– Postprint –

# On Generalization of Definitional Equivalence to Non-Disjoint Languages

Koen Lefever Gergely Székely

October 20, 2018

**Abstract**

For simplicity, most of the literature introduces the concept of definitional equivalence only for disjoint languages. In a recent paper, Barrett and Halvorson introduce a straightforward generalization to non-disjoint languages and they show that their generalization is not equivalent to intertranslatability in general. In this paper, we show that their generalization is not transitive and hence it is not an equivalence relation. Then we introduce another formalization of definitional equivalence due to Andréka and Németi which is equivalent to the Barrett–Halvorson generalization in the case of disjoint languages. We show that the Andréka–Németi generalization is the smallest equivalence relation containing the Barrett–Halvorson generalization and it is equivalent to intertranslatability, which is another definition for definitional equivalence, even for non-disjoint languages. Finally, we investigate which definitions for definitional equivalences remain equivalent when we generalize them for theories in non-disjoint languages.

**Keywords**: First-Order Logic · Definability Theory · Definitional Equivalence · Logical Translation · Logical Interpretation

## 1 Introduction

In mathematics and philosophy of science, there exist several approaches to answer the question "when are two theories, which at first may seem different or even contradictory, the same?" A first and straightforward answer is that two theories are the same if they are logically equivalent, which means that the consequences of their axioms are identical. A classic example of this is Euclidean geometry using Euclids's fifth postulate[1] and Euclidean geometry using Playfair's axiom.[2] The postulate by Euclid and the axiom of Playfair are by

[1] See (Heath 1956, Vol. 1 p. 190).

[2] See (Playfair 1846, p. 29).

themselves not logically equivalent, but in combination with the other axioms of Euclidean geometry, their consequences are the same.

In this paper, we define theories by their set of axioms or postulates. If we however would, as some authors do, assume that theories are closed under consequence, then logical equivalence becomes identity. Logical equivalence may however be too strict, because theories can be expressed in different languages. To remedy this, other equivalences between theories have been proposed, such as definitional equivalence,[3] many-sorted definitional equivalence[4] and categorical equivalence. For a comparison in terms of category theory between definitional equivalence, bi-interpretability, mutual interpretability and sentential equivalence, we refer to (Visser 2006).

In the current paper, we discuss definitional equivalence,[5] and the different ways to define it. A well-known classic example of definitional equivalence is the equivalence between the theory of Boolean Algebras in the language $\wedge$, $\vee$, $\neg$, $0$, $1$ and the theory of Complemented Bounded Distributive Lattices in the language $\leq$. Since those theories use different languages, they cannot prove each other's theorems. However, $\wedge$, $\vee$, $\neg$, $0$, $1$ can be defined in terms of $\leq$, and the other way round $\leq$ can be defined in the language of Boolean Algebras; and in this common language both theories can prove the same theorems.

The idea behind definitional equivalence is that extending a theory by definitions (i.e., introducing formula abbreviations) does not change the theory, but only its presentation (i.e., its language). With this intuition it is natural to consider theories having a common definitional extensions as one theory presented differently. Following Andréka and Németi's Definition 4.2 from (Andréka and Németi 2014), we are going to introduce definitional equivalence as the smallest equivalence relation containing definitional extension.[6]

Even though the intuition that we cannot really change a theory by adding definitions is clear, one may be surprised when learning that theories contradicting each other can be definitionally equivalent. As an illustration and explanation why that is a desired feature, consider the following example: suppose one author would rewrite the theory of subatomic particles, but consistently

[3] To avoid confusion, we call the non-transitive variant of definitional equivalence "definitional mergeability"; see Definition 6.

[4] In (Andréka et al. 2002, Section 6.3), (Madarász 2002, Section 4.3) and (Andréka et al. 2008, Definition 2.4.1), definitional equivalence is generalized to many-sorted definability, where even new entities can be defined and not just new relations between existing entities. In (Barrett and Halvorson 2016*b*), where the corresponding many-sorted definitional mergeability is called "Morita equivalence", the relation between logical equivalence, definitional mergeability, many-sorted definitional mergeability and categorical equivalence is discussed.

[5] Definitional equivalence has also been called *logical synonymity* or *synonymy*, e.g., in (de Bouvère 1965*a*), (Visser 2006), (Friedman and Visser 2014) and (Visser 2015).

[6] More precisely, definitional equivalence is the symmetric transitive closure of definitional extension, see Definitions 4 and 5.

switch the words electron and proton. The "new" theory would contradict standard subatomic theory in many ways, using the criterion of logical equivalence these two theories would not be the same. However, the differences between these two theories would be only terminological, and physicists would soon discover that by a simple translation of terms they could use the results of this "new" theory in the same way as the standard theory despite the apparent contradictions.

The concept of definitional equivalence was first introduced by Montague in (Montague 1956), but there are already some traces of the idea in Quine's (Quine 1946) and in (Tarski et al. 1953) by Tarski, Mostowski and Robinson. Early development of this concept was done in de Bouvère's (de Bouvère 1965*a*) and (de Bouvère 1965*b*). In philosophy of science, it was introduced by Glymour in (Glymour 1970), (Glymour 1977) and (Glymour 1980). See Corcoran's (Corcoran 1980) for notes on the history of definitional equivalence. Definitional equivalence preserves (in the case of finite signatures) important properties such as finite axiomatizability, decidability and categoricity; see (Pinter 1978) and (Visser 2006). Barrett and Halvorson's (Barrett and Halvorson 2016*a*), on which the present paper is partly a commentary, contains more references to examples on the use of definitional equivalence in the context of philosophy of science.

We have recently started in (Lefever and Székely 2018) to use definitional equivalence to study the exact differences and similarities between theories which are *not* equivalent, in that case classical and relativistic kinematics. In that paper, we showed that there exists an interpretation of relativistic kinematics in classical kinematics, but not the other way round. We also showed that special relativity extended with a "primitive ether" is definitionally equivalent to classical kinematics. Those theories are expressed in the same language, and hence have non-disjoint languages.

When studying competing theories,[7] such as classical mechanics versus relativity theories, classical thermodynamics versus statistical thermodynamics, or the phlogiston theory versus the compound theory in chemistry,[8] the main criterion to decide on which theory to chose is empirical adequacy: one theory is better than another if it accounts for more of the data or phenomena than the other. However, this is not always straightforward, since one theory may be better suited to explain one part of the data or phenomena, and another theory may be better suited to explain another part. A classic example of this is the transition from the Ptolemaic geocentric theory to the Copernican heliocentric theory,

[7] For a further discussion of comparing competing theories by introducing conceptual distances based on definitional equivalence, see (Friend et al. 2018).

[8] See (Chang 2012) for a discussion of competing theories from chemistry in philosophy of science.

where the Ptolemaic theory initially provided in many cases more accurate predictions than the Copernican theory.[9] This illustrates that empirical adequacy as a criterion contains some degree of arbitrary. Therefore, we also need other methods to study and compare theories than just empirical adequacy, such as those discussed here. In the case of competing theories, an overlap of the languages can be expected. In general, it is easy to get rid of the overlap between languages by using a disjoint renaming as introduced below in Definition 14 and as illustrated in (Lefever 2017). This however makes the notation heavier, less natural and less intuitive. In some cases the overlap between the languages may not as easily be avoided or may even be systematically important, for example in the case of Fujimoto interpretability where a sub-vocabulary is kept constant, see (Fujimoto 2010).

In Barrett and Halvorson's (Barrett and Halvorson 2016*a*, Definition 2) and (Barrett and Halvorson 2016*b*, Definition in §3.2 on p. 561) definitional equivalence from Hodges' (Hodges 1993, pp. 60-61) is generalized for non-disjoint languages in a straightforward way. They show that their generalization, which we call here *definitional mergeability* to avoid ambiguity, is not equivalent to intertranslatability in general but only for theories in disjointly formulated languages. In this paper, we show in Theorem 1 below that definitional mergeability is not an equivalence relation because it is not transitive.[10] We recall Andréka and Németi's Definition 4.2 from (Andréka and Németi 2014) which is known to be equivalent to definitional mergeability for theories formulated in disjoint languages. Then we show that the Andréka–Németi definitional equivalence is the smallest equivalence relation containing definitional mergeability and that it is equivalent to intertranslatability even for theories formulated in non-disjoint languages. Actually, two theories are definitionally equivalent iff there is a theory that is definitionally mergeable to both of them. Moreover, one of these definitional mergers can be a renaming; see Theorem 4.

Theorem 4.2 of (Andréka and Németi 2014) claims without proof that (i) definitional equivalence, (ii) definitional mergeability, (iii) intertranslatability and (iv) model mergeability (see Definition 7 below) are equivalent in case of disjoint languages. Here, we show that the equivalence of (i) and (iii) and that of (ii) and (iv) hold for arbitrary languages, see Theorems 8 and 7. However, since (i) and (ii) are not equivalent by Theorems 1 and 3, no other equivalence of extends to arbitrary languages. Finally, we introduce a modification of (iv) that is equivalent to (i) and (iii) for arbitrary languages, see Theorem 9.

[9] See Kuhn's (Kuhn 1957).

[10] The non-transitivity of definitional mergeability in the case when new (sorts of) entities can also be defined is already noted in (Andréka et al. 2008, §2.4 p. 55).

## 2 Framework and definitions

For every theory $T$ which might contain constants and functions, there is another theory $T'$ which is formulated in a languages containing only predicates (relation symbols) and connected to $T$ by the two central relations $\overset{\triangle}{\equiv}$ and $\nearrow\!\nwarrow$ investigated in this paper, see (Barrett and Halvorson 2016*a*, Proposition 2 and Theorem 1). Therefore, here we only consider languages containing only predicates.

