Bi-interpretability from a Categorical Viewpoint

Hisashi Aratake*

Research Institute for Mathematical Sciences, Kyoto University, Kyoto, Japan

1 Introduction

First-order categorical logic (FOCL for short) originated as a categorical foundation for model theory (in Makkai & Reyes [8]). As we can see in [7], some classical model-theoretic phenomena can be efficiently described in terms of FOCL. However, most concepts in modern model theory remain to be under categorical consideration. Our principal aim is to explore categorical aspects of model theory.

This short note is intended to give an exposition for model theorists of the author's recent work on categorical characterization of bi-interpretability [1]. From the viewpoint of FOCL, classical first-order theories give rise to **Boolean pretoposes**, i.e. categories equipped with logical operations and quotients of equivalence relations. They are called **classifying pretoposes** of theories (see the definition in Theorem2.12). As Harnik [3] pointed out, a construction of classifying pretoposes can be given via Shelah's eq-construction. In fact, two theories have equivalent classifying pretoposes precisely when they are bi-interpretable in the usual model-theoretic sense. In this note, we will sketch a proof of this theorem. To keep this article accessible to model theorists, we only use elementary category theory and omit most of the details of definitions and proofs. The reader is invited to consult [1] for details.

Our theorem suggests a potential of FOCL for application to model theory. For example, we can define "model-theoretic properties" of pretoposes (e.g. completeness and stability) for those invariant under bi-interpretability. An expected application will be proposed at the end of this note.

Terminologies and notational conventions. Throughout this paper, we consider many-sorted classical first-order theories. For notations and terms, we basically follow Johnstone [5, 6] (in particular, Chapter D1 in vol. 2) with several exceptions. In the following list, $\mathcal{L}$ is a fixed (many-sorted first-order) language.

- $\mathcal{L}$-Sort (resp. $\mathcal{L}$-Rel, $\mathcal{L}$-Func) denotes the set of $\mathcal{L}$-sorts (resp. relation symbols, function symbols).

E-mail address: `aratake@kurims.kyoto-u.ac.jp`

- If we mention finite strings of sorts or variables, the empty case is in consideration.
- A string of sorts is referred to as a **type**, and a string of variables as a **context**. The former is denoted by, say, $\bar{A}$ while the latter is denoted in bold face, say $\boldsymbol{x}$. If $\boldsymbol{x} \equiv x_1, \dots, x_n$ where x_i is a variable of sort A_i, we say that $\boldsymbol{x}$ is of type $\bar{A}$, and write $\boldsymbol{x} : \bar{A}$.
- We occasionally regard a 0-ary function symbol as a constant symbol.
- Let $\varphi(\boldsymbol{x})$ be an $\mathcal{L}$-formula-in-context. We often write $T \models \varphi(\boldsymbol{x})$ to mean $T \models \forall \boldsymbol{x} \varphi(\boldsymbol{x})$.
- We write $\varphi \subseteq \psi$ for $\mathcal{L}$-formulas-in-context φ, ψ, if they are in the same context (say $\boldsymbol{x} : \bar{A}$) and $T \models (\varphi(\boldsymbol{x}) \to \psi(\boldsymbol{x}))$. Although T does not appear in the notation, we will always make it clear what the ambient theory is.
- T-$\mathbf{Mod}(\mathbf{Set})_e$ denotes the category of T-models and elementary embeddings.

2 Preliminaries

2.1 Interpretation

First, we recall the notion of interpretation in model theory while we will not pursue model-theoretic aspects of interpretations (see Hodges [4, Chap. 5 §3]). Throughout this subsection, we suppose that $\mathcal{L}, \mathcal{L}'$ are languages, and that T (resp. T') is an $\mathcal{L}$-theory (resp. an $\mathcal{L}'$-theory).

Definition 2.1. Let $\partial(\boldsymbol{x})$ be an $\mathcal{L}$-formula-in-context and $\Delta(\boldsymbol{x}, \boldsymbol{x}')$ an $\mathcal{L}$-formula-in-context with $\Delta \subseteq \partial(\boldsymbol{x}) \wedge \partial(\boldsymbol{x}')$ (where $\boldsymbol{x}, \boldsymbol{x}'$ are assumed to be disjoint). We say that Δ is a **partial T-equivalence relation** (on ∂) if the following axioms are valid in T:

$$\begin{gathered}\forall \boldsymbol{x} \left[\partial(\boldsymbol{x}) \to \Delta(\boldsymbol{x}, \boldsymbol{x})\right], \\ \forall \boldsymbol{x}\boldsymbol{x}' \left[\Delta(\boldsymbol{x}, \boldsymbol{x}') \to \Delta(\boldsymbol{x}', \boldsymbol{x})\right], \\ \forall \boldsymbol{x}\boldsymbol{x}'\boldsymbol{x}'' \left[\Delta(\boldsymbol{x}, \boldsymbol{x}') \wedge \Delta(\boldsymbol{x}', \boldsymbol{x}'') \to \Delta(\boldsymbol{x}, \boldsymbol{x}'')\right].\end{gathered}$$

We say that an $\mathcal{L}$-formula $\varphi \subseteq \partial$ is **closed under** Δ when $T \models \Delta(\boldsymbol{x}, \boldsymbol{x}') \to [\varphi(\boldsymbol{x}) \leftrightarrow \varphi(\boldsymbol{x}')]$. □

To define interpretations, we need a predefinition:

Definition 2.2 (Pre-interpretation)**.** A **pre-interpretation** I of T in T' consists of the following data:

- For each sort A in $\mathcal{L}$, we have a pair $(\partial^I_A(\boldsymbol{u}), \Delta^I_A(\boldsymbol{u}, \boldsymbol{u}'))$ where ∂^I_A is an $\mathcal{L}'$-formula and Δ^I_A is a partial T'-equivalence relation on ∂^I_A.

