Gaps, Gluts, and Theoretical Equivalence

Carlo Nicolai
King's College London

Abstract. When are two formal theories of broadly logical concepts, such as truth, equivalent? The paper investigates a case study, involving two well-known variants Kripke-Feferman truth. The first, $\mathtt{KF}+\mathtt{CONS}$, features a consistent but partial truth predicate. The second, $\mathtt{KF}+\mathtt{COMP}$, an inconsistent but complete truth predicate. It is well-known that the two truth predicates are dual to each other. We show that this duality reveals a much stricter correspondence between the two theories: they are intertraslatable. Intertranslatability under natural assumptions coincides with definitional equivalence, and is arguably the strictest notion of theoretical equivalence different from logical equivalence. The case of $\mathtt{KF}+\mathtt{CONS}$ and $\mathtt{KF}+\mathtt{COMP}$ raises a puzzle: the two theories can be proved to be strictly related, yet they appear to embody remarkably different conceptions of truth. The puzzle can be solved by reflecting on the scope and limitations of formal notions of theoretical equivalence in certain contexts.

1. Introduction

When are two formal theories of broadly logical concepts, such as truth, equivalent? From the work of logicians and philosophers of science, we know that there are several notions of mutual reduction between formal theories to choose from (Halvorson, 2019; Visser, 2006). Glymour (1970) proposed the (demanding) criterion of theoretical equivalence known as *definitional equivalence* or *intertranslatability*. The criterion roughly states that two theories are equivalent if each theory can define the primitive concepts of the other in a sufficiently natural way. "Natural" here has a definite sense: each theory should recognize that the other theory's definitions of its own primitives are the inverse of its own definitions (see §3 for a precise definition). Several theorists agree that intertranslatability is an unrealistically strict criterion (Weatherall, 2019). For our purposes, however, this strictness is an advantage.

The ever-increasing popularity of truth-theoretic deflationism (Cieśliński, 2017), together with a revived attention to the Liar paradox prompted by new technical tools (Field, 2008; Horsten, 2012; Halbach, 2014), led to a multiplication of formal systems extending some standard syntax theory with a primitive truth predicate governed by suitable axiom. These systems have a twofold nature: on the one hand they embody some conception of truth, including a solution to the difficulty posed by paradox; on the other hand they characterize the truth predicate as a logical tool whose formal properties witness the role that the notion of truth can play in (sustained) reasoning – e.g. in applied mathematics and in the formal sciences. The existence of several such systems leads naturally to the question of how to compare them, both in their formal and philosophical aspects.

In what follows we study the question whether the formal notions of theoretical equivalence devised from logicians and philosophers of science can support an adequate comparison between formal theories of primitive truth. We focus on the case study of one of the most influential cluster of theories of truth, the Kripke-Feferman theory. Kripke-Feferman truth traces back to the work of Feferman on the foundations of predicativism (Feferman, 1991), and it is often presented as an axiomatization in classical logic of the class of fixed-point models proposed by Kripke (1975). Kripke-Feferman truth is not a single theory, but rather a *recipe* to generate theories featuring truth predicates with different properties. We will focus on two theories from the Kripke-Feferman cluster. The first is the theory $\mathtt{KF} + \mathtt{CONS}$, whose truth predicate is consistent but partial (not every sentence is true or false). The second is the theory $\mathtt{KF} + \mathtt{COMP}$, whose truth predicate is inconsistent and complete (every sentence is either true or false). In the light of these differences, it is implausible to consider $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ as theoretically equivalent theories of truth.

Yet, in §4, we will show that $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ are intertranslatable. This is certainly puzzling. Two truth predicates that reflect distinct concepts of truth stand in a relation of theoretical inter-reduction that is considered to be *too strong* by philosophers of science. We will discuss a potential way out of this puzzle in §5. Before this, in §2, we will introduce Kripke-Feferman truth and some of the key properties of the truth predicates of $\mathtt{KF}+\mathtt{CONS}$ and $\mathtt{KF}+\mathtt{COMP}$. Since we are drawing parallels between (formalizations of) scientific theories and theories of primitive truth, we will also recall some theoretical contexts in which Kripke-Feferman truth has been employed. §4 contains the main technical observations of the paper: the well-known phenomenon of the duality between $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ will be recalled, and we will show that it can be lifted to the intertranslatability of the two theories.

