   Here $\mathsf{incon}(T)$ is given some standard arithmetization and, in ACA_0 and GB we employ the usual interpretation of arithmetic. (This is Question 6.14.)

10. It is clear that the Ackermann translation makes sense in Elementary Arithmetic EA aka $I\Delta_0 + \mathsf{Exp}$. So what precise set theory corresponds with EA? (This is Question 6.20.)

11. Do we have a restricted $K : T \to T$, for all sequential T? (This is Question 7.7.)

12. Is there an example of an arrow $\iota : \mathsf{F} \to U$ that is *weakly initial*, but that does not satisfy full induction? (This is Question 8.2.)

13. Prove or refute the analogue of Theorem 9.2 for eqINT. (This is Question 9.3.)

14. It is a great open problem whether S_2, aka $I\Delta_0 + \Omega_1$, is finitely axiomatizable. We do know that S_2 is interpretable in Q. By Theorem 10.2, we know that, if S_2 were an hINT-retraction of Q, then S_2 would be finitely axiomatizable. Hence, our question: prove or refute that S_2 is an hINT-retraction of Q. (This is Question 10.3.)

B More General Notions

In this appendix we sketch some possible natural extensions of the notions of translation and interpretation as discussed in this paper.

B.1 Multidimensional Interpretations

We could employ δ, containing variables $v_0, \ldots, v_{n-1}$, where we use *several* variables to represent *one* object. Suppose $\mathsf{ar}(P) = m$. Then, P would τ-translate to a formula with variables among $v_0, \ldots, v_{nm-1}$. Each subsequent block of m variables would stand for one object. Such translations are called *multidimensional*. They give rise in the obvious way to multidimensional interpretations.

B.2 Many-sorted Predicate Logic

Translations and interpretatons generalize immediately to the many-sorted case. For many purposes, this is a very natural generalization. I suspect that the categories so obtained would have slightly better properties.

B.3 Interpretations with Parameters

Many famous interpretations like 'the Klein model' and Tarski's interpretation of true arithmetic into an extension of the theory of groups are interpretations with parameters. This extension is sufficiently important to at least state the basic definitions.[22]

Let Σ and Θ be signatures. A *relative translation with parameters* $\tau : \Sigma \to \Theta$ is given by a triple $\langle p, \delta, F \rangle$. The variables $v_0, \ldots, v_{p-1}$ will have the role of parameters. The formula δ is a Θ-formula, which contains at most $v_0, \ldots v_p$ free. The mapping F associates to each relation symbol R of Σ with arity n a Θ-formula $F(R)$ with variables among $v_0, \ldots, v_{n+p-1}$. Here the $v_p, \ldots v_{n+p-1}$ represent the argument places.

[22] I am not aware of a statement of this definition anywhere in the literature. One reason for this defect, may be that Tarski in [TMR53] opted for a different style of treatment, to wit: adding constants.

We translate Σ-formulas of to Θ-formulas of as follows:

- $(R(v_{i_0}, \cdots, v_{i_{n-1}}))^\tau := F(R)[v_p := v_{i_0+p}, \cdots, v_{p+n-1} := v_{i_{n-1}+p}]$;
  We have to 'shift' the v_{i_j} to v_{i_j+p} to avoid confusion with the parameters.
- $(\cdot)^\tau$ commutes with the propositional connectives;
- $(\forall v_j\, A)^\tau := \forall v_j\, (\delta[v_p := v_{j+p}] \to A^\tau)$;
- $(\exists v_j\, A)^\tau := \exists v_j\, (\delta[v_p := v_{j+p}] \wedge A^\tau)$.

We can compose relative translations with parameters as follows.

- $p_{\tau\nu} = p_\tau + p_\nu$,
- $\delta_{\tau\nu} := (\delta_\nu \wedge (\delta_\tau)^\nu)$,
- $R_{\tau\nu} = (R_\tau)^\nu$.

The identity translation $\mathsf{id} := \mathsf{id}_\Theta$ is defined in the obvious way, setting $p_{\mathsf{id}} := 0$.