  For each finite string of sorts $\bar{A} = A_1 \cdots A_n$, we put

$$\partial^I_{\bar{A}}(\boldsymbol{u}) \equiv \bigwedge_{i=1}^{n} \partial^I_{A_i}(\boldsymbol{u}_i), \qquad \Delta^I_{\bar{A}}(\boldsymbol{u}, \boldsymbol{u}') \equiv \bigwedge_{i=1}^{n} \Delta^I_{A_i}(\boldsymbol{u}_i, \boldsymbol{u}'_i),$$

where all the contexts $\boldsymbol{u}_i$ and $\boldsymbol{u}'_i$ are supposed to be disjoint.

- For each relation symbol $R \rightarrowtail \bar{A}$ in $\mathcal{L}$, we have an $\mathcal{L}'$-formula $R^I(\boldsymbol{u}) \subseteq \partial^I_{\bar{A}}(\boldsymbol{u})$ which is closed under $\Delta^I_{\bar{A}}$.
- For each function symbol $f\colon \bar{A} \to B$ in $\mathcal{L}$, we have an $\mathcal{L}'$-formula $\Gamma^I_f(\boldsymbol{u}, \boldsymbol{v}) \subseteq \partial^I_{\bar{A}B}(\boldsymbol{u}, \boldsymbol{v})$ such that $\Gamma^I_f(\boldsymbol{u}, \boldsymbol{v})$ is closed under $\Delta^I_{\bar{A}B}$ and the following formulas are valid in T':

$$\Delta^I_{\bar{A}}(\boldsymbol{u}, \boldsymbol{u}') \to \exists \boldsymbol{v} \left[\Gamma^I_f(\boldsymbol{u}, \boldsymbol{v}) \wedge \Gamma^I_f(\boldsymbol{u}', \boldsymbol{v})\right],$$
$$\exists \boldsymbol{u} \left[\Gamma^I_f(\boldsymbol{u}, \boldsymbol{v}) \wedge \Gamma^I_f(\boldsymbol{u}, \boldsymbol{v}')\right] \to \Delta^I_B(\boldsymbol{v}, \boldsymbol{v}').$$

□

By induction on the construction of $\mathcal{L}$-formulas, we can associate each $\mathcal{L}$-formula φ of type $\bar{A}$ with an $\mathcal{L}'$-formula $\varphi^I \subseteq \partial^I_{\bar{A}}$ which is closed under Δ.

Definition 2.3 (Interpretation). A pre-interpretation I of T in T' is said to be an **interpretation** if φ^I is valid in T' for any $\mathcal{L}$-sentence $\varphi \in T$. When I is an interpretation, we denote it by $I\colon T \to T'$. □

Let I be a pre-interpretation of T in T' and let $\mathcal{N}$ a T'-model. Then we can define a canonical $\mathcal{L}$-structure $\mathcal{M}$ by using appropriate quotients $\partial^I(\mathcal{N})/\Delta^I(\mathcal{N})$. For each relation symbol $R \rightarrowtail \bar{A}$, $R^{\mathcal{M}}$ is defined to be the subset $R^I(\mathcal{N})/\Delta^I_{\bar{A}}(\mathcal{N})$ of $\partial^I_{\bar{A}}(\mathcal{N})/\Delta^I_{\bar{A}}(\mathcal{N}) = \bar{A}^{\mathcal{M}}$ [1]. Function symbols are interpreted similarly. We will denote $\mathcal{M}$ by $\mathcal{N}|_I$. The next theorem is basic:

Theorem 2.4 (Reduction theorem, see Hodges [4, Theorem 5.3.2]). In the above notations, for any $\mathcal{L}$-formula $\varphi(\boldsymbol{x})$ and for any tuple $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) \in \partial^I_{\bar{A}}(\mathcal{N})$,

$$\mathcal{N} \models \varphi^I(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) \iff \mathcal{N}|_I \models \varphi([\boldsymbol{a}_1], \ldots, [\boldsymbol{a}_n])$$

where each $[\boldsymbol{a}_i]$ is the equivalence class of $\boldsymbol{a}_i$ in $\partial^I_{A_i}(\mathcal{N})/\Delta^I_{A_i}(\mathcal{N})$. □

Corollary 2.5. With the same assumptions as in the Theorem, the following are equivalent:

(i) I is an interpretation.

(ii) For any T'-model $\mathcal{N}$, $\mathcal{N}|_I$ is a T-model.

