WHEN BI-INTERPRETABILITY IMPLIES SYNONYMY

HARVEY FRIEDMAN
Department of Mathematics, The Ohio State University
and
ALBERT VISSER
Department of Philosophy and Religious Studies, Faculty of Humanities,
Utrecht University

Abstract. Two salient notions of sameness of theories are *synonymy*, aka *definitional equivalence*, and *bi-interpretability*. Of these two *definitional equivalence* is the strictest notion. In which cases can we infer synonymy from bi-interpretability? We study this question for the case of sequential theories. Our result is as follows. Suppose that two sequential theories are bi-interpretable and that the interpretations involved in the bi-interpretation are one-dimensional and identity preserving. Then, the theories are synonymous.

The crucial ingredient of our proof is a version of the Schröder–Bernstein theorem under very weak conditions. We think this last result has some independent interest.

We provide an example to show that this result is optimal. There are two finitely axiomatized sequential theories that are bi-interpretable but not synonymous, where precisely one of the interpretations involved in the bi-interpretation is not identity preserving.

§1. Introduction. When are two theories the same? Are there reasonable ways of abstracting away from the precise choice of the signature? The notions of *synonymy* (or: *definitional equivalence*) and *bi-interpretability* provide two good answers to these questions.

The notion of synonymy was introduced by de Bouvère in 1965 (see [2, 3]). It appears to be the strictest notion of sameness of theories except strict identity of signature and set of theorems. Two theories U and V are synonymous iff there is a theory W that is both a definitional extension of U and of V (where we allow the signatures of U and V to be made disjoint). Equivalently, U and V are synonymous iff there are interpretations $K : U \to V$ and $M : V \to U$, such that V proves that the composition $K \circ M$ is the identity interpretation on V and such that U proves that the composition $M \circ K$ of is the identity interpretation on U. (Thus, synonymy is isomorphism in an appropriate category INT_0 of theories and interpretations.) There are many familiar examples of synonymy. First, a theory is synonymous with a definitional extension. For example, Peano Arithmetic with 0, S, $+$, and $\times$ is synonymous with its extension with $<$ and the additional axiom $x < y \leftrightarrow \exists z\ x + \mathsf{S}z = y$. Then, we have examples that are

Received: January 23, 2014.
2020 *Mathematics Subject Classification*: Primary 03A05, 03B30, 03F25.
Key words and phrases: interpretations, interpretability, Schröder–Bernstein theorem.

doi:10.1017/S1755020325100920

considered by mathematicians to be trivial alternative formulations like the theory of weak partial order and the theory of strong partial order. Finally, there are more substantial examples. For example, Peano Arithmetic, PA, is synonymous with an appropriate theory of strings and concatenation. See [5] and, for a slightly weaker classical result, [25].[1]

The notion of *bi-interpretability* was introduced by Alhbrandt and Ziegler in 1986 (see [1] and, also, [9]). Two theories U and V are bi-interpretable iff there are interpretations $K : U \to V$ and $M : V \to U$, such that there is a V-definable function F, such that V proves that F is an isomorphism between $K \circ M$ and the identity interpretation on V and such that there is a U-definable function G, such that U proves that G is an isomorphism between $M \circ K$ and the identity interpretation on U. (Thus, bi-interpretability is isomorphism in an appropriate category INT_1 of theories and interpretations.)

In terms of models, the notion of bi-interpretability takes the following form. We note that an interpretation $K : U \to V$ gives us a uniform construction of an internal model $\widetilde{K}(\mathcal{M})$ of U in a model $\mathcal{M}$ of V. We find that U and V are bi-interpretable iff, there are interpretations $K : U \to V$ and $M : V \to U$ and formulas F and G, such that, for all models $\mathcal{M}$ of V, the formula F defines in $\mathcal{M}$ an isomorphism between $\mathcal{M}$ and $\widetilde{M}\widetilde{K}(\mathcal{M})$, and, for all models $\mathcal{N}$ of U, the formula G defines in $\mathcal{N}$ an isomorphism between $\mathcal{N}$ and $\widetilde{K}\widetilde{M}(\mathcal{N})$.

Bi-interpretability has a lot of good properties. For example, it preserves automorphism groups, κ-categoricity, finite axiomatizability, etc. Still the stricter notion of synonymy preserves more. For example, synonymy preserves the action of the automorphism group on the domain of the model. Bi-interpretability (without parameters) does preserve the automorphism group modulo isomorphism but does not necessarily preserve the action on the domain. See §7 for an example illustrating this difference. An example of a property of theories that is not preserved under bi-interpretability (not even under definitional equivalence) is having a computable model (see [19]).

Surprisingly, it is not easy to provide natural examples of pairs of theories that are bi-interpretable but not synonymous. For example, Peano Arithmetic PA is *prima facie* bi-interpretable with an appropriate theory of the hereditarily finite sets. However, on closer inspection, these theories are also synonymous (see §6.3). In §7, we give a verified example of two finitely axiomatized sequential theories that are bi-interpretable but not synonymous.

Our interest in this paper is in the relationship between synonymy and bi-interpretability for a special class of theories, *the sequential theories*. These are theories that have coding of sequences. Examples of sequential theories are Buss's theory S^1_2, Elementary Arithmetic EA or EFA, $\mathrm{I}\Sigma_1$, ZF, ZFC. We explain in detail what sequential theories are in §3.

We will show that, for *identity-preserving* interpretations between sequential theories, synonymy and bi-interpretability coincide (§5). In fact the proof works for a somewhat wider class: *the conceptual theories*. This last class was just defined here to pin down (more) precisely what is needed for the theorem.

[1] In fact, it is simpler to prove that these theories are bi-interpretable via one-dimensional, identity preserving interpretations. Then, the results of the present paper immediately yield the synonymy.

A central ingredient of our proof is the Schröder–Bernstein Theorem that turns out to hold under surprisingly weak conditions. We give the proof of the Schröder–Bernstein Theorem under these weak conditions in §4.

1.1. Historical remark. The first preprint of this paper came out in the Logic Group Preprint Series of Utrecht University in 2014. On August 4, 2023, Albert Visser received an e-mail from Leszek Kołodziejczyk reporting that his team found a counter-example to Theorem 5.4 of the preprint and, on September 6, 2023, Visser received an e-mail from Tim Button that he found a mistake in the proof of Theorem 5.4 with suggestions on how to fix it. Fortunately, none of the further results of the paper rested on the mistaken Theorem, but only on the correct Corollary 5.5. So, we eliminated the mistaken theorem. Corollary 5.5 of the earlier version is Theorem 5.3 of the present version.

§2. Basic notions.

In this section, we formulate the basic notions employed in the paper. We keep the definitions here at an informal level. More detailed definitions are given in Appendix A.

2.1. Theories. The primary focus in this paper is on one-sorted theories of first-order predicate logic of relational signature. We take identity to be a logical constant. Our official signatures are relational, however, via the term-unwinding algorithm, we can also accommodate signatures with functions. Many-sorted theories appear as an auxiliary in the study of one-sorted theories. We will only consider theories with a finite number of sorts.

The results of the paper that are stated for one-sorted theories can be lifted to the many-sorted case in a fairly obvious way. We choose to restrict ourselves to the one-sorted case to keep the presentation reasonably light.

In this paper, no restriction is needed on the complexity of the set of axioms of a theory or on the size of the signature.

2.2. Interpretations. We describe the notion of an m-dimensional interpretation for a one-sorted language. An interpretation $K : U \to V$ is given by the theories U and V and a translation τ from the language of U to the language of V. The translation is given by a domain formula $\delta(\vec{x}\,)$, where $\vec{x}$ is a sequence of m variables, and a mapping from the predicates of U to formulas of V, where an n-ary predicate P is mapped to a formula $A(\vec{x}_0, \dots, \vec{x}_{n-1})$, where the $\vec{x}_j$ are appropriately chosen pairwise disjoint sequences of m variables. We lift the translation to the full language in the obvious way, making it commute with the propositional connectives and quantifiers, where we relativize the translated quantifiers to the domain δ. We demand that V proves all the translations of theorems of U.

We can compose interpretations in the obvious way. Note that the composition of an n-dimensional interpretation with an m-dimensional interpretation is $m \times n$-dimensional.

A one-dimensional interpretation is *identity preserving* if translates identity to identity. A one-dimensional interpretation is *unrelativized* if its domain consists of all the objects of the interpreting theory. A one-dimensional interpretation is *direct* if it is unrelativized and preserves identity. Note that all these properties are preserved by composition.

Each interpretation $K : U \to V$ gives us an inner model construction that builds a model $\widetilde{K}(\mathcal{M})$ of U out of a model $\mathcal{M}$ of V. Note that $\widetilde{(\cdot)}$ behaves contravariantly.