We use the following set of basic logical symbols $\mathsf{Log} = \{\,\exists, \wedge, \neg, (, ), =\,\}$ and assume that there is a countable set $\mathsf{Var} = \{\,\mathsf{v}_1, \mathsf{v}_2 \ldots, \mathsf{v}_i, \ldots\,\}$ of variables in a fixed order. A ***signature***[11] *of language* $\mathcal{L}$ is a pair $\langle \mathsf{Pred}_{\mathcal{L}}, \mathsf{ar}_{\mathcal{L}} \rangle$ of the set $\mathsf{Pred}_{\mathcal{L}}$ of ***predicates***[12] (relation symbols) and the ***arity function*** $\mathsf{ar}_{\mathcal{L}}$ which assigns an arity[13] to elements of $\mathsf{Pred}_{\mathcal{L}}$. ***Formulas*** *of language* $\mathcal{L}$ are built up recursively from alphabet $\mathsf{Pred}_{\mathcal{L}} \cup \mathsf{Log} \cup \mathsf{Var}$ in the usual way and their set is denoted by $\mathsf{Form}_{\mathcal{L}}$. A ***theory*** $T$ *of language* $\mathcal{L}$ is a set of formulas of $\mathcal{L}$, that is, $T \subseteq \mathsf{Form}_{\mathcal{L}}$.

**Convention 1.** Whenever we talk about a theory, we always assume that there is a fixed language corresponding to that theory. The languages corresponding to theories $T_i$, $T'$, etc. are respectively denoted by $\mathcal{L}_i$, $\mathcal{L}'$, etc.

A ***model*** $\mathfrak{M} = \langle M, \langle p^{\mathfrak{M}} : p \in \mathit{Pred}_{\mathcal{L}} \rangle\rangle$ *of language* $\mathcal{L}$ consists of a non-empty underlying set $M$, and for every predicate $p$ of $\mathcal{L}$, a relation $p^{\mathfrak{M}} \subseteq M^n$ with the arity $\mathsf{ar}_{\mathcal{L}}(p) = n$.[14] $\mathsf{Mod}(T)$ is *the* ***class of models*** *of theory* $T$,

$$\mathsf{Mod}(T) \overset{\text{def}}{=} \{\mathfrak{M} : \mathfrak{M} \models T\}.$$

**Definition 1.** Two theories $T_1$ *and* $T_2$ *are* ***logically equivalent***, in symbols

$$T_1 \equiv T_2,$$

iff[15] they have the same class of models, i.e., $\mathsf{Mod}(T_1) = \mathsf{Mod}(T_2)$.

**Remark 1.** It is worth noting that, by Gödel's Completeness Theorem, $\equiv$ extensionally corresponds to interderivability, i.e., $T \equiv T'$ iff $T \vdash \varphi'$ for all $\varphi' \in T'$ and $T' \vdash \varphi$ for all $\varphi \in T$.

**Convention 2.** Instead of using meta-variables, we refer to arbitrary elements of $\mathsf{Var}$ by using indexes. When we would like to talk about $n$-many arbitrary variables from $\mathsf{Var}$, we use double indexes $i_1, \ldots, i_n$. Sometimes the list of variables $\mathsf{v}_{i_1}, \ldots, \mathsf{v}_{i_n}$ is abbreviated to $\bar{v}$ and quantifiers $\forall \mathsf{v}_{i_1}, \ldots, \forall \mathsf{v}_{i_n}$ to $\forall \bar{v}$.

[11] A *signature* is also called a *vocabulary*.

[12] Note that we allow $\mathsf{Pred}_{\mathcal{L}}$ to be infinite.

[13] The *arity* is the number of variables in the relation, it is also called the *rank*, *degree*, *adicity* or *valency* of the relation.

[14] The non-empty underlying set $M$ is also called the *universe*, the *carrier* or the *domain* of $\mathfrak{M}$. $M^n$ denotes the Cartesian power of set $M$.

[15] *iff* abbreviates *if and only if*. It is denoted by $\leftrightarrow$ in the object languages (see remark 2 below) and by $\Leftrightarrow$ in the meta-language.

**Definition 2.** ***Semantics***: let $\mathfrak{M}$ be a model, let $M$ be the non-empty underlying set of $\mathfrak{M}$, let $\varphi$ be a formula and let $e : \mathsf{Var} \to M$ be an evaluation of variables, then we inductively define that $e$ *satisfies* $\varphi$ *in* $\mathfrak{M}$, in symbols

$$\mathfrak{M} \models \varphi[e],$$

as:

1. For predicate $p$, $\mathfrak{M} \models p(\mathsf{v}_{i_1}, \mathsf{v}_{i_2}, \ldots, \mathsf{v}_{i_n})[e]$ holds if
$$\big(e(\mathsf{v}_{i_1}), e(\mathsf{v}_{i_2}), \ldots, e(\mathsf{v}_{i_n})\big) \in p^{\mathfrak{M}},$$
2. $\mathfrak{M} \models (\mathsf{v}_i = \mathsf{v}_j)[e]$ holds if $e(\mathsf{v}_i) = e(\mathsf{v}_j)$ holds,
3. $\mathfrak{M} \models \neg\,\varphi[e]$ holds if $\mathfrak{M} \models \varphi[e]$ does not hold,
4. $\mathfrak{M} \models (\psi \wedge \theta)[e]$ holds if both $\mathfrak{M} \models \psi[e]$ and $\mathfrak{M} \models \theta[e]$ hold,
5. $\mathfrak{M} \models \big(\exists\,\mathsf{v}_j\,\psi\big)[e]$ holds if there is an element $b \in M$, such that $\mathfrak{M} \models \psi[e']$ if $e'(\mathsf{v}_j) = b$ and $e'(\mathsf{v}_i) = e(\mathsf{v}_i)$ if $i \neq j$.

Let $\mathsf{v}_{i_1}, \mathsf{v}_{i_2}, \ldots, \mathsf{v}_{i_n}$ be the list of all free variables of $\varphi$ in the order of their first occurrence in $\varphi$ and let $\bar{a}$ be a list $a_1, a_2, \ldots, a_n$ of elements of $M$. Then $\mathfrak{M} \models \varphi[\bar{a}]$ iff $\mathfrak{M}$ satisfies[16] $\varphi$ for all (or equivalently some) evaluation $e$ of variables for which $e(\mathsf{v}_{i_j}) = a_j$ for all $j \in \{1, 2, \ldots, n\}$. That $\varphi$ is true in $\mathfrak{M}$ for all evaluations of variables is denoted by $\mathfrak{M} \models \varphi$. For theory $T$, $\mathfrak{M} \models T$ abbreviates that $\mathfrak{M} \models \varphi$ for all $\varphi \in T$.

**Remark 2.** We use $\varphi \vee \psi$ as an abbreviation for $\neg\,(\neg\,\varphi \wedge \neg\,\psi)$, $\varphi \to \psi$ for $\neg\,\varphi \vee \psi$, $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$ and $\forall \mathsf{v}_i\,\varphi$ for $\neg\,\exists\,\mathsf{v}_i\,\neg\,\phi$.

**Definition 3.** Let $\mathcal{L}$ and $\mathcal{L}^+$ be two languages such that $\mathsf{Form}_{\mathcal{L}} \subset \mathsf{Form}_{\mathcal{L}^+}$. An ***explicit definition*** *of an $n$-ary predicate $p \in \mathit{Pred}_{\mathcal{L}^+} \setminus \mathit{Pred}_{\mathcal{L}}$ in terms of $\mathcal{L}^+$* is a formula of the form

$$\forall \bar{v}\big(p(\bar{v}) \leftrightarrow \varphi(\bar{v})\big),$$

where $\varphi \in \mathsf{Form}_{\mathcal{L}}$.[17]

**Definition 4.** A ***definitional extension***[18] *of a theory $T$ of language $\mathcal{L}$ to language $\mathcal{L}^+$* is a theory $T^+ \equiv T \cup \Delta$, where $\Delta$ is a set of explicit definitions in terms of $\mathcal{L}$ for each predicate $p \in \mathsf{Pred}_{\mathcal{L}^+} \setminus \mathsf{Pred}_{\mathcal{L}}$. In this paper,

$$T \nearrow T^+ \text{ and } T^+ \nwarrow T$$

denote that $T^+$ is a definitional extension of $T$.

[16] $\mathfrak{M} \models \varphi[\bar{a}]$ can also be read as $\varphi[\bar{a}]$ *being true in* $\mathfrak{M}$.

[17] Here $\bar{v}$ gives the variables of $p(\bar{v})$ in the order of occurrence and the free variables of $\varphi$ are among the variables of $\bar{v}$ but not necessarily all of $\bar{v}$.

[18] We follow the definition from (Andréka and Németi 2014, Section 4.1, p.36), (Hodges 1993, p.60) and (Hodges 1997, p.53). In (Barrett and Halvorson 2016*a*, Section 3.1), the logical equivalence relation is not part of the definition.

We use $\Delta_{ij}$ to denote the set of explicit definitions when the language $\mathcal{L}_j$ of theory $T_j$ is defined in terms of the language $\mathcal{L}_i$ of theory $T_i$.

**Definition 5.** Two theories $T, T'$ are ***definitionally equivalent***, in symbols

$$T \overset{\triangle}{\equiv} T',$$

iff there is a chain $T_1, \ldots, T_n$ of theories such that $T = T_1$, $T' = T_n$, and for all $1 \leq i < n$ either $T_i \nearrow T_{i+1}$ or $T_i \nwarrow T_{i+1}$.

**Remark 3.** If a theory is consistent, then all theories which are definitionally equivalent to that theory are also consistent since definitions cannot make consistent theories inconsistent. Similarly, if a theory is inconsistent, then all theories which are definitionally equivalent to that theory are also inconsistent.

**Definition 6.** Let $T_1$ and $T_2$ be two arbitrary theories. $T_1$ and $T_2$ are ***definitionally mergeable***, in symbols

$$T_1 \nearrow\!\!\nwarrow T_2,$$

iff there is a theory $T^+$ which is a common definitional extension of $T_1$ and $T_2$, i.e., $T_1 \nearrow T^+ \nwarrow T_2$.

**Remark 4.** From Definitions 5 and 6, it is immediately clear that being definitionally mergeable is a special case of being definitionally equivalent.

Lemma 1 below establishes that our Definition 6 of definitional mergeability is equivalent to the definition for definitional equivalence in (Barrett and Halvorson 2016*a*, Definition 2).

**Lemma 1.** Let $T_1$ and $T_2$ be two arbitrary theories. Then $T_1 \nearrow\!\!\nwarrow T_2$ iff there are sets of explicit definitions $\Delta_{12}$ and $\Delta_{21}$ such that $T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}$.