Proof. (i) $\iff$ $T' \models \varphi^I$ holds for any $\varphi \in T$,

$\iff$ Every T'-model $\mathcal{N}$ satisfies φ^I for any $\varphi \in T$,

$\iff$ $\mathcal{N}|_I \models \varphi$ holds for any T'-model $\mathcal{N}$ and for any $\varphi \in T$,

$\iff$ (ii).

■

In fact, the above correspondence $\mathcal{N} \mapsto \mathcal{N}|_I$ constitutes a functor

$$(-)|_I\colon T'\text{-}\mathbf{Mod}(\mathbf{Set})_e \to T\text{-}\mathbf{Mod}(\mathbf{Set})_e. \tag{2.1}$$

We will see that this functor can be obtained by composition with a coherent functor (see (2.2)).

We now recall Shelah's eq-construction:

[1] Recall that $R^I(\mathcal{N})$ is closed under $\Delta^I_{\bar{A}}(\mathcal{N})$, and hence this definition makes sense.

Definition 2.6. We construct a language $\mathcal{L}^{\mathrm{eq}}$ and an $\mathcal{L}^{\mathrm{eq}}$-theory T^{eq} as follows: For each (total) T-equivalence relation (Δ on $\bar{A}$, say), we add to $\mathcal{L}$ a new sort Q_Δ and a new function symbol $\varepsilon_\Delta \colon \bar{A} \to Q_\Delta$, obtaining $\mathcal{L}^{\mathrm{eq}}$-Sort and $\mathcal{L}^{\mathrm{eq}}$-Func. The relation symbols of $\mathcal{L}^{\mathrm{eq}}$ are the same as $\mathcal{L}$. We then have defined $\mathcal{L}^{\mathrm{eq}}$. The $\mathcal{L}^{\mathrm{eq}}$-theory T^{eq} consists of the axioms of T together with those of the forms

$$\forall y \exists \boldsymbol{x} (\varepsilon_\Delta(\boldsymbol{x}) = y) \quad \text{and} \quad \forall \boldsymbol{x} \forall \boldsymbol{x}' [\Delta(\boldsymbol{x}, \boldsymbol{x}') \leftrightarrow \varepsilon_\Delta(\boldsymbol{x}) = \varepsilon_\Delta(\boldsymbol{x}')].$$

□

For each partial T-equivalence relation Δ, we associate a total T-equivalence relation

$$\tilde{\Delta}(\boldsymbol{x}, \boldsymbol{x}') \equiv (\partial(\boldsymbol{x}) \wedge \partial(\boldsymbol{x}') \to \Delta(\boldsymbol{x}, \boldsymbol{x}')) \wedge (\neg(\partial(\boldsymbol{x}) \wedge \partial(\boldsymbol{x}')) \to \boldsymbol{x} = \boldsymbol{x}').$$

Returning to our settings, for each $\mathcal{L}$-formula $\varphi(\boldsymbol{x})$, we have an $\mathcal{L}'^{\mathrm{eq}}$-formula

$$\left(\varphi^I / \Delta^I_{\bar{A}}\right)(v) \equiv \exists \boldsymbol{u} \left[\varepsilon_{\tilde{\Delta}^I_{\bar{A}}}(\boldsymbol{u}) = v \wedge \varphi^I(\boldsymbol{u})\right]$$

where v is a variable of sort $Q_{\tilde{\Delta}^I_{\bar{A}}}$. As a result, the interpretation I gives rise to a map $\varphi(\boldsymbol{x}) \mapsto \left(\varphi^I / \Delta^I_{\bar{A}}\right)(v)$ from the set of all $\mathcal{L}$-formulas to that of all $\mathcal{L}'^{\mathrm{eq}}$-formulas.

2.2 Syntactic Category and Classifying Pretopos

Definition 2.7 (Syntactic category)**.** For any $\mathcal{L}$-theory T, we construct a category $\mathcal{C}_T$ (called the **syntactic category** of T) as follows:

- The objects are α-equivalence classes of $\mathcal{L}$-formulas-in-context. The α-equivalence class of $\varphi(\boldsymbol{x})$ is denoted by $\{\boldsymbol{x}.\, \varphi\}$.
- Assuming that the contexts $\boldsymbol{x}, \boldsymbol{y}$ are disjoint, the morphisms from $\{\boldsymbol{x}.\, \varphi\}$ to $\{\boldsymbol{y}.\, \psi\}$ are T-provably-equivalence classes of **T-provably functional formulas**. A T-provably functional formula is an $\mathcal{L}$-formula-in-context $\chi(\boldsymbol{x}, \boldsymbol{y})$ such that the formulas

  $$\chi \to \varphi \wedge \psi, \quad \varphi \to \exists \boldsymbol{y} \chi, \quad \text{and} \quad \chi \wedge \chi[\boldsymbol{z}/\boldsymbol{y}] \to \boldsymbol{y} = \boldsymbol{z}$$

  are provable in T. $[\chi]$ denotes the T-provably-equivalence class of χ.
- If $[\chi] \colon \{\boldsymbol{x}.\, \varphi\} \to \{\boldsymbol{y}.\, \psi\}$ and $[\theta] \colon \{\boldsymbol{y}.\, \psi\} \to \{\boldsymbol{z}.\, \xi\}$ are morphisms in $\mathcal{C}_T$, the composite $[\theta] \circ [\chi] \colon \{\boldsymbol{x}.\, \varphi\} \to \{\boldsymbol{z}.\, \xi\}$ is defined to be $[\exists \boldsymbol{y}(\chi \wedge \theta)]$. □