If we want to use interpretations to analyze sameness of theories, we will need, as we will see, to be able to say when two interpretations are 'equal'. Strict identity of interpretations is too fine grained. It depends too much on arbitrary choices like the selection of bound variables. We specify a first notion of equality between interpretations: two interpretations are *equal* when the *target theory thinks they are*. Modulo this identification, the operations identity and composition give rise to a category INT_0, where the theories are objects and the interpretations arrows.[2]

Let MOD be the category with as objects classes of models and as morphisms all functions between these classes. We define $\mathsf{Mod}(U)$ as the class of all models of U. Suppose $K : U \to V$. Then, $\mathsf{Mod}(K)$ is the function from $\mathsf{Mod}(V)$ to $\mathsf{Mod}(U)$ given by: $\mathcal{M} \mapsto \widetilde{K}(\mathcal{M})$. It is clear that Mod is a *contravariant functor* from INT_0 to MOD.

2.3. Sameness of interpretations. For each sufficiently good notion of sameness of interpretations, there is an associated category of theories and interpretations: the category of interpretations modulo that notion of sameness. Any such a category gives us a notion of isomorphism of theories which can function as a notion of sameness.

We present a *basic list* of salient notions of sameness. For all items in the list, it is easily seen that sameness is preserved by composition. Our list does not have any pretence of being complete. For example, we omitted notions of sameness based on Ehrenfeucht games.

2.3.1. Equality. The interpretations $K, K' : U \to V$ are equal when V 'thinks' K and K' are identical. By the Completeness Theorem, this is equivalent to saying that, for all V-models $\mathcal{M}$, $\widetilde{K}(\mathcal{M}) = \widetilde{K'}(\mathcal{M})$. This notion gives rise to the category INT_0. Isomorphism in INT_0 is *synonymy* or *definitional equivalence*.

2.3.2. i-Isomorphism. An i-isomorphism between interpretations $K, M : U \to V$ is given by a V-formula F. We demand that V verifies that "F is an isomorphism between K and M", or, equivalently, that, for each model $\mathcal{M}$ of V, the function $F^{\mathcal{M}}$ is an isomorphism between $\widetilde{K}(\mathcal{M})$ and $\widetilde{M}(\mathcal{M})$.

Two interpretations $K, K' : U \to V$, are *i-isomorphic* iff there is an i-isomorphism between K and K'. Wilfrid Hodges calls this notion: *homotopy* (see [9, p. 222]).

We can also define the notion of being i-isomorphic semantically. The interpretations $K, K' : U \to V$, are *i-isomorphic* iff there is V-formula F such that for all V-models $\mathcal{M}$, the relation $F^{\mathcal{M}}$ is an isomorphism between $\widetilde{K}(\mathcal{M})$ and $\widetilde{K'}(\mathcal{M})$.

In case the signature of U is finite, being i-isomorphic has a third characterization. The interpretations $K, K' : U \to V$, are *i-isomorphic* iff, for every V-model $\mathcal{M}$, there is an $\mathcal{M}$-definable isomorphism between $\widetilde{K}(\mathcal{M})$ and $\widetilde{K'}(\mathcal{M})$ (see Theorem A.1).

Clearly, if K, K' are equal in the sense of the previous section, they will be i-isomorphic. The notion of i-isomorphism gives rise to a category of interpretations modulo i-isomorphism. We call this category INT_1. Isomorphism in INT_1 is *bi-interpretability*.

[2] For many reasons, the choice for the reverse direction of the arrows would be more natural. However, our present choice coheres with the extensive tradition in degrees of interpretability. So, we opted to adhere to the usual choice.

2.3.3. Isomorphism. Our third notion of sameness of the basic list is that K and K' are the same if, for all models $\mathcal{M}$ of V, the internal models $\widetilde{K}(\mathcal{M})$ and $\widetilde{K'}(\mathcal{M})$ are isomorphic. We will simply say that K and K' are isomorphic. Clearly, i-isomorphism implies isomorphism. We call the associated category INT_2. Isomorphism in INT_2 is *iso-congruence*.

2.3.4. Elementary equivalence. The fourth notion is to say that two interpretations K and K' are the same if, for each $\mathcal{M}$, the internal models $\widetilde{K}(\mathcal{M})$ and $\widetilde{K'}(\mathcal{M})$ are elementarily equivalent. We will say that K and K' are elementarily equivalent. By the Completeness Theorem, this notion can be alternatively defined by saying that K is the same as K' iff, for all U-sentences A, we have $V \vdash A^K \leftrightarrow A^{K'}$.

We call the associated category INT_3. Isomorphism in INT_3 is *elementary congruence* or *sentential congruence*.

2.3.5. Identity of source and target. Finally, we have the option of abstracting away from the specific identity of interpretations completely, simply counting any two interpretations $K, K' : U \to V$ the same. The associated category is INT_4. This is simply the structure of degrees of one-dimensional interpretability. Isomorphism in INT_4 is *mutual interpretability*.

2.4. The many-sorted case. Interpretability can be extended to interpretability between many-sorted theories. However to do that properly, we would need to develop the notion of piecewise interpretation. Since this notion is not needed in the present paper, we just describe interpretations of many-sorted theories in one-sorted theories. These are precisely what one would expect: the interpretation K does not specify just one domain, but, for each sort $\mathfrak{a}$, a domain $\delta_{\mathfrak{a}}$. We allow a different dimension for each sort. The translation of a quantifier $\forall x^{\mathfrak{a}}$ is $\forall \vec{x}\ (\delta_{\mathfrak{a}}(\vec{x}\,) \to \cdots)$. We translate a predicate P of type $\mathfrak{a}_0, \ldots, \mathfrak{a}_{n-1}$ to a formula $A(\vec{x}_0, \ldots \vec{x}_{n-1})$, where the target theory verifies, for $i < n$, the formula $A(\vec{x}_0, \ldots \vec{x}_{n-1}) \to \delta_{\mathfrak{a}_i}(\vec{x}_i)$.[3]

We will consider theories with a designated sort $\mathfrak{o}$ of objects. An interpretation of such a theory into a one-sorted theory is $\mathfrak{o}$-*direct* iff it is one-dimensional for sort $\mathfrak{o}$, and has $\delta_{\mathfrak{o}}(x) := (x = x)$ and translates identity on $\mathfrak{o}$ to identity simpliciter. In other words, the interpretation is direct when we restrict our attention to the single sort $\mathfrak{o}$.

2.5. Parameters. We can extend our notion of interpretation to *interpretation with parameters* as follows. Say our interpretation is $K : U \to V$. In the target theory, we have a parameter domain $\alpha(\vec{z}\,)$, which is V-provably non-empty. The definition of interpretation remains the same but for the fact that the parameters $\vec{z}$. Our condition for K to be an interpretation becomes:

$$U \vdash A \;\;\Rightarrow\;\; V \vdash \forall \vec{z}\ (\alpha(\vec{z}\,) \to A^{K,\vec{z}}\,).$$

We note that an interpretation $K : U \to V$ with parameters provides a parametrized *set* of inner models of U inside a model of V.

§3. Sequentiality and conceptuality. We are interested in theories *with coding*. There are several 'degrees' of coding, like pairing, sequences, etc. We want a notion that allows

[3] Note that the sequence $\vec{x}_i$ has as length the dimension associated with the sort $\mathfrak{a}_i$.

us to build arbitrary sequences of all objects of our domain. The relevant notion is *sequentiality*. We also define a wider notion *conceptuality*. This last notion is proof-generated: it gives us the most natural class of theories for which our proof works. All sequential theories are conceptual, but not vice versa.

We have a simple and elegant definition of sequentiality. A theory U is *sequential* iff it directly interprets *adjunctive set theory* AS. Here AS is the following theory in the language with only one binary relation symbol.

AS1. $\vdash \exists x\, \forall y\;\; y \notin x$,
AS2. $\vdash \forall x, y\, \exists z\, \forall u\, (u \in z \leftrightarrow (u \in x \vee u = y))$.

So the basic idea is that we can define a predicate $\in^\star$ in U such that $\in^\star$ satisfies a very weak set theory involving *all* the objects of U. Given this weak set theory, we can develop a theory of sequences for all the objects in U, which again gives us partial truth-predicates, etc. In short, the notion of sequentiality explicates the idea of a *theory with coding*.

Remark 3.1. *To develop the notion of sequentiality in a proper way for many-sorted theories, we would need the idea of a* piecewise *interpretation. We do not develop the idea of piecewise interpretation here. Fortunately, one can forget the framework and give the definition in a theory-free way. It looks like this. Let U be a theory with sorts $\mathcal{S}$. The theory U is sequential when we can define, for each $\mathfrak{a}, \mathfrak{b} \in \mathcal{S}$, a binary predicate $\in^{\mathfrak{ab}}$ of type $\mathfrak{ab}$ such that*:

a. $U \vdash \bigvee_{\mathfrak{a} \in \mathcal{S}} \exists x^{\mathfrak{a}} \bigwedge_{\mathfrak{b} \in \mathcal{S}} \forall y^{\mathfrak{b}}\, y^{\mathfrak{b}} \notin^{\mathfrak{ba}} x^{\mathfrak{a}}$,
b. $U \vdash \bigwedge_{\mathfrak{a},\mathfrak{b} \in \mathcal{S}} \forall x^{\mathfrak{a}}, y^{\mathfrak{b}} \bigvee_{\mathfrak{c} \in \mathcal{S}} \exists z^{\mathfrak{c}} \bigwedge_{\mathfrak{d} \in \mathcal{S}} \forall u^{\mathfrak{d}}\, (u^{\mathfrak{d}} \in^{\mathfrak{dc}} z^{\mathfrak{c}} \leftrightarrow (u^{\mathfrak{d}} \in^{\mathfrak{da}} x^{\mathfrak{a}} \vee u^{\mathfrak{d}} =^{\mathfrak{db}} y^{\mathfrak{b}}))$.