*Proof.* Let $T_1 \nearrow\!\!\nwarrow T_2$, then there exists a $T^+$ such that $T_1 \nearrow T^+ \nwarrow T_2$. By the definition of definitional extension, there exist sets of explicit definitions $\Delta_{12}$ and $\Delta_{21}$ such that $T_1 \cup \Delta_{12} \equiv T^+$ and $T_2 \cup \Delta_{21} \equiv T^+$, and hence by transitivity of logical equivalence $T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}$.

To prove the other direction: let $T_1$ and $T_2$ be theories such that $T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}$ for some sets $\Delta_{12}$ and $\Delta_{21}$ of explicit definitions. Let $T^+ = T_1 \cup T_2 \cup \Delta_{12} \cup \Delta_{21}$. Hence $T_1 \cup \Delta_{12} \equiv T^+ \equiv T_2 \cup \Delta_{21}$ and $T_1 \nearrow T^+ \nwarrow T_2$, and therefore $T_1 \nearrow\!\!\nwarrow T_2$. □

**Convention 3.** If theories $T_1$ and $T_2$ are definitionally mergeable and their languages are disjoint, i.e., $\mathsf{Pred}_{\mathcal{L}_1} \cap \mathsf{Pred}_{\mathcal{L}_2} = \emptyset$, we write

$$T_1 \overset{\emptyset}{\nearrow\!\!\nwarrow} T_2.$$

**Definition 7.** Theories $T_1$ and $T_2$ are ***model mergeable***,[19] in symbols

$$\mathsf{Mod}(T_1) \nearrow\!\nwarrow \mathsf{Mod}(T_2),$$

iff there is a bijection $\beta$ between $\mathsf{Mod}(T_1)$ and $\mathsf{Mod}(T_2)$ that is defined along two sets $\Delta_{12}$ and $\Delta_{21}$ of explicit definitions such that if $\mathfrak{M} \in \mathsf{Mod}(T_1)$, then

- the underlying sets of $\mathfrak{M}$ and $\beta(\mathfrak{M})$ are the same,
- the relations in $\beta(\mathfrak{M})$ are the ones defined in $\mathfrak{M}$ according to $\Delta_{12}$ and vice versa, the relations in $\mathfrak{M}$ are the ones defined in $\beta(\mathfrak{M})$ according to $\Delta_{21}$.

**Definition 8.** Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two arbitrary languages. A ***translation***

$$\mathsf{tr} : \mathsf{Form}_{\mathcal{L}_1} \to \mathsf{Form}_{\mathcal{L}_2}$$

is a map from formulas of $\mathcal{L}_1$ to that of $\mathcal{L}_2$ which

- for all $p \in \mathsf{Pred}_{\mathcal{L}_1}$, maps every atomic formula $p(\mathsf{v}_1, \mathsf{v}_2, \dots, \mathsf{v}_n)$ to a corresponding formula $\varphi_p(\mathsf{v}_1, \mathsf{v}_2, \dots, \mathsf{v}_n)$ and maps $p(\mathsf{v}_{i_1}, \mathsf{v}_{i_2}, \dots, \mathsf{v}_{i_n})$ to the appropriately substituted version[20] of $\varphi_p$.
- preserves the equality, logical connectives, and quantifiers, i.e.,
  - $\mathsf{tr}(\mathsf{v}_i = \mathsf{v}_j)$ is $\mathsf{v}_i = \mathsf{v}_j$,
  - $\mathsf{tr}(\neg\,\varphi)$ is $\neg\,\mathsf{tr}(\varphi)$,
  - $\mathsf{tr}(\varphi \land \psi)$ is $\mathsf{tr}(\varphi) \land \mathsf{tr}(\psi)$, and
  - $\mathsf{tr}(\exists\,\mathsf{v}_i\,\varphi)$ is $\exists\,\mathsf{v}_i\,\mathsf{tr}(\varphi)$.

**Definition 9.** Let $T_1$ and $T_2$ be theories. An ***interpretation*** *$tr_{12}$ of theory $T_1$ in theory $T_2$* is a translation that maps consequences of $T_1$ into consequences of $T_2$, i.e., $T_1 \models \varphi$ implies $T_2 \models \mathsf{tr}_{12}(\varphi)$ for all $\varphi \in \mathsf{Form}_{\mathcal{L}_1}$.

It is easy to see that mutual interpretability does not imply definitional equivalence, see e.g., (Visser 2006, §4.8.4) for some simple examples, which also show that mutual interpretability does not even preserves decidability. Mutual definability, when no models gets lost during the interpretation, does not imply definitional equivalence either, see (Andréka et al. 2005). However, requiring the equivalence of any formula and its the back and forth translation turns mutual interpretability into a natural equivalent formulation of definitional equivalence:

[19] We use the definition from Andréka and Németi in (Andréka and Németi 2014, p. 40, item iv), which is a variant of the definition by Henkin, Monk, and Tarski in (Henkin et al. 1971, p. 56, Remark 0.1.6).

[20] An appropriate built in mechanism of substitution is needed to avoid "collision of variables", see e.g., (Andréka and Németi 2014, p.37, footnote 34).

**Definition 10.** Theories $T_1$ and $T_2$ are ***intertranslatable***,[21] in symbols

$$T_1 \rightleftarrows T_2,$$

iff there are interpretations $\mathsf{tr}_{12}$ of $T_1$ in $T_2$ and $\mathsf{tr}_{21}$ of $T_2$ in $T_1$ such that

- $T_1 \models \forall \bar{v}\big(\varphi(\bar{v}) \leftrightarrow \mathsf{tr}_{21}\big(\mathsf{tr}_{12}(\varphi(\bar{v}))\big)\big)$
- $T_2 \models \forall \bar{v}\big(\psi(\bar{v}) \leftrightarrow \mathsf{tr}_{12}\big(\mathsf{tr}_{21}(\psi(\bar{v}))\big)\big)$

for every formulas $\varphi(\bar{v})$ and formula $\psi(\bar{v})$ of languages $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively.

For a direct proof that intertranslatability is an equivalence relation, see e.g., (Lefever 2017, Theorem 1, p. 7). This fact also follows from Theorems 3 and 8 below.

**Definition 11.** The *relation defined by formula* $\varphi$ in model $\mathfrak{M}$ is:[22]

$$\|\varphi\|^{\mathfrak{M}} \stackrel{\text{def}}{=} \big\{\bar{a} \in M^n : \mathfrak{M} \models \varphi[\bar{a}]\big\}.$$

**Definition 12.** For every translation $\mathsf{tr}_{12} : \mathsf{Form}_{\mathcal{L}_1} \to \mathsf{Form}_{\mathcal{L}_2}$, let $\mathsf{tr}^*_{12}$ be the map that maps model $\mathfrak{M} = \langle M, \ldots \rangle$ of $\mathcal{L}_2$ to

$$\mathsf{tr}^*_{12}(\mathfrak{M}) \stackrel{\text{def}}{=} \Big\langle M, \big\langle \|\mathsf{tr}_{12}(p_i)\|^{\mathfrak{M}} : p_i \in \mathsf{Pred}_{\mathcal{L}_1} \big\rangle \Big\rangle,$$

that is all predicates $p_i$ of $\mathcal{L}_1$ interpreted in model $\mathsf{tr}^*_{12}(\mathfrak{M})$ as the relation defined by formula $\mathsf{tr}_{12}(p_i)$.

**Lemma 2.** Let $\mathfrak{M}$ be a model of language $\mathcal{L}_2$, let $\varphi$ be a formula of language $\mathcal{L}_1$, and let $e : \mathsf{Var} \to M$ be an evaluation of variables. If $\mathsf{tr}_{12} : \mathsf{Form}_{\mathcal{L}_1} \to \mathsf{Form}_{\mathcal{L}_2}$ is a translation, then

$$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \varphi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\varphi)[e].$$

*Proof.* We are going to prove Lemma 2 by induction on the complexity of $\varphi$. So let us first assume that $\varphi$ is a single predicate $p$ of language $\mathcal{L}_1$.

Let $\bar{u}$ be the $e$-image of the free variables of $p$. Then $\mathsf{tr}^*_{12}(\mathfrak{M}) \models p[e]$ holds exactly if $\mathsf{tr}^*_{12}(\mathfrak{M}) \models p[\bar{u}]$. By Definition 12, this holds iff

$$\Big\langle M, \big\langle \|\mathsf{tr}_{12}(p_i)\|^{\mathfrak{M}} : p_i \in \mathsf{Pred}_{\mathcal{L}_1} \big\rangle \Big\rangle \models p[\bar{u}]. \tag{1}$$

[21] In Henkin, Monk, and Tarski's (Henkin et al. 1985, p. 167, Definition 4.3.42), definitional equivalence is defined as intertranslatability. It can be argued that this is the more fundamental definition of definitional equivalence, because all known interpretation-based notions of equivalence of theories can be put in a uniform format; see (Visser 2006).

[22] $\|\varphi\|^{\mathfrak{M}}$ is basically the same as the *meaning* of formula $\varphi$ in model $\mathfrak{M}$, see (Andréka et al. 2001, p. 194 Definition 34 and p. 231 Example 8).

By Definition 11, $\|\mathsf{tr}_{12}(p)\|^{\mathfrak{M}} = \big\{\bar{a} \in M^n : \mathfrak{M} \models \mathsf{tr}_{12}(p)[\bar{a}]\big\}$. So (1) is equivalent to $\mathfrak{M} \models \mathsf{tr}_{12}(p)[\bar{u}]$.

If $\varphi$ is $\mathsf{v}_i = \mathsf{v}_j$, then we should show that

$$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \mathsf{v}_i = \mathsf{v}_j[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\mathsf{v}_i = \mathsf{v}_j)[e].$$

Since translations preserve mathematical equality by Definition 8, this is equivalent to

$$\mathsf{tr}^*_{12}(\mathfrak{M}) \models (\mathsf{v}_i = \mathsf{v}_j)[e] \Leftrightarrow \mathfrak{M} \models (\mathsf{v}_i = \mathsf{v}_j)[e],$$

which holds because the underlying sets of $\mathsf{tr}^*_{12}(\mathfrak{M})$ and $\mathfrak{M}$ are the same and both sides of the equivalence are equivalent to $e(\mathsf{v}_i) = e(\mathsf{v}_j)$ by Definition 2.