We will follow the notational conventions below:

- We frequently write $\varphi(\boldsymbol{x})$ (or more simply φ) for an object $\{\boldsymbol{x}.\, \varphi\}$.
- We often identify the object $\{\boldsymbol{x}.\, \boldsymbol{x} = \boldsymbol{x}\}$ with the type $\bar{A}$ of $\boldsymbol{x}$. We also identify the object $\{\boldsymbol{x}.\, \top\}$ with the type $\bar{A}$ of $\boldsymbol{x}$.

When we mention a subobject in a syntactic category, thanks to [6, Lemma D1.4.4(iv)], we may assume that it is represented by a formula in the same context.

Proposition 2.8. For any theory T, the syntactic category $\mathcal{C}_T$ is a **Boolean coherent category**, *i.e.*, roughly speaking,

- Any subobject poset $\mathrm{Sub}(\{\boldsymbol{x}.\,\varphi\})$ is a Boolean algebra.
- Each morphism has an image factorization.
- These categorical structures are stable under pullbacks. □

A **coherent functor** is a functor between (Boolean) coherent categories which preserves coherent structures.

Theorem 2.9 ([6, Theorem D1.4.7]). Each T-model $\mathcal{M}$ gives a coherent functor $F_{\mathcal{M}}\colon \mathcal{C}_T \to \mathbf{Set}$ which sends $\{\boldsymbol{x}.\,\varphi\}$ to the definable set $\varphi(\mathcal{M})$.

Moreover, this correspondence yields (a half part of) an equivalence of categories

$$T\text{-}\mathbf{Mod}(\mathbf{Set})_e \simeq \mathfrak{Coh}(\mathcal{C}_T, \mathbf{Set})$$

where $\mathfrak{Coh}(\mathcal{C}_T, \mathbf{Set})$ is the category of coherent functors from $\mathcal{C}_T$ to $\mathbf{Set}$ and natural transformations. □

As we saw at the end of §2.1, an interpretation I gives rise to a map $\varphi(\boldsymbol{x}) \mapsto \left(\varphi^I/\Delta^I_{\bar{A}}\right)(v)$ from the set of all $\mathcal{L}$-formulas to that of all $\mathcal{L}'^{\mathrm{eq}}$-formulas. Now we can associate I with a coherent functor:

Proposition 2.10. An interpretation $I\colon T \to T'$ induces a coherent functor $F_I\colon \mathcal{C}_T \to \mathcal{C}_{T'^{\mathrm{eq}}}$ sending a morphism $[\chi]\colon \{\boldsymbol{x}.\,\varphi\} \to \{\boldsymbol{y}.\,\psi\}$ in $\mathcal{C}_T$ to the following morphism in $\mathcal{C}_{T'^{\mathrm{eq}}}$:

$$\left[\exists \boldsymbol{u} \exists \boldsymbol{v} \left(\chi^I(\boldsymbol{u}, \boldsymbol{v}) \wedge \varepsilon_{\tilde{\Delta}^I_{\bar{A}}}(\boldsymbol{u}) = u' \wedge \varepsilon_{\tilde{\Delta}^I_{\bar{B}}}(\boldsymbol{v}) = v'\right)\right] :\ \{u'.\,\varphi^I/\Delta^I_{\bar{A}}\} \to \{v'.\,\psi^I/\Delta^I_{\bar{B}}\}\,.$$

□

We can make up $F_I\colon \mathcal{C}_T \to \mathcal{C}_{T'^{\mathrm{eq}}}$ into another coherent functor $\mathcal{P}_I\colon \mathcal{C}_{T^{\mathrm{eq}}} \to \mathcal{C}_{T'^{\mathrm{eq}}}$. The category $\mathcal{C}_{T^{\mathrm{eq}}}$ will play a special role in our theory.

Definition 2.11 (Proper theory). A classical theory T is said to be **proper** if there exists a sort D such that $T \models \exists x x'(x \neq x')$ with $x, x' : D$. □

In what follows, all theories are assumed to be proper. The following observation was made by Harnik [3].

Theorem 2.12. Let T be a theory. Then $\mathcal{C}_{T^{\mathrm{eq}}}$ is a **Boolean pretopos**, i.e. a Boolean coherent category having the following properties:

- Any "equivalence relation" has a quotient.
- Finite coproducts exist and are disjoint.