Here '$=^{\mathfrak{db}}$' is not really in the language if $\mathfrak{d} \neq \mathfrak{b}$. In this case, we read $u^{\mathfrak{d}} =^{\mathfrak{db}} y^{\mathfrak{b}}$ simply as $\bot$.

It's a nice exercise to show that, e.g., ACA_0 *and* GB *are sequential.*

Closely related to AS is *adjunctive class theory* ac. We define this theory as follows. The theory ac is two-sorted with sorts $\mathfrak{o}$ (of objects) and $\mathfrak{c}$ (of classes). We have an identity for every sort and one relation symbol $\in$ between objects and classes, i.e., of type $\mathfrak{oc}$. We let $x, y, \ldots$ range over objects and $X, Y, \ldots$ range over classes. We have the following axioms:

ac1. $\vdash \exists X\, \forall x\;\; x \notin X$,
ac2. $\vdash \forall Y, y\, \exists X\, \forall x\, (x \in X \leftrightarrow (x \in Y \vee x = y))$,
ac3. $\vdash X = Y \leftrightarrow \forall z\, (z \in X \leftrightarrow z \in Y)$.

Note that extensionality is cheap since we could treat identity on classes as *defined* by the relation of extensional sameness. The theory ac is much weaker than AS, since it admits finite models. The following theorem is easy to see.

Theorem 3.2. *A theory U is sequential iff there is an $\mathfrak{o}$-direct interpretation of* ac *in U that is one-dimensional in the interpretation of classes.*

A theory U is *conceptual* iff there is an $\mathfrak{o}$-direct interpretation of ac in U. We note that there are conceptual theories that are not sequential. For example, sequential theories always have an infinite domain, but there are conceptual theories with finite models.

Note also that AS is sequential, but ac is *not* conceptual (not even in the appropriate many-sorted formulation).

For more information on sequentiality and conceptuality, see Appendix B.

§4. The Schröder–Bernstein theorem. We start with a brief story of the genesis of our version of the Schröder–Bernstein theorem. The first step was taken by Harvey Friedman who saw that sequential theories should satisfy a version of the Schröder–Bernstein Theorem,[4] which would lead to the desired result on the coincidence of synonymy and bi-interpretability.[5] Albert Visser subsequently wrote down a proof, discovering that one needs even less than sequentiality: the thing to use is adjunctive class theory. Allan van Hulst verified Visser's version of the proof in Mizar as part of his master's project under Freek Wiedijk in 2009. After hearing a presentation by Allan van Hulst, Tonny Hurkens found a simplification of the proof. Hurkens' proof is shorter and conceptually simpler. In our presentation here, we include Hurkens' simplification. We thank Tonny for his gracious permission to do so.

Our proof differs from the usual proofs to keep the demand on resources low. The reader may find it instructive to compare it with a more standard proof of the Schröder–Bernstein theorem (see, e.g., [14, pp. 85–86]).

We work in the theory SB which is ac extended with two unary predicates on objects: A and B and four binary predicates on objects: $\mathsf{E_A}$, $\mathsf{E_B}$, F, and G, plus axioms expressing that $\mathsf{E_A}$ is an equivalence relation on A, $\mathsf{E_B}$ is an equivalence relation on B, F is an injection from $\mathsf{A}/\mathsf{E_A}$ to $\mathsf{B}/\mathsf{E_B}$, and G is an injection from $\mathsf{B}/\mathsf{E_B}$ to $\mathsf{A}/\mathsf{E_A}$. We construct a formula H that SB-provably defines a bijection between $\mathsf{A}/\mathsf{E_A}$ and $\mathsf{B}/\mathsf{E_B}$.

We will employ the usual notations like: $\emptyset$, $\{x_0, \ldots, x_{n-1}\}$, $\subseteq$.

Our definition of what it means that F is a function includes: if $x\mathsf{E_A}x'\mathsf{F}y'\mathsf{E_B}y$, then $x\mathsf{F}y$. Similarly for G. We will treat A as a virtual class and write $x \in \mathsf{A}$, etc.

- A pair of classes (X, Y) is *downwards closed* if,
  1. $X \subseteq \mathsf{A}$ and $Y \subseteq \mathsf{B}$,
  2. if $v\mathsf{G}u$ and $u \in X$, then there is a $v' \in Y$ such that $v'\mathsf{G}u$,
  3. if $u\mathsf{F}v$ and $v \in Y$, then there is a $u' \in X$ such that $u'\mathsf{F}v$.
- We say that (X, Y) is an *x-switch* if
  i. (X, Y) is downwards closed,
  ii. x is a member of X,
  iii. each member of X is in the range of G.
- $x\mathsf{H}y$ iff (there is no x-switch and $x\mathsf{F}y$) or (there is an x-switch and $y\mathsf{G}x$).

LEMMA 4.1 (SB). H *is a function from* $\mathsf{A}/\mathsf{E_A}$ *to* $\mathsf{B}/\mathsf{E_B}$.

Proof. We prove that H preserves equivalences. Suppose $x\mathsf{E_A}x'$, $y\mathsf{E_B}y'$ and $x\mathsf{H}y$. Suppose there is an x-switch (X, Y). It is easy to see that $(X \cup \{x'\}, Y)$ is an x'-switch. It is now immediate that $x'\mathsf{H}y'$. Similarly, if we are given an x'-switch, we may conclude

4 In this paper, we use the version of the Schröder–Bernstein Theorem that is formulated in terms of injections. We note that the theorem really should be called: the Dedekind–Cantor–Schröder–Bernstein Theorem.

5 Harvey Friedman announced the result in an e-mail of January 1, 2009 starting with the words: *I am guessing that you will not believe in this.*

that there is an x-switch. The remaining case where there is neither an x-switch nor an x'-switch is again immediate.

We prove that H is functional. Suppose $x\mathsf{H}y$ and $x\mathsf{H}y'$. If there is an x-switch, we have $y\mathsf{G}x$ and $y'\mathsf{G}x$. So, we are done by the injectivity of G. If there is no x-switch, we have $x\mathsf{F}y$ and $x\mathsf{F}y'$. So, we are done by the functionality of F.

We prove that H is total on A. Consider any x in A. If there is no x-switch, we are done, since x is in the domain of F. If there is an x-switch, then x is in the range of G, and, again, we are done. □

LEMMA 4.2 (SB). H *is injective from* $\mathsf{A}/\mathsf{E}_\mathsf{A}$ *to* $\mathsf{B}/\mathsf{E}_\mathsf{B}$.

Proof. Suppose $x\mathsf{H}y$ and $x'\mathsf{H}y$.

If in both cases the same clauses in the definition of H are active, we are easily done.

Suppose there is no x-switch and there is an x'-switch, say (X', Y'). By the definition of H, we have $x\mathsf{F}y\mathsf{G}x'$. By the downwards closure of (X', Y'), the injectivity of G (viewed as a relation between $\mathsf{B}/\mathsf{E}_\mathsf{B}$ and $\mathsf{A}/\mathsf{E}_\mathsf{A}$), and the functionality of F (viewed as a relation between $\mathsf{A}/\mathsf{E}_\mathsf{A}$ and $\mathsf{B}/\mathsf{E}_\mathsf{B}$), we may conclude that $x\mathsf{E}_\mathsf{A}x''$, for some x'' in X'. Hence, $(X' \cup \{x\}, Y')$ is an x-switch. A contradiction. □

LEMMA 4.3 (SB). H *is surjective.*

Proof. Consider any $y \in \mathsf{B}$. First, suppose y is not in the range of F. Let $y\mathsf{G}x$. Then, $(\{x\}, \{y\})$ is an x-switch, and we have $x\mathsf{H}y$.

Next, suppose $x\mathsf{F}y$ and there is no x-switch. In this case, $x\mathsf{H}y$.

Finally, suppose $x\mathsf{F}y$ and there is an x-switch (X, Y). Let $y\mathsf{G}x'$. Then, we find that $(X \cup \{x'\}, Y \cup \{y\})$ is an x'-switch. Hence, $x'\mathsf{H}y$.

Thus, in all cases, y is in the image of H. □

We have proved the following theorem.

THEOREM 4.1. *In* SB, *we can construct a bijection* H *between* $\mathsf{A}/\mathsf{E}_\mathsf{A}$ *and* $\mathsf{B}/\mathsf{E}_\mathsf{B}$.