Let us now prove the more complex cases by induction on the complexity of formulas.

- If $\varphi$ is $\neg\,\psi$, then we should show that

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \neg\,\psi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\neg\,\psi)[e].$$

  Since $\mathsf{tr}_{12}$ is a translation, it preserves (by Definition 8) the conectives, and therefore this is equivalent to

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \neg\,\psi[e] \Leftrightarrow \mathfrak{M} \models \neg\,\mathsf{tr}_{12}(\psi)[e],$$

  which holds by Definition 2 Item 3 since we have

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \psi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\psi)[e]$$

  by induction.

- If $\varphi$ is $(\psi \wedge \theta)$, then we should show that

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models (\psi \wedge \theta)[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\psi \wedge \theta)[e].$$

  Since $\mathsf{tr}_{12}$ is a translation, it preserves (by Definition 8) the conectives, and therefore $\mathsf{tr}_{12}(\psi \wedge \theta)$ is equivalent to $\mathsf{tr}_{12}(\psi) \wedge \mathsf{tr}_{12}(\theta)$, and hence the above is equivalent to

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models (\psi \wedge \theta)[e] \Leftrightarrow \mathfrak{M} \models \big(\mathsf{tr}_{12}(\psi) \wedge \mathsf{tr}_{12}(\theta)\big)[e],$$

  which holds by Definition 2 Item 4 because both

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \psi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\psi)[e]$$

  and

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \theta[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\theta)[e]$$

  hold by induction.

- If $\varphi$ is $\exists\, \mathsf{v}_j\, \psi$, then we should show that

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \big(\exists\, \mathsf{v}_j\, \psi\big)[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}\big(\exists\, \mathsf{v}_j\, \psi\big)[e]$$

  holds. Since $\mathsf{tr}_{12}$ is a translation, it preserves (by Definition 8) the quantifiers, and hence this is equivalent to

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \big(\exists\, \mathsf{v}_j\, \psi\big)[e] \Leftrightarrow \mathfrak{M} \models \big(\exists\, \mathsf{v}_j\, \mathsf{tr}_{12}(\psi)\big)[e].$$

  By Definition 2 Item 5, both sides of he equivalence hold exactly if there exists an element $b \in M$ such that

  $$\mathsf{tr}^*_{12}(\mathfrak{M}) \models \psi[e'] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{12}(\psi)[e'],$$

  where $e'(\mathsf{v}_j) = b$ and $e'(\mathsf{v}_i) = e(\mathsf{v}_i)$ if $\mathsf{v}_i \neq \mathsf{v}_j$, which holds by induction because the underlying sets of $\mathsf{tr}^*_{12}(\mathfrak{M})$ and $\mathfrak{M}$ are the same. □

**Corollary 1.** Let $\mathsf{tr}_{12} : \mathsf{Form}_{\mathcal{L}_1} \to \mathsf{Form}_{\mathcal{L}_2}$ be a translation. Then the following two statements are equivalent:

- $\mathsf{tr}_{12}$ is an interpretation[23] of theory $T_1$ in theory $T_2$
- $\mathsf{tr}^*_{12}$ maps all models of $T_2$ to models of $T_1$, i.e., $\mathsf{tr}^*_{12} : \mathsf{Mod}(T_2) \to \mathsf{Mod}(T_1)$.

*Proof.* Assume first that $\mathsf{tr}_{12}$ is an interpretation of theory $T_1$ in theory $T_2$. Let $\mathfrak{M}$ be a model of theory $T_2$. We should prove that $\mathsf{tr}^*_{12}(\mathfrak{M})$ is a model of $T_1$, i.e., $\mathsf{tr}^*_{12}(\mathfrak{M}) \models \varphi$ for every $\varphi \in T_1$. By Lemma 2, it is enough to show that $\mathfrak{M} \models \mathsf{tr}_{12}(\varphi)$ for every $\varphi \in T_1$, which is true since $\mathsf{tr}_{12}$ is an interpretation of theory $T_1$ in theory $T_2$.

To prove the other direction, assume that $\mathsf{tr}^*_{12}$ maps all models of $T_2$ to models of $T_1$ and let $T_1 \models \varphi$. We have to prove that $T_2 \models \mathsf{tr}_{12}(\varphi)$, i.e., $\mathfrak{M} \models \mathsf{tr}_{12}(\varphi)$ for every model $\mathfrak{M}$ of $T_2$. Since $\mathsf{tr}^*_{12}$ maps all models of $T_2$ to models of $T_1$, we have $\mathsf{tr}^*_{12}(\mathfrak{M}) \in \mathsf{Mod}(T_1)$. Hence $\mathsf{tr}^*_{12}(\mathfrak{M}) \models \varphi$. Then, by Lemma 2, we have that $\mathfrak{M} \models \mathsf{tr}_{12}(\varphi)$. This completes the proof since $\mathfrak{M}$ was an arbitrary model of $T_2$. □

**Remark 5.** Note that while $\mathsf{tr}_{12}$ is an interpretation of $T_1$ in $T_2$, $\mathsf{tr}^*_{12}$ translates models the other way round from $\mathsf{Mod}(T_2)$ to $\mathsf{Mod}(T_1)$. For an example illustrating this for an interpretation from relativistic kinematics in classical kinematics, see (Lefever 2017, Chapter 7) or (Lefever and Székely 2018, Section 7).

**Definition 13.** Theories $T_1$ and $T_2$ are ***model intertranslatable***, in symbols

$$\mathsf{Mod}(T_1) \rightleftarrows \mathsf{Mod}(T_2),$$

iff there are translations $\mathsf{tr}_{12} : \mathsf{Form}_{\mathcal{L}_1} \to \mathsf{Form}_{\mathcal{L}_2}$ and $\mathsf{tr}_{21} : \mathsf{Form}_{\mathcal{L}_2} \to \mathsf{Form}_{\mathcal{L}_1}$, such that $\mathsf{tr}^*_{12} : \mathsf{Mod}(T_2) \to \mathsf{Mod}(T_1)$ and $\mathsf{tr}^*_{21} : \mathsf{Mod}(T_1) \to \mathsf{Mod}(T_2)$ are bijections which are inverses of each other.

[23] In terms of the framework of (Visser 2006), $\mathsf{tr}_{12}$ is a contravariant functor from the category of direct interpretations to the category of sets.

**Definition 14.** Theory $T$ and theory $T'$ are ***disjoint renamings*** *of each other*, in symbols

$$T \overset{\emptyset}{\simeq} T',$$

iff their languages $\mathcal{L}$ and $\mathcal{L}'$ are disjoint, i.e., $\mathsf{Pred}_{\mathcal{L}} \cap \mathsf{Pred}_{\mathcal{L}'} = \emptyset$, and there is an arity preserving bijection $R^{\emptyset}_{\mathcal{L}\mathcal{L}'} : \mathsf{Pred}_{\mathcal{L}} \to \mathsf{Pred}_{\mathcal{L}'}$, which naturally can be extended to a bijection $\bar{R}^{\emptyset}_{\mathcal{L}\mathcal{L}'} : \mathsf{Form}_{\mathcal{L}} \to \mathsf{Form}_{\mathcal{L}'}$, and $T' = \left\{ \bar{R}^{\emptyset}_{\mathcal{L}\mathcal{L}'}(\varphi) : \varphi \in T \right\}$.

**Remark 6.** Note that disjoint renaming is symmetric but neither reflexive nor transitive. Also, if $T \overset{\emptyset}{\simeq} T'$, then $T \overset{\emptyset}{\nearrow\!\nwarrow} T'$, $T \nearrow\!\nwarrow T'$, $T \overset{\triangle}{\equiv} T'$, $T \rightleftarrows T'$, and if $\mathsf{Pred}_{\mathcal{L}}$ and $\mathsf{Pred}_{\mathcal{L}'}$ are not empty, then $T \neq T'$.

## 3 Properties

**Theorem 1.** Definitional mergeability $\nearrow\!\nwarrow$ is not transitive. Hence it is not an equivalence relation.

The proof is based on (Barrett and Halvorson 2016*a*, Example 5).[24] Note that the proof relies on the languages of theories $T_1$ and $T_2$ being non-disjoint.

*Proof.* Let $p$ and $q$ be unary predicates. Consider the following theories $T_1$, $T_2$ and $T_3$ for which $\mathsf{Pred}_{\mathcal{L}_1} = \mathsf{Pred}_{\mathcal{L}_2} = \{p\}$ and $\mathsf{Pred}_{\mathcal{L}_3} = \{q\}$:

$$\begin{aligned}
T_1 &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, p(\mathsf{v}_1) \,\} \\
T_2 &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, \neg p(\mathsf{v}_1) \,\} \\
T_3 &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, q(\mathsf{v}_1) \,\}
\end{aligned}$$

$T_1$ and $T_2$ are not definitionally mergeable, since they do not have a common extension as they contradict each other.[25]

Let us define $T_1^+$ where $q$ is defined in terms of $T_1$ as $p$ and let us define $T_3^+$ where $p$ is defined in terms of $T_3$ as $q$, i.e.,

$$\begin{aligned}
T_1^+ &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, p(\mathsf{v}_1),\ \forall \mathsf{v}_1 \big(q(\mathsf{v}_1) \leftrightarrow p(\mathsf{v}_1)\big) \,\} \\
T_3^+ &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, q(\mathsf{v}_1),\ \forall \mathsf{v}_1 \big(p(\mathsf{v}_1) \leftrightarrow q(\mathsf{v}_1)\big) \,\}.
\end{aligned}$$

Then $T_1$ and $T_3$ are definitionally mergeable because $T_1 \nearrow T_1^+$, $T_3 \nearrow T_3^+$, and $T_1^+ \equiv T_3^+$.

[24] Using a propositional constant (i.e., 0-ary relation) $p$, we can have the following smaller counterexample: $T_1 \overset{\text{def}}{=} \{p\}$, $T_2 \overset{\text{def}}{=} \{\neg p\}$ and $T_3 \overset{\text{def}}{=} \emptyset$.