2. Kripke-Feferman truth

The system $\mathtt{KF}$ is formulated in the language $\mathscr{L}_{\mathrm{Tr}}$ obtained by extending the language $\mathscr{L}_{\mathbb{N}}$ of arithmetic with a unary truth predicate Tr applying to Gödel codes of sentences of $\mathscr{L}_{\mathrm{Tr}}$. It is convenient to formulate $\mathscr{L}_{\mathbb{N}}$ in a relational signature: we assume only a finite number of primitive recursive relations. An axiomatization of first-order arithmetic with these features can be found in (Hájek and Pudlák, 2017, §I(e)). The expression $\dot{x}$ stands for the code of a constant symbol c_x associated with any x. It is well-known that there are primitive recursive injective functions sending each x to c_x, e.g. the numeral function.

KF extends classical logic with equality with the basic axioms of a relational version of PA, the induction schema for $\mathtt{IND}(\mathscr{L}_{\mathrm{Tr}})$ for formulae of the entire language $\mathscr{L}_{\mathrm{Tr}}$, and the following truth-theoretic axioms:

(KF1) $\forall x_1 \dots x_n (\mathrm{Tr} \ulcorner R(\dot{x}_1, ..., \dot{x}_n) \urcorner \leftrightarrow R(x_1, ..., x_n))$

(KF2) $\forall x_1 \dots x_n (\mathrm{Tr} \ulcorner \neg R(\dot{x}_1, ..., \dot{x}_n) \urcorner \leftrightarrow \neg R(x_1, ..., x_n))$

(KF3) $\forall x (\mathrm{Tr} \ulcorner \mathrm{Tr}\, \dot{x} \urcorner \leftrightarrow \mathrm{Tr}\, x)$

(KF4) $\forall t (\mathrm{Tr} \ulcorner \neg \mathrm{Tr}\, \dot{x} \urcorner \leftrightarrow \mathrm{Tr} \underset{\cdot}{\neg} x)$

(KF5) $\forall \varphi \forall \psi (\mathrm{Tr}\, (\varphi \underset{\cdot}{\wedge} \psi) \leftrightarrow (\mathrm{Tr}\, \varphi \wedge \mathrm{Tr}\, \psi))$

(KF6) $\forall \varphi \forall \psi (\mathrm{Tr} \underset{\cdot}{\neg} (\varphi \underset{\cdot}{\wedge} \psi) \leftrightarrow (\mathrm{Tr} \underset{\cdot}{\neg} \varphi \vee \mathrm{Tr} \underset{\cdot}{\neg} \psi))$

(KF7) $\forall v \forall \varphi (\mathrm{Tr} \underset{\cdot}{\forall} v \varphi \leftrightarrow \forall y\, \mathrm{Tr}\, \varphi(\dot{y}/v))$

(KF8) $\forall v \forall \varphi (\mathrm{Tr} \underset{\cdot}{\neg} \underset{\cdot}{\forall} v \varphi \leftrightarrow \exists y\, \mathrm{Tr} \underset{\cdot}{\neg} \varphi(\dot{y}/v))$

(KF9) $\forall \varphi (\mathrm{Tr} \underset{\cdot}{\neg} \underset{\cdot}{\neg} \varphi \leftrightarrow \mathrm{Tr}\, \varphi)$

In KF5-KF8, the quantification $\forall\varphi$... abbreviates $\forall x (\mathtt{Sent}_{\mathscr{L}_{\mathrm{Tr}}}(x) \rightarrow$ ... Other conventions follow Halbach (2014), including the under-dotting convention for syntactic functions.

The theory KFI is obtained by replacing the $\mathscr{L}_{\mathrm{Tr}}$-induction schema of KF with the axiom of *internal induction*:[1]

(I-IND($\mathscr{L}_{\mathrm{Tr}}$))

$$\forall x \Big(\mathtt{Sent}(\underset{\cdot}{\forall} v x) \rightarrow (\mathrm{Tr}\, x(0/v) \wedge \forall y (\mathrm{Tr}\, x(\dot{y}/v) \rightarrow \mathrm{Tr}\, x(\dot{\mathrm{S}} y/v) \rightarrow \forall y \mathrm{Tr}\, x(\dot{y}/v)) \Big)$$

The theory $\mathrm{KF} \restriction \mathscr{L}_{\mathbb{N}}$ is like KF but features only the axiom schema of induction restricted to formulae of $\mathscr{L}_{\mathbb{N}}$.

In this paper we will focus on two extensions of KF obtained by adding, respectively, the axioms

(CONS) $\forall x (\mathrm{Tr} \underset{\cdot}{\neg} x \rightarrow \neg \mathrm{Tr}\, x)$,

(COMP) $\forall x (\neg \mathrm{Tr}\, x \rightarrow \mathrm{Tr} \underset{\cdot}{\neg} x)$.