EXAMPLE 4.2. *Consider a model of* SB. *We write A for the interpretation of* A, *etc. We note that the function H constructed by the proof of the theorem depends on our choice of classes. Suppose, e.g., that our objects are the integers, A is the set of even integers, B is the set of odd integers, our equivalence relations are identity on the given virtual class, F is the successor function domain-restricted to A, and G is the successor function domain-restricted to B. In the case that our classes are all possible classes of numbers, the pair of the class of all even numbers and all odd numbers is an* a*-switch, for each even* a. *So $H = G^{-1}$, i.e., the predecessor function domain-restricted to A. In the case that our classes are the finite classes, there is no* a*-switch for any even* a. *So, $H = F$.*

For us, the following obvious corollary is relevant.

COROLLARY 4.1. *Let T be a conceptual theory. Suppose we have formulas Ax, By, E_A, E_B, F, G, where T proves that E_A is an equivalence relation on $\{x \mid Ax\}$, that E_B is an equivalence relation on $\{y \mid By\}$, that F is an injection from $\{x \mid Ax\}/E_A$ to $\{y \mid By\}/E_B$, and that G is an injection from $\{y \mid By\}/E_B$ to $\{x \mid Ax\}/E_A$. Then we can find a formula H that T-provably defines a bijection between the virtual classes $\{x \mid Ax\}/E_A$ and $\{y \mid By\}/E_B$.*

Our corollary can be rephrased as follows. Suppose that T is conceptual and we have one-dimensional interpretations $K, M : \mathsf{EQ} \to T$, where EQ *is the theory of equality.*

Suppose further that $F : K \to M$ and $G : M \to K$ are injections. Then we can find a bijection $H : K \to M$, and, thus, K and M are i-isomorphic.

We note that if T is sequential, we can drop the demand that K and M are one-dimensional, since every interpretation in a sequential theory is i-isomorphic with a one-dimensional interpretation.

§5. From bi-interpretability to synonymy. In this section, we prove our main result. If two theories are bi-interpretable via identity-preserving interpretations, then they are synonymous.

THEOREM 5.1. *Let U and V be any theories with $K : U \to V$ and $M : V \to U$. Suppose that, for any model $\mathcal{M}$ of V, we have $\widetilde{M}\widetilde{K}(\mathcal{M}) = \mathcal{M}$ (in other words, $K \circ M = \mathsf{id}_V$ in INT_0). Suppose further that, for any model $\mathcal{N}$ of U, the model $\widetilde{K}\widetilde{M}(\mathcal{N})$ is elementarily equivalent to $\mathcal{N}$ (in other words, $M \circ K = \mathsf{id}_U$ in INT_3). Then U and V are synonymous.*

Here is a different formulation: if K, M witness that V is an INT_0-retract of U and that U is an INT_3-retract of V, then U and V are synonymous.

Proof. Consider any model $\mathcal{N}$ of U. We have $\widetilde{K}\widetilde{M}\widetilde{K}\widetilde{M}(\mathcal{N}) = \widetilde{K}\widetilde{M}(\mathcal{N})$. Let $\mathcal{P} := \widetilde{K}\widetilde{M}(\mathcal{N})$. So, $\widetilde{K}\widetilde{M}(\mathcal{P}) = \mathcal{P}$. We note that the identity of $\widetilde{K}\widetilde{M}(\mathcal{P})$ and $\mathcal{P}$ is witnessed by such statements as $\forall x\, \delta_{M \circ K}(x)$ and $\forall \vec{x}\, (P_{M \circ K}\vec{x} \leftrightarrow P\vec{x}\,)$. Since $\mathcal{P}$ is elementarily equivalent to $\mathcal{N}$, we have $\widetilde{K}\widetilde{M}(\mathcal{N}) = \mathcal{N}$. So $M \circ K = \mathsf{id}_U$ in INT_0. □

There is, of course, also a model-free proof of the result.

THEOREM 5.2. *Suppose that $K : U \to V$ and $M : V \to U$ witness that V is an INT_1-retract of U and that U is an INT_3-retract of V. Suppose further that M is direct. Then U and V are synonymous.*

Proof. Since M is direct, it follows that, in V, we have $\delta_{K \circ M} = \delta_K$. We replace K by a definably isomorphic direct interpretation K'. Suppose F is the promised isomorphism between $K \circ M$ and id_V. We take, for P of arity n, in the signature of U:

- $P_{K'}(v_0, \dots, v_{n-1}) :\leftrightarrow \exists \vec{u}_0 \in \delta_K, \dots, \exists \vec{u}_{n-1} \in \delta_K$
  $(\bigwedge_{i<n} \vec{u}_i\, F\, v_i \wedge P_K(\vec{u}_0, \dots, \vec{u}_{n-1}))$.

Clearly, we have an isomorphism $F' : K \to K'$, based on the same underlying formula as F. Hence K', M witness that U is an INT_3-retract of V.

We note that $K' \circ M$ is direct. Suppose R is an m-ary predicate of V. The theory T proves:

$$\begin{aligned} R_{K' \circ M}(x_0, \dots, x_{m-1}) &\leftrightarrow (R_M(x_0, \dots, x_{m-1}))^{K'} \\ &\leftrightarrow \exists \vec{y}_0 \in \delta_K, \dots, \exists \vec{y}_{m-1} \in \delta_K \\ &\qquad (\bigwedge_{j<m} \vec{y}_j\, F\, x_j \wedge (R_M(\vec{y}_0, \dots, \vec{y}_{m-1}))^K) \\ &\leftrightarrow \exists \vec{y}_0 \in \delta_K, \dots, \exists \vec{y}_{m-1} \in \delta_K \\ &\qquad (\bigwedge_{j<m} \vec{y}_j\, F\, x_j \wedge R_{K \circ M}(\vec{y}_0, \dots, \vec{y}_{m-1})) \\ &\leftrightarrow R(x_0, \dots, x_{m-1}). \end{aligned}$$

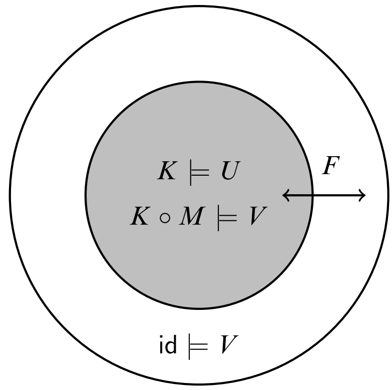

Figure 1. Illustration of the Proof of Theorem 5.2.

Thus, we find: $K' \circ M = \mathsf{id}_V$ in INT_0, in other words, V is an INT_0-retract of U. We apply Theorem 5.1 to K', M to obtain the desired result that U and V are synonymous (Figure 1). □

We are mainly interested in the following corollary.

Corollary 5.1. *Suppose U and V are bi-interpretable and one of the witnessing interpretations is direct. Then U and V are synonymous.*

We now prove our main theorem.

Theorem 5.3. *Suppose V is conceptual and that $K : U \to V$ and $M : V \to U$ form a bi-interpretation of U and V. Let K and M both be identity-preserving. Then, U and V are synonymous.*

Proof. We note that, in V, δ_K is a (virtual) subclass of the full domain. Hence, we have a definable injection from δ_K to the full domain.

Again $\delta_{K \circ M} = \delta_M^K \cap \delta_K$ is a (virtual) subclass of δ_K. Moreover, we have a definable bijection F between the full domain and $\delta_{K \circ M}$. Hence, we have a definable injection from the full domain into δ_K.

We apply the Schröder–Bernstein Theorem to the full domain and δ_K providing us with a bijection G between the full domain and δ_K. We define a new interpretation $K' : U \to V$, by setting:

- $P_{K'}(v_0, \dots, v_{n-1}) :\leftrightarrow \exists w_0 \in \delta_K, \dots, \exists w_{n-1} \in \delta_K. \; (\bigwedge_{i<n} v_i \, G \, w_i \wedge P_K(w_0, \dots, w_{n-1}))$.

Clearly, K' is direct and isomorphic to K, so K' and M form a bi-interpretation of U and V. By Corollary 5.1, we may conclude that U and V are synonymous (Figure 2). □

We note that in the circumstances of Theorem 5.3, it follows that U is also conceptual. In §7, we provide an example to illustrate that one cannot drop the demand of identity preservation for any of the two interpretations. We provide two finitely axiomatised sequential theories that are bi-interpretable, but not synonymous. One of the two witnesses of the bi-interpretation is identity preserving.

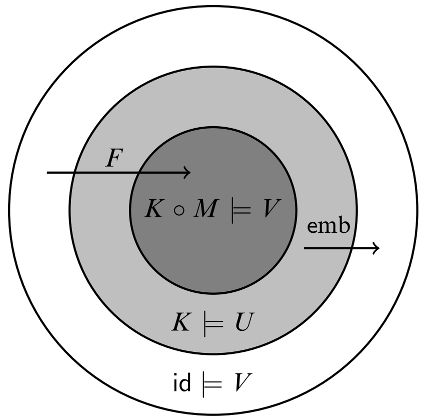

Figure 2. Illustration of the Proof of Theorem 5.3.

§6. Applications. In this section, we provide a number of applications of Theorem 5.3.