[25] $\exists!$ is an abbreviation for "there exists exactly one", i.e.,

$$\exists!\, \mathsf{v}_i\, \varphi(\mathsf{v}_i) \Leftrightarrow \exists\, \mathsf{v}_i \left(\varphi(\mathsf{v}_i) \wedge \neg \exists\, \mathsf{v}_j \left(\varphi(\mathsf{v}_j) \wedge \mathsf{v}_i \neq \mathsf{v}_j\right)\right).$$

Let us now define $T_2^+$ where $q$ is defined in terms of $T_2$ as $\neg\, p$ and let us define $T_3^\times$ where $p$ is defined in terms of $T_3$ as $\neg\, q$, i.e.,

$$\begin{aligned} T_2^+ &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, \neg\, p(\mathsf{v}_1),\ \forall \mathsf{v}_1\, \big(q(\mathsf{v}_1) \leftrightarrow \neg\, p(\mathsf{v}_1)\big) \,\} \\ T_3^\times &\overset{\text{def}}{=} \{\, \exists!\, \mathsf{v}_1\, \mathsf{v}_1 = \mathsf{v}_1,\ \forall \mathsf{v}_1\, q(\mathsf{v}_1),\ \forall \mathsf{v}_1\, \big(p(\mathsf{v}_1) \leftrightarrow \neg\, q(\mathsf{v}_1)\big) \,\}. \end{aligned}$$

Then $T_2$ and $T_3$ are definitionally mergeable because $T_2 \nearrow T_2^+$, $T_3 \nearrow T_3^\times$, and $T_2^+ \equiv T_3^\times$.

Therefore, being definitionally mergeable is not transitive and hence not an equivalence relation as $T_1 \nearrow\!\nwarrow T_3 \nearrow\!\nwarrow T_2$ but $T_1$ and $T_2$ are not definitionally mergeable. □

**Lemma 3.** Definitional extension is a transitive relation, i.e.,

$$\text{if } T_1 \nearrow T_2 \nearrow T_3, \text{ then } T_1 \nearrow T_3.$$

*Proof.* By $T_1 \nearrow T_2$, there exists a set of explicit definitions $\Delta_{12}$ which defines all predicates of the language of $T_2$ by predicates from the language of $T_1$. Similarly, by $T_2 \nearrow T_3$, there exists a set of explicit definitions $\Delta_{23}$ which defines all predicates of the language of $T_3$ by those of $T_2$. $\Delta_{12}$ generates a translation $\mathsf{tr}_{21} : \mathsf{Form}_2 \to \mathsf{Form}_1$. Consequently, we can define a a set of explicit definitions $\Delta_{13}$ by rewriting all definitions in $\Delta_{23}$ in terms of $\mathcal{L}_1$, i.e.,

$$\Delta_{13} \overset{\text{def}}{=} \Delta_{12} \cup \big\{ \forall \bar{v}\big(p(\bar{v}) \leftrightarrow \mathsf{tr}_{21}(\varphi(\bar{v}))\big) : \forall \bar{v}\big(p(\bar{v}) \leftrightarrow \varphi(\bar{v})\big) \in \Delta_{23} \big\}.$$

$\Delta_{13}$ is a set of explicit definitions and $T_1 \cup \Delta_{13} \equiv T_3$. Thus $T_1 \nearrow T_3$. □

**Theorem 2.** If theories $T_1$, $T_2$ and $T_3$ are formulated in disjoint languages and $T_1 \nearrow\!\nwarrow T_2$ and $T_2 \nearrow\!\nwarrow T_3$, then $T_1$ and $T_3$ are also mergeable, i.e.,

$$T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T_2 \overset{\emptyset}{\nearrow\!\nwarrow} T_3 \text{ and } \mathsf{Pred}_{\mathcal{L}_1} \cap \mathsf{Pred}_{\mathcal{L}_3} = \emptyset, \text{ then } T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T_3.$$

*Proof.* Let $T_1$, $T_2$ and $T_3$ be theories such that both $\mathsf{Pred}_{\mathcal{L}_1} \cap \mathsf{Pred}_{\mathcal{L}_3} = \emptyset$ and $T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T_2 \overset{\emptyset}{\nearrow\!\nwarrow} T_3$ holds. By definition, we have that there exist sets $\Delta_{12}$, $\Delta_{21}$, $\Delta_{23}$ and $\Delta_{32}$ of explicit definitions, such that

$$T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}, \text{ i.e., } \mathsf{Mod}(T_1 \cup \Delta_{12}) = \mathsf{Mod}(T_2 \cup \Delta_{21}), \tag{2}$$

and

$$T_2 \cup \Delta_{23} \equiv T_3 \cup \Delta_{32}, \text{ i.e., } \mathsf{Mod}(T_2 \cup \Delta_{23}) = \mathsf{Mod}(T_3 \cup \Delta_{32}). \tag{3}$$

By Lemma 3 and the assumption that the languages are disjoint, we have that $T_1 \cup \Delta_{12} \cup \Delta_{23}$ is a definitional extension of $T_1$ and $T_3 \cup \Delta_{32} \cup \Delta_{21}$ is a definitional extension of $T_3$.[26] Therefore, it is enough to prove that they are equivalent, i.e.,

$$\mathsf{Mod}(T_1 \cup \Delta_{12} \cup \Delta_{23}) = \mathsf{Mod}(T_3 \cup \Delta_{32} \cup \Delta_{21}).$$

[26] That is, $T_1 \cup \Delta_{12} \cup \Delta_{23} \equiv T_1 \cup \Delta_{13}$ and $T_3 \cup \Delta_{32} \cup \Delta_{21} \equiv T_3 \cup \Delta_{31}$ for some admissible sets of explicit definitions $\Delta_{13}$ and $\Delta_{31}$.

If $\mathfrak{M} \in \mathsf{Mod}(T_1 \cup \Delta_{12} \cup \Delta_{23})$ is a model, then $\mathfrak{M} \models T_1 \cup \Delta_{12} \cup \Delta_{23}$. Therefore, $\mathfrak{M} \models T_2 \cup \Delta_{21}$ by (2) and also $\mathfrak{M} \models T_3 \cup \Delta_{32}$ because of (3) and the fact that $\mathfrak{M} \models \Delta_{23}$. Hence $\mathfrak{M} \models T_3 \cup \Delta_{32} \cup \Delta_{21}$. Consequently,

$$\mathsf{Mod}(T_1 \cup \Delta_{12} \cup \Delta_{23}) \subseteq \mathsf{Mod}(T_3 \cup \Delta_{32} \cup \Delta_{21}).$$

An analogous calculation shows that

$$\mathsf{Mod}(T_1 \cup \Delta_{12} \cup \Delta_{23}) \supseteq \mathsf{Mod}(T_3 \cup \Delta_{32} \cup \Delta_{21}).$$

Therefore, $\mathsf{Mod}(T_1 \cup \Delta_{12} \cup \Delta_{23}) = \mathsf{Mod}(T_3 \cup \Delta_{32} \cup \Delta_{21})$ and this is what we wanted to prove. □

**Theorem 3.** Definitional equivalence $\overset{\triangle}{\equiv}$ is an equivalence relation.

*Proof.* To show that definitional equivalence is an equivalence relation, we need to show that it is reflexive, symmetric and transitive:

- $\overset{\triangle}{\equiv}$ is reflexive because for every theory $T \nearrow T$ since the set of explicit definitions $\Delta$ can be the empty set, and hence $T \overset{\triangle}{\equiv} T$.
- $\overset{\triangle}{\equiv}$ is symmetric: if $T \overset{\triangle}{\equiv} T'$, then there exists a chain $T \dots T'$ of theories connected by $\equiv$, $\nearrow$ and $\nwarrow$. The reverse chain $T' \dots T$ has the same kinds of connections, and hence $T' \overset{\triangle}{\equiv} T$.
- $\overset{\triangle}{\equiv}$ is transitive: if $T_1 \overset{\triangle}{\equiv} T_2$ and $T_2 \overset{\triangle}{\equiv} T_3$, then there exists chains $T_1 \dots T_2$ and $T_2 \dots T_3$ of theories connected by $\equiv$, $\nearrow$ and $\nwarrow$. The concatenated chain $T_1 \dots T_2 \dots T_3$ has the same kinds of connections, and hence $T_1 \overset{\triangle}{\equiv} T_3$. □

**Lemma 4.** If $T \overset{\triangle}{\equiv} T'$, then there is a chain of definitional mergers such that

$$T \nearrow\!\!\nwarrow T_1 \nearrow\!\!\nwarrow T_2 \nearrow\!\!\nwarrow \dots \nearrow\!\!\nwarrow T_n \nearrow\!\!\nwarrow T'.$$

*Proof.* Since definitional extension is reflexive, the finite chain of steps given by Definition 5 for definitional equivalence can be extended by adding extra extension steps $\nearrow$ or $\nwarrow$ wherever needed in the chain, that is when we have subsequent $\nearrow$-steps or $\nwarrow$-steps, and perhaps at the beginning or at the end of the chain. □

**Lemma 5.** Let $T_i$ and $T_j$ be two theories for which $T_i \nearrow\!\!\nwarrow T_j$. Then

- if $T_j \overset{\emptyset}{\simeq} T_j'$ and $\mathsf{Pred}_{\mathcal{L}_i} \cap \mathsf{Pred}_{\mathcal{L}_j'} = \emptyset$, then $T_i \overset{\emptyset}{\nearrow\!\!\nwarrow} T_j'$
- if $T_i \overset{\emptyset}{\simeq} T_i'$ , $T_j \overset{\emptyset}{\simeq} T_j'$ and $\mathsf{Pred}_{\mathcal{L}_i'} \cap \mathsf{Pred}_{\mathcal{L}_j'} = \emptyset$, then $T_i' \overset{\emptyset}{\nearrow\!\!\nwarrow} T_j'$.