The Liar paradox shows that KF + CONS and KF + COMP are mutually inconsistent.

[1] We prefer to present the full (non-abbreviated) version because of the restriction of unary formulae.

KF can be seen as an axiomatization of Kripke's fixed point semantics Kripke (1975). Let $\Phi\colon \mathscr{P}\omega \longrightarrow \mathscr{P}\omega$ be the arithmetical operator associated with the Kripke truth set (Halbach, 2014, p. 190).[2] A fixed point $X \subseteq \omega$ of Φ is *consistent* if there is no sentence $\varphi \in \mathscr{L}_{\mathrm{Tr}}$ such that $\{\varphi, \neg\varphi\} \subset X$; a *complete* fixed point is such that either $\varphi \in X$ or $\neg\varphi \in X$ for any sentence φ of $\mathscr{L}_{\mathrm{Tr}}$. Standard models of KF + CONS are precisely the consistent fixed points, and standard models of KF + COMP are the complete ones:

FACT 1.

(i) S is a consistent fixed point of Φ iff $(\mathbb{N}, S) \vDash \mathtt{KF} + \mathtt{CONS}$;

(ii) S is a complete fixed point of Φ iff $(\mathbb{N}, S) \vDash \mathtt{KF} + \mathtt{COMP}$.

2.1. **Gaps and Gluts.** KF + CONS and KF + COMP embody different concepts of truth. For future reference, we recall some simple facts separating the two truth predicates. They disagree on almost any key principle available in the theories, such as paradoxical sentences, axioms, and rules of inference.

The truth predicate of KF + CONS is partial and does not declare any sentence to be both true and false. This entails that that there are sentences that are consequences of KF + CONS and yet they are declared not true by the theory. For our purposes, let us define the Liar sentence λ as the sentence $\neg\mathrm{Tr}\, l$, for which the identity $l = \ulcorner\neg\mathrm{Tr}\, l\urcorner$ provable in PA. If $\mathrm{Tr}\, l$, then CONS entails that $\neg\mathrm{Tr}\, l$, that is λ. But $\mathrm{Tr}\, l$ also entails $\neg\lambda$. Therefore, λ is a theorem of KF + CONS. This phenomenon extends to *instances* of some axioms of KF + CONS, such the instance

$$\mathrm{Tr}\, \underset{.}{\neg} l \rightarrow \neg\mathrm{Tr}\, l$$

of CONS. They are declared not true by the theory. The argument is quite straightforward: assuming $\mathrm{Tr}\, (\ulcorner\neg\mathrm{Tr}\, \underset{.}{\neg} l \vee \neg\mathrm{Tr}\, l\urcorner)$ in KF + CONS, by distributing the truth predicate and applying KF3, 4, 9 one gets $\mathrm{Tr}\, l \vee \mathrm{Tr}\, l$, that is $\mathrm{Tr}\, l$, which contradicts $\neg\mathrm{Tr}\, l$.[3] A noticeable feature of KF + CONS is that it derives a universally quantified version of the modal axioms K:

$$\forall\varphi\forall\psi(\mathrm{Tr}\,\varphi \wedge \mathrm{Tr}\,(\varphi \rightarrow \psi) \rightarrow \mathrm{Tr}\,\psi). \tag{1}$$

The proof of (1) in KF + CONS relies essentially on CONS and on the compositional axioms: if $\mathrm{Tr}\,\varphi$ and $\mathrm{Tr}\,(\neg\varphi \vee \psi)$, then $\mathrm{Tr}\,\varphi$ holds together with $\mathrm{Tr}\,\neg\varphi$ or with $\mathrm{Tr}\,\psi$

[2] I will not distinguish between sentences and their codes in the following semantic considerations.

[3] Some authors, such as Field, claim that classical gap theories (such as KF + CONS) are bound to declare their non-logical axioms untrue (Field, 2008, Ch.7). As we have just seen, only some instances of such axioms behave in this way. Since the meaning of the external universal quantifier is governed by classical logic, and the logic of the truth predicate is nonclassical, the fact that the link between a universally quantified axiom and the truth of its instances is broken may not be so devastating as having truth axioms that are not true.

by compositionality. The first option contradicts CONS, while the second gives us $\mathrm{Tr}\,\psi$. Either way, (1) follows. The sentence (1) can be seen as the claim that KF + CONS formalizes the assertion that its (classical) rules of inference are truth preserving. This is especially adequate if, without loss of generality, one formulates KF + CONS in a Hilbert system for classical logic in which the only rule of inference is Modus Ponens (Enderton, 2001, Ch. 2.4).