6.1. Natural numbers and integers. The theory PA^- is the theory of the non-negative part of a discretely ordered commutative ring (see Richard Kaye's book [12, chap. 2] and Emil Jeřábek paper [10] Let DOCR be the theory of a discretely ordered commutative rings. The theories PA^- and DOCR are bi-interpretable. The interpretation of PA^- in DOCR is *restriction to the non-negative part*. Kaye uses the well-known pairs construction as an interpretation of DOCR in PA^-. This construction is two-dimensional and employs an equivalence relation. As shown by Jeřábek in [10], we have a polynomial pairing function $\langle \cdot, \cdot \rangle$ in PA^-. We can define our domain as $\langle 0, x \rangle$, (representing the non-positive numbers) and $\langle x, 0 \rangle$, for all x, representing the non-negative numbers.[6] Since, PA^- has the subtraction axiom, we can define the desired operations as a matter of course. Moreover, the verification that we defined an identity preserving bi-interpretations is immediate. Thus, by Theorem 5.3, we may conclude that PA^- and DOCR are synonymous.

6.2. Natural numbers and rational numbers. Julia Robinson, in her seminal paper [22], shows that the natural numbers are definable in $\mathbb{Q}$, the field of the rationals. This gives us an identity preserving interpretation of $\mathsf{Th}(\mathbb{N})$ in $\mathsf{Th}(\mathbb{Q})$. Conversely, we can find an identity preserving interpretation of $\mathsf{Th}(\mathbb{Q})$ in $\mathsf{Th}(\mathbb{N})$ by using the Cantor pairing and by just considering pairs $\langle m, n \rangle$, where m and n have no common divisor except 1. Addition and multiplication are defined in the usual way. We can easily define internal isomorphisms witnessing that these interpretations form an identity preserving bi-interpretation. Hence, $\mathsf{Th}(\mathbb{N})$ in $\mathsf{Th}(\mathbb{Q})$ are synonymous by Theorem 5.3.

6.3. Finite sets and numbers. We consider the theory $\mathsf{ZF}^+_{\mathsf{fin}} := (\mathsf{ZF} - \mathsf{INF}) + \neg\,\mathsf{INF} + \mathsf{TC}$. This is ZF in the usual formulation minus the axiom of infinity, plus the negation of the axiom of infinity and the axiom TC that tells us that every set has a transitive closure. Kaye and Wong in their paper [13] provide a careful verification that $\mathsf{ZF}^+_{\mathsf{fin}}$

6 Alternatively, we can use the even and odd numbers as a domain. Note that in PA^- these need not be all the numbers. Jeřábek in [10] shows that the odds and the evens are mutually exclusive in PA^-.

and PA are synonymous. By Theorem 5.3, it is sufficient to show that the Ackermann interpretation of $\mathsf{ZF}^+_{\mathsf{fin}}$ in PA and the von Neumann interpretation of PA in $\mathsf{ZF}^+_{\mathsf{fin}}$ form a bi-interpretation. For further information about the related theory $\mathsf{ZF}_{\mathsf{fin}} = (\mathsf{ZF} - \mathsf{INF}) + \neg\,\mathsf{INF}$, see [7].

6.4. Sets with or without urelements. Benedikt Löwe shows that a certain version of ZF with a countable set of urelements is synonymous with ZF (see [15]. Again this result is easily obtained using Theorem 5.3.

§7. Frege meets Cantor: An example. In this section, we provide an example of two finitely axiomatized, sequential theories that are bi-interpretable but not synonymous. One of the two interpretations witnessing bi-interpretability is identity preserving. The example given is meaningful: it is the comparison of a Frege-style weak set theory and a Cantor-style weak set theory.[7]

The theory $\mathsf{ACF}_\flat$ is the one-sorted version of adjunctive class theory with Frege relation.[8] Our theory has unary predicates ob and cl, and binary predicates $\in$ and F. Here F is the Frege relation.

We will write $x:\mathsf{ob}$ for $\mathsf{ob}(x)$, $\forall x:\mathsf{ob}\cdots$ for $\forall x\,(\mathsf{ob}(x) \to \cdots)$, $\exists x:\mathsf{ob}\cdots$ for $\exists x\,(\mathsf{ob}(x) \wedge \cdots)$. Similarly for cl. We have the following axioms:

$\mathsf{ACF}_\flat 1.$ $\vdash \forall x\,(x:\mathsf{ob} \vee x:\mathsf{cl})$,
$\mathsf{ACF}_\flat 2.$ $\vdash \forall x\,\neg\,(x:\mathsf{ob} \wedge x:\mathsf{cl})$,
$\mathsf{ACF}_\flat 3.$ $\vdash \forall x, y\,(x \in y \to (x:\mathsf{ob} \wedge y:\mathsf{cl}))$,
$\mathsf{ACF}_\flat 4.$ $\vdash \forall x, y\,(x\mathsf{F}y \to (x:\mathsf{ob} \wedge y:\mathsf{cl}))$,
$\mathsf{ACF}_\flat 5.$ $\vdash \exists x:\mathsf{cl}\,\forall y:\mathsf{ob}\;\; y \notin x$,
$\mathsf{ACF}_\flat 6.$ $\vdash \forall x:\mathsf{cl}\,\forall y:\mathsf{ob}\,\exists z:\mathsf{cl}\,\forall w:\mathsf{ob}\,(w \in z \leftrightarrow (w \in x \vee w = y))$.
$\mathsf{ACF}_\flat 7.$ $\vdash \forall x, y:\mathsf{cl}\,(\forall z:\mathsf{ob}\,(z \in x \leftrightarrow z \in y) \to x = y)$.
$\mathsf{ACF}_\flat 8.$ $\vdash \forall x:\mathsf{ob}\,\exists y:\mathsf{cl}\; x\mathsf{F}y$,
$\mathsf{ACF}_\flat 9.$ $\vdash \forall x:\mathsf{ob}\,\forall y, y':\mathsf{cl}\,((x\mathsf{F}y \wedge x\mathsf{F}y') \to y = y')$,
$\mathsf{ACF}_\flat 10.$ $\vdash \forall x:\mathsf{cl}\,\exists y:\mathsf{ob}\; y\mathsf{F}x$.

We provide one-dimensional interpretations witnessing that AS and $\mathsf{ACF}_\flat$ are bi-interpretable. Note that by Theorem B.1, it follows that $\mathsf{ACF}_\flat$ is sequential.

In the context of AS, we write:

- $\mathsf{pair}(x, y, z) :\leftrightarrow \exists u, v\,\forall w\,((w \in u \leftrightarrow w = x) \wedge$
  $(w \in v \leftrightarrow (w = x \vee w = y)) \wedge (w \in z \leftrightarrow (w = u \vee w = v)))$,
- $\mathsf{Pair}(x) :\leftrightarrow \exists y, z\,\mathsf{pair}(y, z, x)$,
- $\pi_0(z, x) :\leftrightarrow \exists y\,\mathsf{pair}(x, y, z)$, $\pi_1(z, y) :\leftrightarrow \exists x\,\mathsf{pair}(x, y, z)$,
- $\mathsf{empty}(x) :\leftrightarrow \forall y\; y \notin x$, $\mathsf{inhab}(x) :\leftrightarrow \neg\,\mathsf{empty}(x)$,
- $x \approx y :\leftrightarrow \forall z\,(z \in x \leftrightarrow z \in y)$.

[7] Since the first preprint of this paper came out in 2014, other separating examples were found (see, e.g., [6]).

[8] It would be more natural to give the example for the two-sorted theory ACF and to use the notion of piecewise interpretation. The definitions of the interpretations and the verification that they form a bi-interpretation would be simpler. In fact, we would avoid the use of coding in our definitions. However, we would need to develop more of the theory of many-sorted interpretations to handle the superior approach smoothly. This is beyond the scope of our present paper.

We can verify the usual properties of pairing. The π_i are functional on Pair. We will write them using functional notation. We should remember that they are undefined outside Pair. We first define an interpretation $L : \mathsf{ACF}_\flat \to \mathsf{AS}$.

- $\delta_L(x) :\leftrightarrow \mathsf{Pair}(x)$,
- $\mathsf{ob}_L(x) :\leftrightarrow \mathsf{Pair}(x) \wedge \mathsf{empty}(\pi_0(x))$,
- $\mathsf{cl}_L(x) :\leftrightarrow \mathsf{Pair}(x) \wedge \mathsf{inhab}(\pi_0(x))$,
- $x =_L y :\leftrightarrow (x, y : \mathsf{ob}_L \wedge \pi_1(x) = \pi_1(y)) \vee (x, y : \mathsf{cl}_L \wedge \pi_1(x) \approx \pi_1(y))$,
- $x \in_L y :\leftrightarrow x : \mathsf{ob}_L \wedge y : \mathsf{cl}_L \wedge \pi_1(x) \in \pi_1(y)$,
- $x \,\mathsf{F}_L\, y : \leftrightarrow x : \mathsf{ob}_L \wedge y : \mathsf{cl}_L \wedge \pi_1(x) \approx \pi_1(y)$.

It is easy to see that the specified translation does carry an interpretation of $\mathsf{ACF}_\flat$ in AS, as promised. Next, we define an interpretation $K : \mathsf{AS} \to \mathsf{ACF}_\flat$.