*Proof.* Since $T_i \nearrow\!\nwarrow T_j$, there are by Lemma 1 sets $\Delta_{ij}$ and $\Delta_{ji}$ of explicit definitions such that $T_i \cup \Delta_{ij} \equiv T_j \cup \Delta_{ji}$:

$$\Delta_{ij} = \left\{\forall \bar{v}\, (p(\bar{v}) \leftrightarrow \varphi_p(\bar{v})) : p \in \mathsf{Pred}_{\mathcal{L}_j} \setminus \mathsf{Pred}_{\mathcal{L}_i}\right\},$$

i.e., $\varphi_p$ is the definition of predicate $p$ from $\mathsf{Pred}_{\mathcal{L}_j} \setminus \mathsf{Pred}_{\mathcal{L}_i}$.

$$\Delta_{ji} = \left\{\forall \bar{v}\, (q(\bar{v}) \leftrightarrow \varphi_q(\bar{v})) : q \in \mathsf{Pred}_{\mathcal{L}_i} \setminus \mathsf{Pred}_{\mathcal{L}_j}\right\},$$

i.e., $\varphi_q$ is the definition of predicate $q$ from $\mathsf{Pred}_{\mathcal{L}_i} \setminus \mathsf{Pred}_{\mathcal{L}_j}$. We can now define $\Delta_{ij'}$ and $\Delta_{j'i}$ in the following way:

$$\begin{aligned}\Delta_{ij'} \stackrel{\text{def}}{=} &\left\{\forall \bar{v}\ \left(R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)(\bar{v}) \leftrightarrow \varphi_p(\bar{v})\right) : p \in \mathsf{Pred}_{\mathcal{L}_j} \setminus \mathsf{Pred}_{\mathcal{L}_i}\right\}\\ &\qquad \cup \left\{\forall \bar{v}\ \left(R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)(\bar{v}) \leftrightarrow p(\bar{v})\right) : p \in \mathsf{Pred}_{\mathcal{L}_j} \cap \mathsf{Pred}_{\mathcal{L}_i}\right\},\end{aligned}$$

i.e., in $\Delta_{ij'}$ the renaming $R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)$ of predicate $p$ from $\mathsf{Pred}_{\mathcal{L}_j}$ is defined with the same formula $\varphi_j$ as $p$ was defined in $\Delta_{ij}$.

$$\begin{aligned}\Delta_{j'i} \stackrel{\text{def}}{=} &\left\{\forall \bar{v}\ \left(q(\bar{v}) \leftrightarrow \bar{R}^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(\varphi_q(\bar{v}))\right) : q \in \mathsf{Pred}_{\mathcal{L}_i} \setminus \mathsf{Pred}_{\mathcal{L}_j}\right\}\\ &\qquad \cup \left\{\forall \bar{v}\ \left(q(\bar{v}) \leftrightarrow R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(q)(\bar{v})\right) : q \in \mathsf{Pred}_{\mathcal{L}_i} \cap \mathsf{Pred}_{\mathcal{L}_j}\right\},\end{aligned}$$

i.e., in $\Delta_{j'i}$ predicate $q$ from $\mathsf{Pred}_{\mathcal{L}_i}$ is defined with the renaming $\bar{R}^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(\varphi_q)$ of the formula $\varphi_q$ that was used in $\Delta_{ji}$ to define $q$.

Then $T_i \cup \Delta_{ij'} \equiv T'_j \cup \Delta_{j'i}$, and hence we have proven that $T_i \overset{\emptyset}{\nearrow\!\nwarrow} T'_j$.

Similarly, we can define $\Delta_{i'j'}$ and $\Delta_{j'i'}$ as:

$$\begin{aligned}\Delta_{i'j'} \stackrel{\text{def}}{=} &\left\{\forall \bar{v}\left(R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)(\bar{v}) \leftrightarrow \bar{R}^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(\varphi_p(\bar{v}))\right) : p \in \mathsf{Pred}_{\mathcal{L}_j} \setminus \mathsf{Pred}_{\mathcal{L}_i}\right\}\\ &\qquad \cup \left\{\forall \bar{v}\ \left(R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)(\bar{v}) \leftrightarrow R^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(p)(\bar{v})\right) : p \in \mathsf{Pred}_{\mathcal{L}_j} \cap \mathsf{Pred}_{\mathcal{L}_i}\right\},\end{aligned}$$

i.e., in $\Delta_{i'j'}$ the renaming $R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(p)$ of predicate $p$ from $\mathsf{Pred}_{\mathcal{L}_j}$ is defined with the renaming $\bar{R}^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(\varphi_p)$ of the formula $\varphi_p$ that was used in $\Delta_{ij}$ to define $p$.

$$\begin{aligned}\Delta_{j'i'} \stackrel{\text{def}}{=} &\left\{\forall \bar{v}\left(R^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(q)(\bar{v}) \leftrightarrow \bar{R}^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(\varphi_q(\bar{v}))\right) : q \in \mathsf{Pred}_{\mathcal{L}_i} \setminus \mathsf{Pred}_{\mathcal{L}_j}\right\}\\ &\qquad \cup \left\{\forall \bar{v}\ \left(R^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(q)(\bar{v}) \leftrightarrow R^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(q)(\bar{v})\right) : q \in \mathsf{Pred}_{\mathcal{L}_i} \cap \mathsf{Pred}_{\mathcal{L}_j}\right\},\end{aligned}$$

i.e., in $\Delta_{j'i'}$ the renaming $R^{\emptyset}_{\mathcal{L}_i\mathcal{L}'_i}(q)$ of predicate $q$ from $\mathsf{Pred}_{\mathcal{L}_i}$ is defined with the renaming $\bar{R}^{\emptyset}_{\mathcal{L}_j\mathcal{L}'_j}(\varphi_q)$ of the formula $\varphi_q$ that was used in $\Delta_{ji}$ to define $q$.

Then $T'_i \cup \Delta_{i'j'} \equiv T'_j \cup \Delta_{j'i'}$, and hence we have proven that $T'_i \overset{\emptyset}{\nearrow\!\nwarrow} T'_j$. □

**Theorem 4.** Theories $T_1$ and $T_2$ are definitionally equivalent iff there is a theory $T_2'$ which is the disjoint renaming of $T_2$ to a language which is also disjoint from the language of $T_1$ such that $T_2'$ and $T_1$ are definitionally mergeable, i.e.,

$$T_1 \overset{\triangle}{\equiv} T_2 \Leftrightarrow \text{there is a theory } T_2' \text{ such that } T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T_2' \text{ and } T_2' \overset{\emptyset}{\simeq} T_2.$$

*Proof.* Let $T_1$ and $T_2$ be definitional equivalent theories. From Lemma 4, we know that there exists a finite chain of definitional mergers

$$T_1 \nearrow\!\nwarrow \widetilde{T}_1 \nearrow\!\nwarrow \dots \nearrow\!\nwarrow \widetilde{T}_n \nearrow\!\nwarrow T_2.$$

For all $i$ in $\{1,\dots,n\}$, let $\widetilde{T}_i'$ be a renaming of $\widetilde{T}_i$such that $\mathsf{Pred}_{\mathcal{L}_1} \cap \mathsf{Pred}_{\widetilde{\mathcal{L}}_i'} = \emptyset$ and for all $j$ in $\{1,\dots n\}$, if $i \neq j$ then $\mathsf{Pred}_{\widetilde{\mathcal{L}}_i'} \cap \mathsf{Pred}_{\widetilde{\mathcal{L}}_j'} = \emptyset$. Let $T_2'$ be a renaming of $T_2$ such that $\mathsf{Pred}_{\mathcal{L}_1} \cap \mathsf{Pred}_{\mathcal{L}_2'} = \emptyset$, $\mathsf{Pred}_{\mathcal{L}_2} \cap \mathsf{Pred}_{\mathcal{L}_2'} = \emptyset$ and for all $j$ in $\{1,\dots n\}$, $\mathsf{Pred}_{\widetilde{\mathcal{L}}_j'} \cap \mathsf{Pred}_{\mathcal{L}_2} = \emptyset$.

By Lemma 5, $\widetilde{T}_1', \dots, \widetilde{T}_n', T_2'$ is another chain of mergers from $T_1$ to $T_2$

$$T_1 \overset{\emptyset}{\nearrow\!\nwarrow} \widetilde{T}_1' \overset{\emptyset}{\nearrow\!\nwarrow} \dots \widetilde{T}_n' \overset{\emptyset}{\nearrow\!\nwarrow} T_2' \overset{\emptyset}{\simeq} T_2,$$

where all theories in the chain have languages which are disjoint from the languages of all the other theories in the chain, except for $T_1$ and $T_2$ which may have languages which are non-disjoint.

By Theorem 2, the consecutive mergers from $T_1$ to $T_2'$ can be compressed into one merger. So $T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T_2' \overset{\emptyset}{\simeq} T_2$ and this is what we wanted to prove.

To show the converse direction, let us assume that $T_1$ and $T_2$ are such theories that there is a disjoint renaming theory $T_2'$ of $T_2$ for which $T_1 \nearrow\!\nwarrow T_2'$. As $T_2'$ is a disjoint renaming of $T_2$, we have by Remark 6 that $T_2' \overset{\emptyset}{\nearrow\!\nwarrow} T_2$. Therefore, there is a chain $T^+$, $T^\times$ of theories such that $T_1 \nearrow T^+ \nwarrow T_2' \nearrow T^\times \nwarrow T_2$. Hence $T_1 \overset{\triangle}{\equiv} T_2$. □

**Corollary 2.** Two theories are definitionally equivalent iff they can be connected by two definitional mergers:

$$T_1 \overset{\triangle}{\equiv} T_2 \Leftrightarrow \text{there is a theory } T \text{ such that } T_1 \overset{\emptyset}{\nearrow\!\nwarrow} T \overset{\emptyset}{\nearrow\!\nwarrow} T_2.$$

Consequently, the chain $T_1, \dots, T_n$ in Definition 5 can always be chosen to be at most length four.

*Proof.* This follows immediately from Theorem 4 and Remark 6. □

**Theorem 5.** Definitional equivalence is the finest equivalence relation containing definitional mergeability. In fact $\overset{\triangle}{\equiv}$ is the transitive closure of $\nearrow\!\nwarrow$.