Moreover, the canonical functor $\gamma\colon \mathcal{C}_T \to \mathcal{C}_{T^{\mathrm{eq}}}$ gives a **pretopos completion** of $\mathcal{C}_T$: for any (Boolean) pretopos $\mathcal{P}$ and for any coherent functor $F\colon \mathcal{C}_T \to \mathcal{P}$, there exists a unique coherent functor $G\colon \mathcal{C}_{T^{\mathrm{eq}}} \to \mathcal{P}$ (up to natural isomorphism) such that $F \simeq G\gamma$:

$$\begin{tikzcd}
\mathcal{C}_T \arrow[r, "\gamma"] \arrow[dr, "F"'] & \mathcal{C}_{T^{\mathrm{eq}}} \arrow[d, dashed, "G"] \\
& \mathcal{P}
\end{tikzcd}$$

In fact, the following equivalences hold:

$$T\text{-}\mathbf{Mod}(\mathbf{Set})_e \simeq \mathfrak{Coh}(\mathcal{C}_T, \mathbf{Set}) \simeq \mathfrak{Coh}(\mathcal{C}_{T^{\mathrm{eq}}}, \mathbf{Set}) \simeq T^{\mathrm{eq}}\text{-}\mathbf{Mod}(\mathbf{Set})_e.$$

The **classifying pretopos** $\mathcal{P}_T$ of T is defined to be $\mathcal{C}_{T^{\mathrm{eq}}}$. □

Corollary 2.13. Combining Proposition 2.10 and Theorem 2.12, we can obtain from an interpretation $I\colon T \to T'$ another coherent functor $\mathcal{P}_I\colon \mathcal{P}_T \to \mathcal{P}_{T'}$. □

Observe that the restriction functor $(-)|_I\colon T'\text{-}\mathbf{Mod}(\mathbf{Set})_e \to T\text{-}\mathbf{Mod}(\mathbf{Set})_e$ which appeared in (2.1) can be described as the composite of the functors below:

$$\begin{tikzcd}
\mathfrak{Coh}(\mathcal{C}_{T'}, \mathbf{Set}) \arrow[r, "\sim"] \arrow[d, "\wr"] & \mathfrak{Coh}(\mathcal{P}_{T'}, \mathbf{Set}) \arrow[r, "(-)\circ\mathcal{P}_I"] & \mathfrak{Coh}(\mathcal{P}_T, \mathbf{Set}) \arrow[r, "\sim"] & \mathfrak{Coh}(\mathcal{C}_T, \mathbf{Set}) \arrow[d, "\wr"] \\
T'\text{-}\mathbf{Mod}(\mathbf{Set})_e \arrow[rrr, "(-)|_I"] & & & T\text{-}\mathbf{Mod}(\mathbf{Set})_e
\end{tikzcd} \tag{2.2}$$

3 Bi-interpretability

Let I and J be interpretations of T in T'.

Definition 3.1 (Homotopy between interpretations). A **homotopy** $h\colon I \Rightarrow J$ is a family $\{h_A\}_A$ of $\mathcal{L}'$-formulas with $h_A \subseteq \partial^I_A \wedge \partial^J_A$ such that

- Each h_A is an $\mathcal{L}'$-formula closed under $\Delta^I_A \wedge \Delta^J_A$.
- Each h_A induces an isomorphism $\tilde{h}_A\colon \partial^I_A/\Delta^I_A \xrightarrow{\sim} \partial^J_A/\Delta^J_A$ in $\mathcal{P}_{T'}$ similarly to the definition of $F_I([\chi])$ in Proposition 2.10.
- For any atomic $\mathcal{L}$-formula $\varphi(\boldsymbol{x})$, these isomorphisms induce an isomorphism $\varphi^I/\Delta^I_{\bar{A}} \xrightarrow{\sim} \varphi^J/\Delta^J_{\bar{A}}$.

If we have two homotopies $h, k\colon I \Rightarrow J$ and moreover if their components h_A and k_A are T'-equivalent for any A, then they will be identified.

We say that I and J are **homotopic** when there exists a homotopy from I to J. This is an equivalence relation on interpretations of T in T'. □

Proposition 3.2. A homotopy $h\colon I \Rightarrow J$ induces a natural isomorphism $\mathcal{P}_h\colon \mathcal{P}_I \xRightarrow{\sim} \mathcal{P}_J$.

$$\begin{tikzcd}
T \arrow[r, bend left, "I"{name=U}] \arrow[r, bend right, "J"'{name=D}] \arrow[from=U, to=D, phantom, "\Downarrow h"] & T' & \rightsquigarrow & \mathcal{P}_T \arrow[r, bend left, "\mathcal{P}_I"{name=U2}] \arrow[r, bend right, "\mathcal{P}_J"'{name=D2}] \arrow[from=U2, to=D2, phantom, "\Downarrow \mathcal{P}_h"] & \mathcal{P}_{T'}
\end{tikzcd}$$

□

Definition 3.3 (Composition of interpretations). Let $I\colon T \dashrightarrow T'$ and $J\colon T' \dashrightarrow T''$ be interpretations. Define an interpretation $JI\colon T \dashrightarrow T''$ as follows:

- For each sort A in $\mathcal{L}$, we put $\partial_A^{JI} \equiv (\partial_A^I)^J$ and $\Delta_A^{JI} \equiv (\Delta_A^I)^J$.
- For each relation symbol R in $\mathcal{L}$, we put $R^{JI} \equiv (R^I)^J$.
- For each function symbol f in $\mathcal{L}$, we put $\Gamma_f^{JI} \equiv (\Gamma_f^I)^J$.