- $\delta_K(x) :\leftrightarrow x : \mathsf{ob}$,
- $x =_K y \leftrightarrow x, y : \mathsf{ob} \wedge x = y$,
- $x \in_K y \leftrightarrow x, y : \mathsf{ob} \wedge \exists z : \mathsf{cl}\,(x \in z \wedge y \mathsf{F} z)$.

We note that K is identity preserving. The verification that our interpretations do indeed specify a bi-interpretation is entirely routine. For completeness' sake, we provide the computations involved in Appendix C.

We show that AS and $\mathsf{ACF}_\flat$ are not synonymous—not even when we allow parameters. We build the following model $\mathcal{M}$ of AS. The domain is inductively specified as the smallest set M such that if X is a finite subset of M, then $\langle 0, X\rangle$ and $\langle 1, X\rangle$ are in M. Let $m, n, \ldots$ range over M. We define: $m \in^\star n$ iff $n = \langle i, X\rangle$ and $m \in X$. It is easily seen that $\mathcal{M}$ is indeed a model of AS. Clearly, for any finite subset X_0 of M, we can find an automorphism σ of $\mathcal{M}$ of order 2 that fixes X_0 and fixes only finitely many elements of M.

Suppose AS and $\mathsf{ACF}_\flat$ were synonymous. Let $\mathcal{N}$ be the internal model of $\mathsf{ACF}_\flat$ in $\mathcal{M}$ given by the synonymy. Say the interpretation is P, involving a finite set of parameters X_0. Let σ be an automorphism of order 2 on $\mathcal{M}$ that fixes X_0 and that fixes at most finitely many objects. Consider the classes $\{p, \sigma p\}^{\mathcal{N}}$, where p is in $\mathsf{ob}^{\mathcal{N}}$. (Note that σ must send $\mathcal{N}$-objects to $\mathcal{N}$-objects.) Clearly there is an infinity of such classes. By extensionality, these classes are fixed by σ. This contradicts the fact that σ has only finitely many fixed points.

It is well known that if two models are bi-interpretable (without parameters) then their automorphism groups are isomorphic. Our example shows that the *action* of these automorphism groups on the elements can be substantially different.

§A. Definitions. In this appendix, we provide detailed definitions of translations, interpretations and morphisms between interpretations.

A.1. Translations. Translations are the heart of our interpretations. In fact, they are often confused with interpretations, but we will not do that officially. In practice, it is often convenient to conflate an interpretation and its underlying translation.

A proto-formula is a λ-term $\lambda\vec{x}.A(\vec{x})$, where the variables of the formula A are among those in $\vec{x}$. We will think of proto-formulas modulo α-conversion and we will follow the conventions of the λ-calculus for substitution, where we avoid variable-capture by renaming the bound variables. The arity is a proto-formula is the length of $\vec{x}$.

We define more-dimensional, one-sorted, one-piece relative translations without parameters. Let Σ and Θ be one-sorted signatures. A translation $\tau : \Sigma \to \Theta$ is given by a triple $\langle m, \delta, F\rangle$. Here δ will be an m-ary proto-formula of signature Θ. The mapping F associates to each relation symbol R of Σ with arity n an $m \times n$-ary proto-formula of signature Θ.

We demand that predicate logic proves $F(R)(\vec{x}_0, \dots, \vec{x}_{n-1}) \to (\delta(\vec{x}_0) \wedge \cdots \delta(\vec{x}_{n-1}))$. Of course, given any candidate proto-formula $F(R)$ not satisfying the restriction, we can obviously modify it to satisfy the restriction.

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

- $(R(x_0, \dots, x_{n-1}))^\tau := F(R)(\vec{x}_0, \dots, \vec{x}_{n-1})$. We use sloppy notation here. The single variable x_i of the source language needs to have no obvious connection with the sequence of variables $\vec{x}_i$ of the target language that represents it. We need some conventions to properly handle the association $x_i \mapsto \vec{x}_i$. We do not treat these details here. We demand that the $\vec{x}_i$ are fully disjoint when the x_i are different.
- $(\cdot)^\tau$ commutes with the propositional connectives;
- $(\forall x\, A)^\tau := \forall \vec{x}\, (\delta(\vec{x}\,) \to A^\tau)$;
- $(\exists x\, A)^\tau := \exists \vec{x}\, (\delta(\vec{x}\,) \wedge A^\tau)$.

Here are some convenient conventions and notations.

- We write δ_τ for 'the δ of τ' and F_τ for 'the F of τ'.
- We write R_τ for $F_\tau(R)$.
- We write $\vec{x} \in \delta_\tau$ for: $\delta_\tau(\vec{x}\,)$.

There are some natural operations on translations. The identity translation $\mathsf{id} := \mathsf{id}_\Theta$ is one-dimensional and it is defined by:

- $\delta_{\mathsf{id}} := \lambda x.(x = x)$,
- $R_{\mathsf{id}} := \lambda \vec{x}.R\vec{x}$.

We can compose relative translations as follows. Suppose τ is an m-dimensional translation from Σ to Θ, and ν is a k-dimensional translation from Θ to Ξ. We define:

- We suppose that with the variable x we associate under τ the sequence $x_0, \dots, x_{m-1}$ and under ν we send x_i to $\vec{x}_i$. $\delta_{\tau\nu}(\vec{x}_0, \dots, \vec{x}_{m-1}) := (\delta_\nu(\vec{x}_0) \wedge \cdots \wedge \delta_\nu(\vec{x}_{m-1}) \wedge (\delta_\tau(x))^\nu)$,
- Let R be n-ary. Suppose that under τ we associate with x_i the sequence $x_{i,0}, \dots, x_{i,m-1}$ and that under ν we associate with $x_{i,j}$ the sequence $\vec{x}_{i,j}$. We take: $R_{\tau\nu}(\vec{x}_{0,0}, \dots \vec{x}_{n-1,m-1}) = \delta_\nu(\vec{x}_{0,0}) \wedge \cdots \wedge \delta_\nu(\vec{x}_{n-1,m-1}) \wedge (R_\tau(x_0, \dots, x_{n-1}))^\nu$.

A one-dimensional translation τ *preserves identity* if $(x =_\tau y) = (x = y)$. A one-dimensional translation τ *is unrelativized* if $\delta_\tau(x) = (x = x)$. A one-dimensional translation τ is *direct* if it is unrelativized and preserves identity. Note that all these properties are preserved by composition.

Consider a model $\mathcal{M}$ with domain M of signature Θ and k-dimensional translation $\tau : \Sigma \to \Theta$. Suppose that $N := \{\vec{m} \in M^k \mid \mathcal{M} \models \delta_\tau \vec{m}\}$. Then τ specifies an internal model $\mathcal{N}$ of $\mathcal{M}$ with domain N and with $\mathcal{N} \models R(\vec{m}_0, \dots, \vec{m}_{n-1})$ iff $\mathcal{M} \models R_\tau(\vec{m}_0, \dots, \vec{m}_{n-1})$. We will write $\widetilde{\tau}(\mathcal{M})$ for the internal model of $\mathcal{M}$ given by τ. We treat the mapping $\tau, \mathcal{M} \mapsto \widetilde{\tau}\mathcal{M}$ as a partial function that is defined precisely if

$\delta_\tau^{\mathcal{M}}$ is non-empty. Let Mod or $\widetilde{(\cdot)}$ be the function that maps τ to $\widetilde{\tau}$. We have:

$$\mathsf{Mod}(\tau \circ \rho)(\mathcal{M}) = (\mathsf{Mod}(\rho) \circ \mathsf{Mod}(\tau))(\mathcal{M}).$$

So, Mod behaves contravariantly.

A.2. 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 and assures us that δ_τ is inhabited. We will write $K = \langle U, \tau, V\rangle$ for the interpretation supported by τ. We write $K : U \to V$ for: K is an interpretation of the form $\langle U, \tau, V\rangle$. If M is an interpretation, τ_M will be its second component, so $M = \langle U, \tau_M, V\rangle$, for some U and V.

Par abus de langage, we write 'δ_K' for δ_{τ_K}; we write 'R_K' for R_{τ_K} and we write 'A^K' for A^{τ_K}, etc. Here are the definitions of three central operations on interpretations.

- Suppose T has signature Σ. We define: $\mathsf{id}_T : T \to T$ is $\langle T, \mathsf{id}_\Sigma, T\rangle$.
- Suppose $K : U \to V$ and $M : V \to W$. We define: $M \circ K : U \to W$ is $\langle U, \tau_M \circ \tau_K, W\rangle$.

It is easy to see that we indeed correctly defined interpretations between the theories specified.

A.3. Equality of interpretations. Two interpretations are *equal* when the *target theory thinks they are*. Specifically, we count two interpretations $K, K' : U \to V$ as equal if they have the same dimension, say m, and:

- $V \vdash \forall \vec{x}\ (\delta_K(\vec{x}\,) \leftrightarrow \delta_{K'}(\vec{x}\,))$,
- $V \vdash \forall \vec{x}_0, \dots, \vec{x}_{n-1} \in \delta_K\ (R_K(\vec{x}_0, \dots, \vec{x}_{n-1}) \leftrightarrow R_{K'}(\vec{x}_0, \dots, \vec{x}_{n-1}))$.