*Proof.* From Remark 4, we know that $\overset{\triangle}{\equiv}$ is an extension of $\nearrow\!\nwarrow$. To prove that $\overset{\triangle}{\equiv}$ is the transitive closure of $\nearrow\!\nwarrow$, it is enough to show that $T_1 \overset{\triangle}{\equiv} T_2$ holds if there is a chain $T'_1, \ldots, T'_n$ of theories such that $T_1 = T'_1$, $T_2 = T'_n$, and $T'_i \nearrow\!\nwarrow T'_{i+1}$ for all $1 \leq i < n$. By Theorem 4, there is a theory $T'$ such that $T_1 \nearrow\!\nwarrow T' \overset{\emptyset}{\simeq} T_2$. By Remark 6, $T_1 \nearrow\!\nwarrow T' \nearrow\!\nwarrow T_2$ which proves our statement. □

It is known that, for disjoint languages, being definitionally mergeable and intertranslatability are equivalent, see e.g., (Barrett and Halvorson 2016*a*, Theorems 1 and 2). Now we show that, for disjoint languages, definitional equivalence also coincides with these concepts, i.e.:

**Theorem 6.** Let $T$ and $T'$ be theories formulated in disjoint languages. Then

$$T \overset{\triangle}{\equiv} T' \Leftrightarrow T \overset{\emptyset}{\nearrow\!\nwarrow} T' \Leftrightarrow T \rightleftarrows T'.$$

*Proof.* Since $T \overset{\emptyset}{\nearrow\!\nwarrow} T' \Leftrightarrow T \rightleftarrows T'$ is proven by (Barrett and Halvorson 2016*a*, Theorems 1 and 2), we only have to prove that $T \overset{\triangle}{\equiv} T' \Leftrightarrow T \overset{\emptyset}{\nearrow\!\nwarrow} T'$.

Let theories $T$ and $T'$ be definitionally equivalent theories in disjoint languages, i.e., $\mathsf{Pred}_{\mathcal{L}} \cap \mathsf{Pred}_{\mathcal{L}'} = \emptyset$. Since they are definitionally equivalent, there exists, by Theorem 4 a chain which consists of a single mergeability and a renaming step between $T$ and $T'$. Since $T$ and $T'$ are disjoint, and since renaming by Remark 6 is also a disjoint merger, these two steps can by Theorem 2 be reduced to one step $T \overset{\emptyset}{\nearrow\!\nwarrow} T'$, and this is what we wanted to prove.

The converse direction follows straightforwardly from the definitions. □

Let us now look at the relation between syntax and semantics, and consider models of theories.

**Theorem 7.** Let $T_1$ and $T_2$ be arbitrary theories, then $T_1$ and $T_2$ are mergeable iff they are model mergeable, i.e.,

$$T_1 \nearrow\!\nwarrow T_2 \;\Leftrightarrow\; \mathsf{Mod}(T_1) \nearrow\!\nwarrow \mathsf{Mod}(T_2).$$

*Proof.* Let $T_1$ and $T_2$ be arbitrary theories.

Let us first assume that $T_1 \nearrow\!\nwarrow T_2$ and prove that $\mathsf{Mod}(T_1) \nearrow\!\nwarrow \mathsf{Mod}(T_2)$. We know from Lemma 1 that there exist sets of explicit definitions $\Delta_{12}$ and $\Delta_{21}$ such that $T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}$. Therefore, by Definition 1, $\mathsf{Mod}(T_1 \cup \Delta_{12}) = \mathsf{Mod}(T_2 \cup \Delta_{21})$. We construct map $\beta$ between $\mathsf{Mod}(T_1)$ and $\mathsf{Mod}(T_2)$ by expanding models of $T_1$ using the explicit definitions in $\Delta_{12}$, which since $\mathsf{Mod}(T_1 \cup \Delta_{12}) = \mathsf{Mod}(T_2 \cup \Delta_{21})$ will be a model of $T_1 \cup \Delta_{12}$, and then by taking the reduct to the languages of $T_2$. This map associating the appropriate reducts to the models of $T_2 \cup \Delta_{21}$ is the inverse of the map from $\mathsf{Mod}(T_2)$ to $\mathsf{Mod}(T_2 \cup \Delta_{21})$ associating the definitional expansion of the models of $T_2$ using definitions $\Delta_{21}$. Since definitional expansions are bijections, we have that $\beta$ is also a bijection. Through this construction, the relations in $\beta(\mathfrak{M})$ are the ones defined in $\mathfrak{M}$ according to $\Delta_{12}$ and vice versa, the relations in $\mathfrak{M}$ are the ones defined in $\beta(\mathfrak{M})$ according to $\Delta_{21}$, and clearly the underlying set of $\mathfrak{M}$ and $\beta(\mathfrak{M})$ are the same. Hence $\mathsf{Mod}(T_1) \nearrow\!\nwarrow \mathsf{Mod}(T_2)$.

Let us now assume that $\mathsf{Mod}(T_1) \nearrow\!\nwarrow \mathsf{Mod}(T_2)$ and prove that $T_1 \nearrow\!\nwarrow T_2$. We know by Definition 7 that there is a bijection $\beta$ between $\mathsf{Mod}(T_1)$ and $\mathsf{Mod}(T_2)$ that is defined along two sets $\Delta_{12}$ and $\Delta_{21}$ of explicit definitions such that if $\mathfrak{M} \in \mathsf{Mod}(T_1)$, then

1. the underlying set of $\mathfrak{M}$ and $\beta(\mathfrak{M})$ are the same,
2. the relations in $\beta(\mathfrak{M})$ are the ones defined in $\mathfrak{M}$ according to $\Delta_{12}$ and vice versa, the relations in $\mathfrak{M}$ are the ones defined in $\beta(\mathfrak{M})$ according to $\Delta_{21}$.

Let $\mathfrak{M}^+$ be a model of $T_1 \cup \Delta_{12}$ and let $\mathfrak{M}$ be its reduct to the language of $T_1$. Then $\beta(\mathfrak{M})$ is a model of $T_2$ having the same underlying set by item 1. Let $\beta(\mathfrak{M})^+$ be the expansion of $\beta(\mathfrak{M})$ by definitions in $\Delta_{21}$. By item 2, we have that $\beta(\mathfrak{M})^+ = \mathfrak{M}^+$. Clearly, $\beta(\mathfrak{M})$ is a model of $T_2 \cup \Delta_{21}$. Consequently, $\mathsf{Mod}(T_1 \cup \Delta_{12}) \subseteq \mathsf{Mod}(T_2 \cup \Delta_{21})$. An analogous argument can show that $\mathsf{Mod}(T_1 \cup \Delta_{12}) \supseteq \mathsf{Mod}(T_2 \cup \Delta_{21})$. Therefore, $\mathsf{Mod}(T_1 \cup \Delta_{12}) = \mathsf{Mod}(T_2 \cup \Delta_{21})$, and thus by Definition 1, $T_1 \cup \Delta_{12} \equiv T_2 \cup \Delta_{21}$. Consequently, $T_1 \nearrow\!\nwarrow T_2$. □

**Theorem 8.** Let $T_1$ and $T_2$ be arbitrary theories. Then $T_1$ and $T_2$ are definitionally equivalent iff they are intertranslatable, i.e.,

$$T_1 \overset{\triangle}{\equiv} T_2 \;\Leftrightarrow\; T_1 \rightleftarrows T_2.$$

*Proof.* Let us first assume that $T_1 \overset{\triangle}{\equiv} T_2$. Let $T'$ be a disjoint renaming of $T_2$ to a language which is also disjoint from the language of $T_1$. By Remark 6 and the transitivity of $\overset{\triangle}{\equiv}$, we have $T_1 \overset{\triangle}{\equiv} T' \overset{\triangle}{\equiv} T_2$. By Theorem 6, $T_1 \rightleftarrows T' \rightleftarrows T_2$. Consequently, $T_1 \rightleftarrows T_2$ because relation $\rightleftarrows$ is transitive.

To prove the converse, let us assume that $T_1 \rightleftarrows T_2$. Let $T'$ again be a disjoint renaming of $T_2$ to a language which is also disjoint from the language of $T_1$. By Remark 6 and the transitivity of $\rightleftarrows$, we have $T_1 \rightleftarrows T' \rightleftarrows T_2$. By Theorem 6, $T_1 \overset{\triangle}{\equiv} T' \overset{\triangle}{\equiv} T_2$. Thus, $T_1 \overset{\triangle}{\equiv} T_2$ because relation $\overset{\triangle}{\equiv}$ is transitive. □

**Theorem 9.** Let $T_1$ and $T_2$ be arbitrary theories, then $T_1$ and $T_2$ are intertranslatable iff their models are intertranslatable, i.e.,

$$T_1 \rightleftarrows T_2 \Leftrightarrow \mathsf{Mod}(T_1) \rightleftarrows \mathsf{Mod}(T_2)$$

*Proof.* Let us first assume that $T_1 \rightleftarrows T_2$ and prove that $\mathsf{Mod}(T_1) \rightleftarrows \mathsf{Mod}(T_2)$. Let $\mathsf{tr}_{12}$ and $\mathsf{tr}_{21}$ be the corresponding interpretations between $T_1$ and $T_2$. It is enough to show that $\mathsf{tr}^*_{12} : \mathsf{Mod}(T_2) \to \mathsf{Mod}(T_1)$ and $\mathsf{tr}^*_{21} : \mathsf{Mod}(T_1) \to \mathsf{Mod}(T_2)$ are bijections and are inverses of each other.

For all $\mathfrak{M} \in \mathsf{Mod}(T_1)$,

$$\mathfrak{M} \models \forall \bar{v}\big(\varphi(\bar{v}) \leftrightarrow \mathsf{tr}_{21}\big(\mathsf{tr}_{12}(\varphi(\bar{v}))\big)\big).$$

By Definition 2 and Remark 2, this is equivalent to

$$\mathfrak{M} \models \varphi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{21}(\mathsf{tr}_{12}(\varphi))[e]$$

for all evaluations $e : \mathsf{Var} \to M$.

By applying Lemma 2 twice,

$$\mathfrak{M} \models \mathsf{tr}_{21}(\mathsf{tr}_{12}(\varphi))[e] \Leftrightarrow \mathsf{tr}^*_{21}(\mathfrak{M}) \models \mathsf{tr}_{12}(\varphi)[e] \Leftrightarrow \mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M})) \models \varphi[e].$$

Consequently,

$$\mathfrak{M} \models \varphi[e] \Leftrightarrow \mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M})) \models \varphi[e].$$

Since $M$ is the underlying set of both $\mathfrak{M}$ and $\mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M}))$, this implies that $\mathfrak{M} = \mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M}))$. A completely analogous proof shows that $\mathfrak{N} = \mathsf{tr}^*_{21}(\mathsf{tr}^*_{12}(\mathfrak{N}))$ for all models $\mathfrak{N}$ of $T_2$.