These data indeed satisfy the conditions to be an interpretation. □

Definition 3.4 (Bi-interpretability of theories). We say that two theories T and T' are **bi-interpretable** when there exist two interpretations $I\colon T \dashrightarrow T'$ and $J\colon T' \dashrightarrow T$ such that

- JI is homotopic to $I_T\colon T \dashrightarrow T$ (the identity interpretation),
- IJ is homotopic to $I_{T'}\colon T' \dashrightarrow T'$. □

We now describe our main result on bi-interpretability.

Theorem 3.5 (Categorical characterization of bi-interpretability). T and T' are bi-interpretable precisely when their classifying pretoposes are equivalent, i.e. $\mathcal{P}_T \simeq \mathcal{P}_{T'}$. □

Two proofs of this theorem are provided in [1]. One uses *two-dimensional* category theory, and we do not present its method here. The other involves the notions of *Morita extension* and *Morita span*, which are recently introduced by Barrett & Halvorson [2]. They introduced these to give a plausible definition of theoretical equivalence which generalizes definitional equivalence. The equivalence is defined by existence of a Morita span (see §3.1).

Our proof shows that these three conditions, i.e. bi-interpretability, equivalence of classifying pretoposes and existence of a Morita span, are all equivalent. In §3.2, we will briefly describe the latter proof.

3.1 Morita Extension

Let $\mathcal{L}$ and $\mathcal{L}^+$ be languages with $\mathcal{L} \subseteq \mathcal{L}^+$. We put for convenience

$$(\mathcal{L}^+ \setminus \mathcal{L})\text{-Sort} = \mathcal{L}^+\text{-Sort} \setminus \mathcal{L}\text{-Sort}, \qquad (\mathcal{L}^+ \setminus \mathcal{L})\text{-Rel} = \mathcal{L}^+\text{-Rel} \setminus \mathcal{L}\text{-Rel},$$
$$(\mathcal{L}^+ \setminus \mathcal{L})\text{-Func} = \mathcal{L}^+\text{-Func} \setminus \mathcal{L}\text{-Func}.$$

Definition 3.6 (Explicit definitions).

(1) Suppose that $R : \bar{A} \in (\mathcal{L}^+ \setminus \mathcal{L})$-Rel. An **explicit definition of R in terms of $\mathcal{L}$** is an $\mathcal{L}^+$-sentence of the form

$$\forall \boldsymbol{x}(R(\boldsymbol{x}) \leftrightarrow \varphi(\boldsymbol{x})),$$

where $\varphi(\boldsymbol{x})$ is an $\mathcal{L}$-formula in the same canonical context as R.

(2) Suppose that $f\colon \bar{A} \to B \in (\mathcal{L}^+ \setminus \mathcal{L})$-Func. An **explicit definition of f in terms of $\mathcal{L}$** is an $\mathcal{L}^+$-sentence of the form

$$\forall \boldsymbol{x} \forall y (f(\boldsymbol{x}) = y \leftrightarrow \varphi(\boldsymbol{x}, y)),$$

where $\varphi(\boldsymbol{x}, y)$ is an $\mathcal{L}$-formula in the same canonical context as f. For an arbitrary explicit definition of a function symbol, we refer to the $\mathcal{L}$-sentence

$$\forall \boldsymbol{x} \exists ! y \varphi(\boldsymbol{x}, y)$$

as the **admissibility condition** for the explicit definition. If f is actually a constant symbol, i.e. of arity zero, the above formulas have the following forms:

$$\forall y (y = c \leftrightarrow \varphi(y)),$$
$$\exists ! y \varphi(y).$$

□

Definition 3.7 (Sort definitions)**.** Suppose that $S \in (\mathcal{L}^+ \setminus \mathcal{L})$-Sort, $A \in \mathcal{L}$-Sort and $\bar{A}$ is a type in $\mathcal{L}$. In the following, s (resp. $x, \boldsymbol{x}$) is a variable (or a context) of the sort S (resp. $A, \bar{A}$).

(1) A sort definition of S as a **product sort** in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall \boldsymbol{x} \exists ! s \bigwedge_i \pi_i(s) = x_i,$$

where all $\pi_i\colon S \to A_i$ are in $(\mathcal{L}^+ \setminus \mathcal{L})$-Func.

(2) A sort definition of S as a **coproduct sort** in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall s \left[\bigvee_i \exists ! x_i (\rho_i(x_i) = s) \right] \wedge \bigwedge_{i \neq j} \forall x_i \forall x_j (\rho_i(x_i) \neq \rho_j(x_j)),$$

where all $\rho_i\colon A_i \to S$ are in $(\mathcal{L}^+ \setminus \mathcal{L})$-Func.

(3) A sort definition of S as a **subsort** in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall x \left[\varphi(x) \leftrightarrow \exists s (\iota_\varphi(s) = x) \right] \wedge \forall s \forall s' \left[\iota_\varphi(s) = \iota_\varphi(s') \to s = s' \right],$$

where $\varphi(x)$ is an $\mathcal{L}$-formula-in-context and $\iota_\varphi\colon S \to A$ is in $(\mathcal{L}^+ \setminus \mathcal{L})$-Func.