Consider a translation with parameters $\tau : \Sigma \to \Theta$. Let $\mathcal{M} = \langle M, I\rangle$ be a model of signature Θ. Let f be an assignment for $\mathcal{M}$. Let

$$\Delta := \{m{\in}M \mid \mathcal{M}, f[v_p := m] \models \delta_\tau\}.$$

Suppose $\mathcal{M}, f$ satisfies $\exists v_p\ \delta$ and the τ-translations of the identity axioms in the Σ-language.

Suppose that $f(v_\ell) \in \Delta$, for all $\ell \geq p$. We can define a new model assignment pair $\langle \mathcal{N}, g\rangle := \tau^{\langle \mathcal{M}, f\rangle}$ as follows.

- We define, for m, m' in Δ, $m \simeq m' :\Leftrightarrow \mathcal{M}, f[v_p := m, v_{p+1} := m'] \models E_\tau$. Clearly $\simeq$ is an equivalence relation. We write $[m]$ for the $\simeq$-equivalence class of m. We take N, the domain of $\mathcal{N}$, to be $\Delta/\simeq$.
- $g(v_i) := [f(v_{i+p})]$.
- Suppose $\mathsf{ar}(P) = k$. We find that $\simeq$ is a congruence (on Δ) w.r.t. the relation R_0, given by:

  $$R_0(\vec{m}) :\Leftrightarrow \mathcal{M}, f[v_p := m_0, \ldots, v_{k+p-1} := m_{k-1}] \models P^\tau.$$

  Thus, it makes sense to define, for $\vec{n}$ a sequence of elements of N,

  $$P^{\mathcal{N}}(\vec{n}) :\Leftrightarrow \exists m_0{\in}n_0 \ldots \exists m_{k-1}{\in}n_{k-1}\ R_0(m_0, \cdots, m_{k-1}).$$

One can show, for $\langle \mathcal{N}, g\rangle = \langle \mathcal{M}, f\rangle^\tau$, that:

$$\mathcal{N}, g \models A \Leftrightarrow \mathcal{M}, f \models A^\tau.$$

Further, one can show that $\tau^{\nu^{\mathcal{M},f}}$ exists iff $(\tau\nu)^{\mathcal{M},f}$ exists and that if both exist, they are isomorphic.

A *relative interpretation with parameters* is a quadruple $\langle U, \tau, C, V\rangle$, where U is a theory of signature Σ, V is a theory of signature Θ, where $\tau : \Sigma \to \Theta$ is a translation with parameters and where C is a Θ-formula containing at most $v_0, \ldots, v_{p_\tau - 1}$ free. The formula C provides a constraint on the parameters.[23] We demand:

- $V \vdash \exists v_0 \cdots \exists v_{p_\tau - 1}\, C$,
- for any Σ-sentence A, if $U \vdash A$, then $V \vdash \forall v_0 \cdots \forall v_{p_\tau - 1}\, (C \to A^\tau)$.

The identity interpretation is defined in the obvious way. The only new aspect of composition of interpretations is the transformation of the constraining formula C. This works as follows:

- $C_{M \circ K} := C_M \wedge (v_{p_M}, \ldots, v_{p_M + p_K - 1} : \delta_M) \wedge C_K^M$.

There are several choices for the MOD functor. We can make $\mathsf{MOD}(T)$ the set of models of T and simply say that, for $K : U \to V$, $\mathsf{MOD}(K)$ is a total binary relation from models of V to models of U. We can also take $\mathsf{MOD}(T)$ to be a set of model/assignment pairs $\langle \mathcal{M}, f\rangle$ such that $\mathcal{M} \models T$. In this case, $\mathsf{MOD}(K)$ is a *partial* map from $\mathsf{MOD}(V)$ to $\mathsf{MOD}(U)$, to wit,

$$\mathsf{MOD}(K)(\langle \mathcal{M}, f\rangle) = \langle \mathcal{M}, f\rangle^{\tau_K}.$$

We still have some freedom in stipulating to which category the MOD-functor is a functor, since we can put further conditions on partial maps Φ from $\mathsf{MOD}(V)$ to $\mathsf{MOD}(U)$. E.g., we can demand that for every $\mathcal{M} \models V$, there be an f, such that $\Phi(\langle \mathcal{M}, f\rangle)$ is defined.

[23] In the paper [MPS90], no constraining formula is used. The reason that this can be avoided is the fact that the authors study *local* interpretability.