Consequently, $\mathsf{tr}^*_{12}$ and $\mathsf{tr}^*_{21}$ are everywhere defined and they are inverses of each other because when we combine them we get the identity, and hence they are bijections, which is what we wanted to prove.

Let us now assume that $\mathsf{Mod}(T_1) \rightleftarrows \mathsf{Mod}(T_2)$ and prove that $T_1 \rightleftarrows T_2$. By Definition 13 and Corollary 1, we know that there are interpretations $\mathsf{tr}_{12}$ and $\mathsf{tr}_{21}$ between $T_1$ and $T_2$ such that the induced maps $\mathsf{tr}^*_{12} : \mathsf{Mod}(T_1) \to \mathsf{Mod}(T_2)$ and $\mathsf{tr}^*_{21} : \mathsf{Mod}(T_2) \to \mathsf{Mod}(T_1)$ are bijections which are inverses of each other, and thus $\mathfrak{M} = \mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M}))$ for all models $\mathfrak{M}$ of $T_1$. Since $M$ is the underlying set of $\mathfrak{M}$, and $\mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M}))$, we have that

$$\mathfrak{M} \models \varphi[e] \Leftrightarrow \mathsf{tr}^*_{12}(\mathsf{tr}^*_{21}(\mathfrak{M})) \models \varphi[e].$$

From this, by applying Lemma 2 twice, we get

$$\mathfrak{M} \models \varphi[e] \Leftrightarrow \mathfrak{M} \models \mathsf{tr}_{21}(\mathsf{tr}_{12}(\varphi))[e].$$

for all evaluations $e : \mathsf{Var} \to M$. By Definition 2 and Remark 2, the above is equivalent to

$$\mathfrak{M} \models \forall \bar{v}\big(\varphi(\bar{v}) \leftrightarrow \mathsf{tr}_{21}\big(\mathsf{tr}_{12}(\varphi(\bar{v}))\big)\big).$$

A completely analogous proof shows that

$$\mathfrak{N} \models \forall \bar{v}\big(\psi(\bar{v}) \leftrightarrow \mathsf{tr}_{12}\big(\mathsf{tr}_{21}(\psi(\bar{v}))\big)\big),$$

from which follows by Definition 10 that $T_1 \rightleftarrows T_2$. □

**Remark 7.** If we use the notations of this paper, Theorem 4.2 of (Andréka and Németi 2014) claims, without proof, that (i) definitional equivalence, (ii) definitional mergeability, (iii) intertranslatability and (iv) model mergeability are equivalent in case of disjoint languages. In this paper, we have not only proven these statements, but we also showed which parts can be generalized to arbitrary languages and which cannot. In detail:

- item (i) is equivalent to item (iii) by Theorem 6, and we have generalized this equivalence to theories in arbitrary languages by Theorem 8,
- the equivalence of items (ii) and (iv) have been generalized to theories in arbitrary languages by Theorem 7,
- items (i) and (ii) are indeed equivalent for theories in disjoint languages by Theorem 6; however, they are not equivalent for theories in non-disjoint languages by the counterexample in Theorem 1,
- in Definition 13, we have introduced a model theoretic counterpart of intertranslatability which, by Theorem 9, is equivalent to it even if the languages are not disjoint.

# 4 Conclusion

Since definitional mergeability is not transitive, by Theorem 1, and thus not an equivalence relation, the Barrett–Halvorson generalization is not a well-founded criterion for definitional equivalence when the languages of theories are not disjoint. Contrary to this, the Andréka–Németi generalization of definitional equivalence is an equivalence relation, by Theorem 3. It is also equivalent to intertranslatability, by Theorem 8, and to model intertranslatability, by Theorem 9, even for non-disjoint languages. Therefore, the Andréka–Németi generalization is more suitable to be used as the extension of definitional equivalence between theories of arbitrary languages. It is worth noting, however, that the two generalizations are really close to each-other since the Andréka–Németi generalization is the transitive closure of the Barrett-Halvorson one, see Theorem 5. Moreover, they only differ in at most one disjoint renaming, see Theorems 4 and 6, and as long as we restrict ourselves to theories which all have mutually disjoint languages, Barrett–Halvorson's definition is transitive by Theorem 2.

We hope to have provided a building block for a framework for comparing theories, and to have clarified the relations between the different ways in which theories can be equivalent.

## Acknowledgements

The writing of the current paper was induced by questions by Marcoen Cabbolet and Sonja Smets during the public defence of (Lefever 2017). We are also grateful to Hajnal Andréka, Michèle Friend, Mohamed Khaled, Amedé Lefever, István Németi and Jean Paul Van Bendegem for enjoyable discussions and feedback while writing this paper, as well as to the two anonymous referees who made valuable remarks which helped significantly improving our paper.

## References

Andréka, H., Madarász, J. X. and Németi, I. (2005), 'Mutual definability does not imply definitional equivalence, a simple example', *Mathematical Logic Quarterly* **51,6**, 591–597.

Andréka, H., Madarász, J. X. and Németi, I. (2008), Defining new universes in many-sorted logic, Research report, Alfréd Rényi Institute of Mathematics, Hungar. Acad. Sci., Budapest.
**URL:** *https://www.researchgate.net/publication/242602426*

Andréka, H., Madarász, J. X., Németi, I., with contributions from: Andai, A., Sági, G., Sain, I. and Tőke, C. (2002), *On the logical structure of relativity theories*, Research report, Alfréd Rényi Institute of Mathematics, Hungar. Acad. Sci., Budapest.
**URL:** *https://old.renyi.hu/pub/algebraic-logic/Contents.html*

Andréka, H. and Németi, I. (2014), 'Definability theory course notes'.
**URL:** *https://old.renyi.hu/pub/algebraic-logic/DefThNotes0828.pdf*

Andréka, H., Németi, I. and Sain, I. (2001), Algebraic logic, *in* 'Handbook of Philosophical Logic Volume II', Springer Verlag, pp. 133–248.

Barrett, T. W. and Halvorson, H. (2016*a*), 'Glymour and Quine on theoretical equivalence', *Journal of Philosophical Logic* **45**(5), 467–483.

Barrett, T. W. and Halvorson, H. (2016*b*), 'Morita equivalence', *The Review of Symbolic Logic* **9**(3), 556–582.

Chang, H. (2012), *Is Water $H_2O$? Evidence, Realism and Pluralism*, Springer, Dordrecht.

Corcoran, J. (1980), 'On definitional equivalence and related topics', *History and Philosophy of Logic* **1**(1-2), 231–234.

de Bouvère, K. L. (1965*a*), 'Logical synonymy.', *Indagationes Mathematicae* **27**, 622–629.

de Bouvère, K. L. (1965*b*), Synonymous theories., *in* 'The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley', North Holland, pp. 402–406.

Feferman, S. (1960), 'Arithmetization of metamathematics in a general setting.', *Fundamenta Mathematicae* **49**, 35–92.

Friedman, H. A. and Visser, A. (2014), 'When bi-interpretability implies synonymy'.

Friend, M., Khaled, M., Lefever, K. and Székely, G. (2018), 'Distances between formal theories.'.
**URL:** *https://arxiv.org/abs/1807.01501*

Fujimoto, K. (2010), 'Relative truth definability of axiomatic truth theories.', *Bulletin of Symbolic Logic* **16(3)**, 305–344.

Glymour, C. (1970), 'Theoretical realism and theoretical equivalence', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* **1970**, 275–288.

Glymour, C. (1977), 'Symposium on space and time: The epistemology of geometry', *Noûs* **11**(3), 227–251.

Glymour, C. (1980), *Theory and Evidence*, Princeton.

Heath, T. L. (1956), *The Thirteen Books of Euclid's Elements ([Facsimile. Original publication: Cambridge University Press, 1908] 2nd ed.)*, Dover Publications.

Henkin, L., Monk, J. and Tarski, A. (1971), *Cylindric Algebras Part I*, North-Holland.

Henkin, L., Monk, J. and Tarski, A. (1985), *Cylindric Algebras Part II*, North-Holland.

Hodges, W. (1993), *Model Theory*, Cambridge University Press.

Hodges, W. (1997), *A Shorter Model Theory*, Cambridge University Press.

Kuhn, T. (1957), *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought.*, Harvard University Press.

Lefever, K. (2017), Using Logical Interpretation and Definitional Equivalence to compare Classical Kinematics and Special Relativity Theory, PhD thesis, Vrije Universiteit Brussel.

Lefever, K. and Székely, G. (2018), 'Comparing classical and relativistic kinematics in first-order-logic', *Logique et Analyse* **61**(241), 57–117.

Madarász, J. X. (2002), Logic and Relativity (in the light of definability theory), PhD thesis, Eötvös Loránd Univ., Budapest.

Montague, R. (1956), Contributions to the axiomatic foundations of set theory, PhD thesis, Berkeley.

Pinter, C. C. (1978), 'Properties preserved under definitional equivalence and interpretations.', *Zeitschr. f. math. Logik und Grundlagen d. nlath.* **24**, 481–488.

Playfair, J. (1846), *Elements of Geometry*, W. E. Dean.

Quine, W. V. (1946), 'Concatenation as a basis for arithmetic.', *The Journal of Symbolic Logic* **11(4)**, 105–114.

Tarski, A., Mostowski, A. and Robinson, R. (1953), *Undecidable Theories*, Elsevier.

Visser, A. (2006), Categories of theories and interpretations, *in* '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', ASL, A.K. Peters, Ltd., Wellesley, Mass., pp. 284–341.

Visser, A. (2015), 'Extension & interpretability', *Logic Group preprint series* **329**.
**URL:** *https://dspace.library.uu.nl/handle/1874/319941*

KOEN LEFEVER
Centre for Logic and Philosophy of Science
Vrije Universiteit Brussel
koen.lefever@vub.be
http://lefever.space/

GERGELY SZÉKELY
Alfréd Rényi Institute for Mathematics
Hungarian Academy of Sciences
szekely.gergely@renyi.mta.hu
http://www.renyi.hu/~turms/

This is a post-peer-review, pre-copyedit version of an article published in the *Journal of Philosophical Logic*.

The final authenticated version is available online at: http://dx.doi.org/10.1007/s10992-018-9491-0.