By contrast, the truth predicate of KF + COMP is inconsistent. Consider again the sentence λ. Reasoning in KF + COMP: if λ, then $\mathrm{Tr}\,\ulcorner\lambda\urcorner$ by COMP; also, λ entails $\neg\mathrm{Tr}\,l$. Therefore $\neg\lambda$, that is $\mathrm{Tr}\,\ulcorner\lambda\urcorner$, by classical logic. Now suppose $\neg\mathrm{Tr}\,\underset{\cdot}{\neg} l$. By KF4, this entails λ. But we have just established $\neg\lambda$. Therefore, $\mathrm{Tr}\,\ulcorner\neg\lambda\urcorner$ after all. By KF5, $\mathrm{Tr}\,\ulcorner\lambda\wedge\neg\lambda\urcorner$.

By a straightforward induction on the complexity of the sentence A of $\mathscr{L}_{\mathrm{Tr}}$, all instances of the schema

$$(\mathrm{Tr}\text{-}\mathtt{IN}) \qquad A \rightarrow \mathrm{Tr}\,\ulcorner A\urcorner$$

can be seen to be theorems of KF + COMP. The schema (Tr-IN) guarantees that, unlike what happens in KF + CONS, all axioms of KF + COMP are deemed true by the theory. However, the theory's defining axiom COMP has instances that are *provably false*. We have seen that $\mathrm{Tr}\,\ulcorner\lambda\urcorner$ is provable in KF + COMP. This entails, within KF + COMP, the sentence $\mathrm{Tr}\,\ulcorner\neg\mathrm{Tr}\,l\urcorner \wedge \mathrm{Tr}\,\ulcorner\neg\mathrm{Tr}\,\underset{\cdot}{\neg} l\urcorner$ (the second conjunct employs KF9 and KF4). By compositionality, this entails

$$(2) \qquad \mathrm{Tr}\,\ulcorner\neg(\neg\neg\mathrm{Tr}\,l \vee \mathrm{Tr}\,\underset{\cdot}{\neg} l)\urcorner,$$

which expresses that the instance of COMP involving the Liar sentence is deemed false (i.e. its negation is true) by the theory. Finally, KF+COMP regards its rule of inference to fail to preserve truth. Since KF+COMP proves $\mathrm{Tr}\,\ulcorner\neg\lambda\urcorner$, it also proves $\mathrm{Tr}\,\ulcorner\neg\lambda\vee 0\neq 0\urcorner$. However, KF2 entails that $\neg\mathrm{Tr}\,\ulcorner 0\neq 0\urcorner$. This simply means that the negation of the instance

$$\mathrm{Tr}\,\ulcorner\lambda\urcorner \wedge \mathrm{Tr}\,\ulcorner\lambda\rightarrow 0\neq 0\urcorner \rightarrow \mathrm{Tr}\,\ulcorner 0\neq 0\urcorner.$$

of (1) is provable in KF + COMP.

2.2. **Paradox, Semantics, and Incompleteness.** As it is argued in Fischer et al. (2021), KF (and variants theoreof) is a theory of *scientific* truth. The theoretical status of KF, they argue, depends entirely on the success of its applications outside logic (broadly construed). This is unlike nonclassical theories that are fully characterized by its logical property of intersubstitutivity between A and $\mathrm{Tr}\,\ulcorner A\urcorner$; these theories have a different epistemological status because their constitutive principles are conceptually necessary. We will assume this view of KF in what follows. Qua theories of scientific truth, it is plausible to analyze KF + CONS and KF + COMP by means of standard measures of theoretical equivalence. This is what will be done in the next section. In this subsection we recall some theoretical contexts in which KF and its variants have played a significant role.

The origins of KF trace back to Feferman's predicativist view in the philosophy of mathematics. The limits of predicativity had been already investigated by means of the ramified analytical hierarchy in the sixties by Feferman himself and Schütte (Feferman, 1964; Schütte, 1965). Feferman (1991) provides a non-hierarchical framework to capture the limits of predicativity given the natural numbers. The result is a version of KF endowed with a special substitution rule – dubbed $\mathrm{Ref}^*(\mathrm{PA}(P))$ in Feferman (1991). The theory KF is the simpler version of the reflective closure of PA – i.e. in Feferman's view, the theory capturing the $\mathscr{L}_{\mathbb{N}}$-statements that are implicit in the acceptance of PA –, and it is equivalent to the union of all $\alpha < \varepsilon_0$ iterations of Tarskian truth predicates over PA. This ordinal ε_0 is not ad hoc: it's the supremum of the ordinals that can be proved to be well-founded in the theory whose reflective closure is investigated, i.e. PA. KF then elegantly captures by means of a single, self-applicable truth predicate, the iterations of Tarskian truth along the wellordering that are licensed by the base theory PA.