(4) A sort definition of S as a **quotient sort** in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall x \forall x' \left[\varepsilon_\Delta(x) = \varepsilon_\Delta(x') \leftrightarrow \Delta(x, x') \right] \wedge \forall s \exists x \left[\varepsilon_\Delta(x) = s \right],$$

where $\Delta(x, x')$ is an $\mathcal{L}$-formula-in-context and $\varepsilon_\Delta\colon A \to S$ is in $(\mathcal{L}^+ \setminus \mathcal{L})$-Func. The **admissibility condition** for the definition is the $\mathcal{L}$-sentence which expresses that Δ is an equivalence relation on the sort A. □

Definition 3.8 (Morita extensions, see Barrett & Halvorson [2]). Let T be an $\mathcal{L}$-theory and T^+ an $\mathcal{L}^+$-theory with $T \subseteq T^+$. We say that T^+ is a **Morita extension** of T if the following conditions hold:

- Every $S \in (\mathcal{L}^+ \setminus \mathcal{L})$-Sort is defined as either a product sort $\prod_j A_j$, a coproduct sort $\coprod_j A_j$, a subsort A_φ or a quotient sort A/Δ in terms of $\mathcal{L}$. This means that $(\mathcal{L}^+ \setminus \mathcal{L})$-Func contains the involved function symbols $\pi_i, \rho_i, \iota_\varphi, \varepsilon_\Delta$, that T^+ contains the involved sort definitions and that the involved admissibility conditions for quotient sorts are valid in T.
- Every $f \in (\mathcal{L}^+ \setminus \mathcal{L})$-Func other than the above $\pi_i, \rho_i, \iota_\varphi, \varepsilon_\Delta$ is explicitly defined in terms of $\mathcal{L}$. This means that T^+ contains the involved explicit definitions and that the involved admissibility conditions are valid in T.
- Every $R \in (\mathcal{L}^+ \setminus \mathcal{L})$-Rel is explicitly defined in terms of $\mathcal{L}$. This means that T^+ contains the involved explicit definitions.
- $T^+ \setminus T$ contains no axioms other than those mentioned above. □

Morita extension is a generalized notion of definitional extension which admits sort definitions. In contrast to eq-construction, here we are allowed to define a new sort by using only sorts (not types). On the other hand, Morita extension has some similar properties with eq-construction (e.g. it is conservative).

Morita span. We say that an $\mathcal{L}'$-theory T' is a **copy** of an $\mathcal{L}$-theory T if $\mathcal{L}'$ is obtained from $\mathcal{L}$ by renaming some (possibly no) symbols in $\mathcal{L}$ to symbols not contained in $\mathcal{L}$ and T' is obtained by replacing these symbols appearing in the axioms in T with the corresponding symbols in $\mathcal{L}'$.

Definition 3.9 (Morita span). Let T (resp. T') be an $\mathcal{L}$-theory (resp. an $\mathcal{L}'$-theory). A **Morita span** from T to T' is a family of finitely many theories $T_0, \dots, T_n, T'_0, \dots, T'_m$ (possibly $n, m = 0$) which satisfy the following conditions:

- T_i is an $\mathcal{L}_i$-theory for each $i = 0, \dots, n$, and T'_j is an $\mathcal{L}'_j$-theory for each $j = 0, \dots, m$.
- T_{i+1} is a Morita extension of T_i for each $i = 0, \dots, n-1$, and T'_{j+1} is a Morita extension of T'_j for each $j = 0, \dots, m-1$.
- T_0 is a copy of T, and T'_0 is a copy of T'.
- $\mathcal{L}_n$ and $\mathcal{L}'_m$ are identical and T_n and T'_m are **logically equivalent**, i.e. they make the same sentences valid.

$$\begin{tikzcd}[column sep=small]
 & & & \overset{\substack{\text{logically}\\ \text{equivalent}}}{T_n = T'_m} & & & \\
 & & \cdots \arrow[ur, phantom, sloped, "\subseteq"] & & \cdots \arrow[ul, phantom, sloped, "\supseteq"] & & \\
 & T_1 \arrow[ur, phantom, sloped, "\subseteq"] & & & & T'_1 \arrow[ul, phantom, sloped, "\supseteq"] & \\
T \arrow[r, leftrightarrow, "\text{copy}"'] & T_0 \arrow[u, phantom, sloped, "\subseteq"] & & & & T'_0 \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, leftrightarrow, "\text{copy}"'] & T'
\end{tikzcd}$$

□

A proof for existence of a Morita span to define an equivalence relation on theories is postponed to Corollary 3.11.

3.2 Sketch of Proof

Notice two trivial facts:

- Any copy of T is bi-interpretable with T.
- Two bi-interpretable theories have equivalent classifying pretoposes, for Proposition 3.2 ensures that a bi-interpretation gives an equivalence between the classifying pretoposes.