Modulo this identification, the operations identity and composition give rise to a category INT_0, where the theories are objects and the interpretations arrows.[9]

Let MOD be the category with as objects classes of models and as morphisms all functions between these classes. We define $\mathsf{Mod}(U)$ as the class of all models of U. Suppose $K : U \to V$. Then, $\mathsf{Mod}(K)$ is the function from $\mathsf{Mod}(V)$ to $\mathsf{Mod}(U)$ given by: $\mathcal{M} \mapsto \widetilde{K}(\mathcal{M}) := \widetilde{\tau_K}(\mathcal{M})$. It is clear that Mod is a *contravariant functor* from INT_0 to MOD.

A.4. Maps between interpretations. Consider $K, M : U \to V$. Suppose K is m-dimensional and M is k-dimensional. A V-definable, V-provable morphism from K to M is a triple $\langle K, F, M\rangle$, where F is an $m + k$-ary proto-formula.[10] We write $\vec{x}\ F\ \vec{y}$ for $F(\vec{x}, \vec{y}\,)$. We demand that F has the following properties:

- $V \vdash \vec{x}\ F\ \vec{y} \to (\vec{x} \in \delta_K \wedge \vec{y} \in \delta_M)$.
- $V \vdash \vec{x} =_K \vec{u}\ F\ \vec{v} =_M \vec{y} \to \vec{x}\ F\ \vec{y}$.
- $V \vdash \forall \vec{x} \in \delta_K\ \exists \vec{y} \in \delta_M\ \vec{x}\ F\ \vec{y}$.
- $V \vdash (\vec{x}\ F\ \vec{y} \wedge \vec{x}\ F\ \vec{z}\,) \to \vec{y} =_M \vec{z}$.
- $V \vdash (\vec{x}_0 F \vec{y}_0 \wedge \cdots \wedge \vec{x}_{n-1} F \vec{y}_{n-1} \wedge R_K(\vec{x}_0, \dots, \vec{x}_{n-1})) \to R_M(\vec{y}_0, \dots, \vec{y}_{n-1})$.

9 For many reasons, the choice for the reverse direction of the arrows would be more natural. However, our present choice coheres with the extensive tradition in degrees of interpretability. So, we opted to adhere to the present choice.

10 Since, in this stage, we are looking at definitions without parameters we could, perhaps, better speak of *V-0-definable*. Parameters may be added but in the context where we consider theories rather than models some extra details are needed to make everything work smoothly.

We will call the arrows between interpretations: *i-maps* or *i-morphisms*. We write $F : K \to M$ for: $\langle K, F, M\rangle$ is a V-provable, V-definable morphism from K to M. Remember that the theories U and V are part of the data for K and M. We consider $F, G : K \to M$ as *equal* when they are V-provably the same.

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

- $V \vdash \forall \vec{y} \in \delta_M \, \exists \vec{x} \in \delta_K \; \vec{x} \; F \; \vec{y}$,
- $V \vdash (\vec{x} \; F \; \vec{y} \wedge \vec{z} \; F \; \vec{y}\,) \to \vec{x} =_K \vec{z}$,
- $V \vdash (\vec{x}_0 F \vec{y}_0 \wedge \cdots \wedge \vec{x}_{n-1} F \vec{y}_{n-1} \wedge R_M(\vec{y}_0, \dots, \vec{y}_{n-1})) \to R_K(\vec{x}_0, \dots, \vec{x}_{n-1})$.

We call such isomorphisms: i-isomorphisms. By a simple compactness argument, one may prove the following.

THEOREM A.1. *Suppose the signature of U is finite. Consider $K, M : U \to V$. Suppose that, for every model $\mathcal{N}$ of V, there is an $\mathcal{N}$-definable isomorphism between $\widetilde{K}(\mathcal{N})$ and $\widetilde{M}(\mathcal{N})$. Then, K and M are i-isomorphic.*

A.5. Adding parameters. We can add parameters in the obvious way. An interpretation $K : U \to V$ with parameters will have a k-dimensional parameter domain α (officially a proto-formula), where $V \vdash \exists \vec{x} \; \alpha \vec{x}$. We allow the extra variables $\vec{x}$ to occur in the translations of the U formulas. We have to take the appropriate measures to avoid variable-clashes. The condition for K to be an interpretation changes into: $V \vdash \forall \vec{x} \; (\alpha \vec{x} \to A^{K, \vec{x}})$, where A is a theorem of U.

We note that, in the presence of parameters, the function $\widetilde{K}$ associates a class of models of U to a model of V.

Similar adaptations are needed to define i-isomorphisms with parameters.

§B. Background for sequentiality and conceptuality. The notion of sequential theory was introduced by Pavel Pudlák in his paper [20]. Pudlák uses his notion for the study of the degrees of local multi-dimensional parametric interpretability. He proves that sequential theories are prime in this degree structure. In [21], sequential theories provide the right level of generality for theorems about consistency statements.

The notion of sequential theory was independently invented by Friedman who called it *adequate theory* (see Smoryński's survey [23]). Friedman uses the notion to provide the Friedman characterization of interpretability among finitely axiomatized sequential theories (see also [26, 27]). Moreover, he shows that ordinary interpretability and faithful interpretability among finitely axiomatized sequential theories coincide (see also [28, 30]).

The story of the weak set theory AS can be traced in the following papers: [4, 16, 18, 21, 24], [17, appendix III], [32, 33]. The connection between AS and sequentiality is made in [17, 21].

For further work concerning sequential theories, see, e.g., [8, 11, 28–30, 34]. The paper [34] gives surveys many aspects of sequentiality.

A theorem that is relevant in this paper is Theorem 10.7 of [31].

THEOREM B.1. *Sequentiality is preserved to* INT_1*-retracts for one-dimensional interpretations. In other words: if V is sequential and if U is a one-dimensional retract in* INT_1 *of V, then U is sequential.*

Proof. Suppose $K : U \to V$ and $M : V \to U$ are one-dimensional and $M \circ K$ is i-isomorphic to id_U via F. Let $\in^\star$ be the V-formula witnessing the sequentiality of V. We define the U-formula $\in^*$ witnessing the sequentiality of U by: $x \in^* y$ iff y is in δ_M and, for some z with $z\, F\, x$, we have $(z \in^\star y)^M$. □

This result holds only for one-dimensional interpretability. There are examples of non-sequential theories that are bi-interpretable with a sequential theory. Since bi-interpretablity is such a good notion of sameness of theories, one could argue that the failure of closure of sequential theories under bi-interpretability is a defect and that we need a slightly more general notion to fully reflect the intuitions that sequentiality is intended to capture. For an elaboration of this point, see [34].

We can easily adapt Theorem B.1, to obtain the following.

THEOREM B.2. *Conceptuality is preserved to* INT_1*-retracts.*

§C. Verification of bi-interpretability. We verify that the interpretations K and L of §7 do indeed form a bi-interpretation. We first compute $M := (L \circ K) : \mathsf{AS} \to \mathsf{AS}$. We find:

- $\delta_M(x) \leftrightarrow (\delta_L(x) \wedge (x \in \delta_K)^L) \leftrightarrow (x : \mathsf{Pair} \wedge \mathsf{empty}(\pi_0(x)))$,
- We have (using the contextual information that x and y are in δ_M):

$$\begin{aligned} x =_M y &\leftrightarrow (x =_K y)^L \\ &\leftrightarrow \pi_1(x) = \pi_1(y). \end{aligned}$$

- We have (using the contextual information that x and y are in δ_M):

$$\begin{aligned} x \in_M y &\leftrightarrow (\exists z{:}\mathsf{cl}\,(x \in z \wedge y\mathsf{F}z))^L \\ &\leftrightarrow \exists z{:}\mathsf{pair}\,(\mathsf{inhab}(\pi_0(z)) \wedge \pi_1(x) \in \pi_1(z) \wedge \pi_1(y) \approx \pi_1(z)) \\ &\leftrightarrow \pi_1(x) \in \pi_1(y). \end{aligned}$$

Clearly π_1 is the desired isomorphism from M to $\mathsf{id}_{\mathsf{AS}}$.

In the other direction, let $N := (K \circ L) : \mathsf{ACF}_\flat \to \mathsf{ACF}_\flat$. We first note a simple fact about K. Since F is functional on classes we will use functional notation for it. We have, for $u, v : \mathsf{ob}$,

$$\begin{aligned} u \approx_K v &\leftrightarrow \forall w\,(w \in_K u \leftrightarrow w \in_K v) \\ &\leftrightarrow \forall w\,(w \in \mathsf{F}(u) \leftrightarrow w \in \mathsf{F}(v)) \\ &\leftrightarrow \mathsf{F}(u) = \mathsf{F}(v). \end{aligned}$$

We have:

- $\delta_N(x) \leftrightarrow (x \in \delta_K(x) \wedge (\delta_L(x))^K) \leftrightarrow (x : \mathsf{ob} \wedge \mathsf{Pair}^K(x))$,
- $\mathsf{ob}_N(x) \leftrightarrow (\mathsf{empty}(\pi_0(x)))^K$,
- $\mathsf{cl}_N(x) \leftrightarrow (\mathsf{inhab}(\pi_0(x)))^K$,
- $x =_N y \leftrightarrow ((\mathsf{ob}_N(x) \wedge \mathsf{ob}_N(y) \wedge \pi_1^K(x) = \pi_1^K(y)) \vee$ $(\mathsf{cl}_N(x) \wedge \mathsf{cl}_N(y) \wedge \mathsf{F}(\pi_1^K(x)) = \mathsf{F}(\pi_1^K(y))))$,
- $x \in_N y \leftrightarrow (\mathsf{ob}_N(x) \wedge \mathsf{cl}_N(y) \wedge \pi_1^K(x) \in_K \pi_1^K(y))$.
- $x\, \mathsf{F}_N\, y :\leftrightarrow (\mathsf{ob}_N(x) \wedge \mathsf{cl}_N(y) \wedge \mathsf{F}(\pi_1^K(x)) = \mathsf{F}(\pi_1^K(y)))$.

We define $G : N \to \mathsf{id}_{\mathsf{ACF}_\flat}$ as follows:

- $x\, G\, y :\leftrightarrow (\mathsf{ob}_N(x) \wedge \mathsf{ob}(y) \wedge \pi_1^K(x) = y) \vee (\mathsf{cl}_N(x) \wedge \mathsf{cl}(y) \wedge \mathsf{F}(\pi_1^K(x)) = y)$.

 https://doi.org/10.1017/S1755020325100920

We note that K is identity preserving. Thus, e.g., we find that pair^K is a true pairing on ob. It is easy to see that, in $\mathsf{ACF}_\flat$, the virtual classes ob_N and cl_N form a partition of δ_N. It is now trivial to check that G is indeed an isomorphism.

Acknowledgments. We thank Tonny Hurkens who simplified our proof of the Schröder–Bernstein Theorem. We present his version here with his permission.

We thank Allan van Hulst who verified our proof of the Schröder–Bernstein Theorem in Mizar.

We are grateful to Peter Aczel, Ali Enayat, and Wilfrid Hodges for stimulating conversations and/or e-mail correspondence. We thank the anonymous referee for his helpful reports.

We are grateful to Piotr Gruza and Leszek Kołodziejczyk who spotted a mistake in an earlier version of this paper and to Tim Button who, independently, spotted the same mistake.

Funding. H.M.F.'s research on this paper was partially supported by the John Templeton Foundation grant ID no. 36297. The opinions expressed here are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

BIBLIOGRAPHY

[1] Ahlbrandt, G., & Ziegler, M. (1986). Quasi finitely axiomatizable totally categorical theories. *Annals of Pure and Applied Logic*, **30**(1), 63–82.

[2] de Bouvère, K. L. (1965). Logical synonymy. *Indagationes Mathematicae*, **27**, 622–629.

[3] ———. (1965). Synonymous theories. In Addison, J. W., Henkin, L., and Tarski, A., editors. *The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*. Amsterdam: North Holland, pp. 402–406.

[4] Collins, G. E., & Halpern, J. D. (1970). On the interpretability of arithmetic in set theory. *Notre Dame Journal of Formal Logic*, **11**(4), 477–483.

[5] Corcoran, J., Frank, W., & Maloney, M. (1974). String theory. *The Journal of Symbolic Logic*, **39**(4), 625–636.

[6] Enayat, A., & Łełyk, M. (2024). Categoricity-like properties in the first order realm. *Journal for the Philosophy of Mathematics*, **1**, 63–98.

[7] Enayat, A., Schmerl, J., & Visser, A. (2010). ω-models of finite set theory. In Kennedy, J., and Kossak, R., editors. *Set Theory, Arithmetic and Foundations of Mathematics: Theorems, Philosophies*, Number 36 in ASL Lecture Notes in Logic. New York: ASL and Cambridge University Press, pp. 43–65.

[8] Hájek, P., & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin: Perspectives in Mathematical Logic. Springer.

[9] Hodges, W. (1993). *Model Theory*, Encyclopedia of Mathematics and its Applications, 42. Cambridge: Cambridge University Press.

[10] Jeřábek, E. (2012). Sequence encoding without induction. *Mathematical Logic Quarterly*, **58**(3), 244–248.

[11] Joosten, J. J., & Visser, A. (2000). The interpretability logic of *all* reasonable arithmetical theories. *Erkenntnis*, **53**(1–2), 3–26.

[12] Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford: Oxford Logic Guides. Oxford University Press.

[13] Kaye, R., & Wong, T. L. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, **48**(4), 497–510.

[14] Levy, A. (1979). *Basic Set Theory*, Perspectives in Mathematical Logic. Berlin Heidelberg New York: Springer-Verlag.

[15] Löwe, B. (2006). Set theory with and without urelements and categories of interpretations. *Notre Dame Journal of Formal Logic*, **47**(1), 83–91.

[16] Montagna, F., & Mancini, A. (1994). A minimal predicative set theory. *Notre Dame Journal of Formal Logic*, **35**(2), 186–203.

[17] Mycielski, J., Pudlák, P., & Stern, A. S. (1990). *A Lattice of Chapters of Mathematics (Interpretations between Theorems)*, volume 84 of Memoirs of the American Mathematical Society. Providence: AMS.

[18] Nelson, E. (1986). *Predicative Arithmetic*. Princeton: Princeton University Press.

[19] Pakhomov, F. (2022). How to escape Tennenbaum's theorem, preprint, arXiv:2209.00967.

[20] Pudlák, P. (1983). Some prime elements in the lattice of interpretability types. *Transactions of the American Mathematical Society*, **280**, 255–275.

[21] ———. (1985). Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, **50**(2), 423–441.

[22] Robinson, J. (1949). Definability and decision problems in arithmetic. *Journal of Symbolic Logic*, **14**(2), 98–114.

[23] Smoryński, C. (1985). Nonstandard models and related developments. In Harrington, L. A., Morley, M. D., Scedrov, A., and Simpson, S. G., editors. *Harvey Friedman's Research on the Foundations of Mathematics*. Amsterdam: North Holland, pp. 179–229.

[24] Szmielew, W., & Tarski, A. (1952). Mutual interpretability of some essentially undecidable theories. In Graves, L. M., Smith, P. A., Hille, E. & Zariski, O., editors. *Proceedings of the International Congress of Mathematicians (Cambridge, Massachusetts, 1950)*, Vol. 1. Providence: American Mathematical Society, p. 734.

[25] Van Orman, W., & Quine. (1946). Concatenation as a basis for arithmetic. *The Journal of Symbolic Logic*, **11**(4), 105–114

[26] Visser, A. (1990). Interpretability logic. In Petkov, P. P., editors. *Mathematical Logic, Proceedings of the Heyting 1988 Summer School in Varna, Bulgaria*. Boston: Plenum Press, pp. 175–209.

[27] ———. (1992). An inside view of EXP. *The Journal of Symbolic Logic*, **57**(1), 131–165.

[28] ———. (1993). The unprovability of small inconsistency. *Archive for Mathematical Logic*, **32**(4), 275–298.

[29] ———. (1998). An overview of interpretability logic. In Kracht, M., de Rijke, M., Wansing, H., and Zakharyaschev, M., editors. *Advances in Modal Logic*, volume 1, 87 of CSLI Lecture Notes. Stanford: Center for the Study of Language and Information, pp. 307–359.

[30] ———. (2005). Faith & falsity: A study of faithful interpretations and false Σ_1^0-sentences. *Annals of Pure and Applied Logic*, **131**(1–3), 103–131.

[31] ———. (2006). Categories of theories and interpretations. In Enayat, A., Kalantari, I., and Moniri, M., editors. *Logic in Tehran. Proceedings of the Workshop*

and Conference on Logic, Algebra and Arithmetic, Held October 18–22, 2003, Volume 26 of Lecture Notes in Logic. Wellesley: ASL, A.K. Peters, Ltd., pp. 284–341.

[32] ———. (2008). Pairs, sets and sequences in first order theories. *Archive for Mathematical Logic*, **47**(4), 299–326.

[33] ———. (2009). Cardinal arithmetic in the style of Baron von Münchhausen. *Review of Symbolic Logic*, **2**(3), 570–589.

[34] ———. (2013). What is sequentiality? In Cégielski, P., Cornaros, C., and Dimitracopoulos, C., editors. *New Studies in Weak Arithmetics*, volume 211 of CSLI Lecture Notes. Stanford: CSLI Publications and Presses Universitaires du Pôle de Recherche et d'Enseingement Supérieur Paris-est, pp. 229–269.

DEPARTMENT OF MATHEMATICS
THE OHIO STATE UNIVERSITY
MATHEMATICS BUILDING
231 WEST 18TH AVENUE
COLUMBUS, OH 43210, USA
E-mail: friedman@math.ohio-state.edu

DEPARTMENT OF PHILOSOPHY AND RELIGIOUS STUDIES
FACULTY OF HUMANITIES
UTRECHT UNIVERSITY
JANSKERKHOF 13
3512 BL UTRECHT, THE NETHERLANDS
E-mail: a.visser@uu.nl