We now sketch a proof of Theorem 3.5. The proof proceeds by showing the following implications:

$$\begin{aligned} \text{Bi-interpretability} &\implies \text{Equivalence of classifying pretoposes} \\ &\implies \text{Existence of a Morita span} \implies \text{Bi-interpretability} \end{aligned}$$

The first implication was explained above. For the third implication, the general case reduces to a Morita extension:

Theorem 3.10. Any two theories constituting a Morita extension are bi-interpretable. As a consequence, so are any two theories which can be connected by a Morita span.

Proof. Let T^+ be a Morita extension of T. We have to get an interpretation $J\colon T^+ \to T$. For $\prod_i A_i, A_\varphi, A/\Delta \in (\mathcal{L}^+ \setminus \mathcal{L})$-Sort, their interpretations under J are defined as expected. For the coproduct case, using properness, we can also take an $\mathcal{L}^{\mathrm{eq}}$-formula defining a desired coproduct in $\mathcal{P}_T$. Therefore, we have obtained a pre-interpretation J. This is indeed an interpretation and, together with the canonical interpretation $I\colon T \to T^+$, constitutes a bi-interpretation. ■

The second implication needs substantial works. To obtain a desired Morita span from T to T', we construct sequences of Morita extensions

$$T \subseteq T_1 \subseteq \cdots \subseteq T_5 \quad \text{and} \quad T' \subseteq T'_1 \subseteq \cdots \subseteq T'_5,$$

where T_i (resp T'_j) is an $\mathcal{L}_i$-theory (resp. an $\mathcal{L}'_j$-theory) and the following conditions hold:

- There is a bijection between $\mathcal{L}_5$ and $\mathcal{L}'_5$.
- If we identify $\mathcal{L}_5$ with $\mathcal{L}'_5$ by using the bijection, T_5 and T'_5 are logically equivalent.

The essential idea is as follows: if we have a categorical equivalence $\mathcal{P}_T \simeq \mathcal{P}_{T'}$ and skeletons $\mathcal{S}$, $\mathcal{S}'$ (i.e. complete systems of representatives of isomorphism classes of objects) of $\mathcal{P}_T$, $\mathcal{P}_{T'}$ respectively, then a bijection between $\mathcal{S}$ and $\mathcal{S}'$ is induced. We add to T (resp. T') some product sorts, subsorts and quotient sorts step by step such that, for each formula ψ in $\mathcal{S}$ (resp. in $\mathcal{S}'$), there exists a sort in $\mathcal{L}_3$ (resp. in $\mathcal{L}'_3$) representing ψ. Slightly modifying T_3, T'_3, we obtain T_4, T'_4 with $\mathcal{L}_4$-Sort $\overset{\sim}{\to}$ $\mathcal{L}'_4$-Sort. At last we add to T_4, T'_4 explicit definitions of relations and functions such that the above bijections extends to $\mathcal{L}_5 \overset{\sim}{\to} \mathcal{L}'_5$.

These constructions indeed give logically equivalent T_5, T'_5 while we suppressed lots of technical details. Once we have proved Theorem 3.5, we also fulfill a promise made at the end of §3.1.

Corollary 3.11. Existence of a Morita span defines an equivalence relation on theories. □

Since (λ-)stability is invariant under bi-interpretability, we have the following somewhat new definition:

Definition 3.12 (Stability of pretoposes). A pretopos $\mathcal{P}$ is (λ-)**stable** when it is (equivalent to) the classifying pretopos of some (λ-)stable theory. □

It is known that completeness of a theory can be described categorically via the notion of *two-valued pretopos*. On the other hand, we do not know how stability of a pretopos can be described in the language of pretoposes. If we succeed in characterizing stability categorically, this will suggest more extensive uses of category theory in modern model theory. For example, we expect that 2-categorical constructions for pretoposes will give new model-theoretic constructions for good theories. So we will pursue this direction.

References

[1] H. Aratake. "Bicategory of Theories and Interpretations". In preparation.

[2] T. W. Barrett and H. Halvorson. "Morita Equivalence". In: **The Review of Symbolic Logic** 9.3 (2016), pp. 556–582. arXiv: `1506.04675 [math.LO]`.

[3] V. Harnik. "Model Theory vs. Categorical Logic: Two Approaches to Pretopos Completion (a.k.a. T^{eq})". In: **Models, Logics, and Higher-Dimensional Categories: A Tribute to the Work of Mihály Makkai**. Ed. by B. Hart et al. CRM Proceedings & Lecture Notes 53. American Mathematical Society, 2011, pp. 79–106.

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

[5] P. T. Johnstone. **Sketches of an Elephant: A Topos Theory Compendium**. Vol. 1. Oxford Logic Guides 43. Clarendon Press, 2002.

[6] P. T. Johnstone. **Sketches of an Elephant: A Topos Theory Compendium**. Vol. 2. Oxford Logic Guides 44. Clarendon Press, 2002.

[7] M. Makkai. "Duality and Definability in First Order Logic". In: **Memoirs of the American Mathematical Society 105**.503 (1993). x+106 pp.

[8] M. Makkai and G. E. Reyes. **First Order Categorical Logic: Model-Theoretical Methods in the Theory of Topoi and Related Categories**. Lecture Notes in Mathematics 611. Springer-Verlag, 1977.