Kripke-Feferman truth can play a role in articulating the formal properties of a *semantic* notion of truth. According to Michael Glanzberg (see e.g. Glanzberg (2015)), our implicit grasp of the semantic properties of a language can be made explicit by an act of reflection, involving an explicit characterization of the notion of truth for the language. One such characterization is the formulation of the theory KF + CONS. Glanzberg countenances hierarchies of KF + CONS-axioms to model the open-ended nature of the act of reflecting on one's implicit grasp of the semantic properties of a language. Glanzberg's perspective can be seen as the semanticist's framework to study languages endowed with a self-applicable truth predicate.[4] This entails that a semantic theorizing may reveal that it is not KF + CONS, but rather KF + COMP the best theory to model truth ascription.

Kripke-Feferman truth has been employed to provide a diagnosis of the Liar paradox. Reinhardt (1986) famously advocated KF + CONS as a theoretical tool to uncover the truth-theoretic content of fixed-point semantics. Theorems of KF + CONS of the form $\mathrm{Tr}\ulcorner A \urcorner$ are members of the extension of the truth predicate of all consistent fixed points. Therefore, one may employ KF + CONS as an efficient reasoning tool to uncover truths without giving in to the clumsiness of nonclassical conditionals. Somewhat similarly, Maudlin (2004) argued that KF + CONS is the basis of a theory of

[4] This nice way of describing the the semantic role of KF is due to Johannes Stern.

truth and permissible assertability. Roughly speaking, theorems of KF + CONS such as the Liar sentence that are not true are nonetheless assertible. The view is then completed by a careful consideration of the norms relating truth and assertability that are compatible with such formal properties.

Besides Feferman's study of the implicit commitment of formal theories, the phenomenon of incompleteness of mathematical theories prompted more daring philosophical questions. One such question is whether the incompleteness theorems show that the human mind can be mechanized. Roger Penrose formulated an interesting argument against mechanism based on the notions of truth and absolute provability Penrose (1994). Penrose wasn't careful in calibrating the exact list of principles of truth employed in the argument. This task has been recently taken up by logicians and philosophers (Koellner, 2018; Stern, 2018). In particular, Stern analyzes Penrose's argument by formalizing it in KF + CONS: he shows that mechanism can be refuted in KF + CONS, although this refutation cannot fall into the extension of the truth predicate of KF + CONS (as we have seen, this pattern is quite common in KF + CONS).

3. Theoretical Equivalence

Qua theories of scientific truth, KF + COMP and KF + COMP should be evaluated by means of standard theory choice criteria. Comparing scientific theories can also be done formally. There is a rich variety of notions of formal inter-theoretic reductions which have been studied in the philosophy of science: a comprehensive overview of such notions can be found for instance in Halvorson (2019). In what follows we will apply some standard notions of inter-theoretic reductions to KF+CONS and KF+COMP. This section contains the necessary background.

Given first-order theories T and W, a *relative translation* τ of $\mathscr{L}_T$ into $\mathscr{L}_W$ – formulated in a relational signature – can be described as a pair (δ, F) where δ is a $\mathscr{L}_W$-formula with one free variable – the domain of the translation – and F is a (finite) mapping that takes n-ary relation symbols of $\mathscr{L}_T$ and returns formulas of $\mathscr{L}_W$ with n free variables. The description of the translation τ is completed, modulo suitable renaming of bound variables, by the following inductive clauses:

- $(R(x_1, ..., x_n))^\tau :\leftrightarrow F(R)(x_1, ..., x_n)$;
- τ commutes with propositional connectives;
- $(\forall x\ A(x))^\tau\ :\leftrightarrow\ \forall x\ (\delta(x) \rightarrow A^\tau)$.

Definition 1. *An interpretation K is specified by a triple (T, τ, W), where τ is a translation of $\mathscr{L}_T$ in $\mathscr{L}_W$, such that for all formulas $\varphi(x_1, .., x_n)$ of $\mathscr{L}_T$ with the free variables displayed, we have:*

$$\textit{if } T \vdash \varphi(x_1, ..., x_n), \textit{ then } W \vdash \bigwedge_{i=1}^{n} \delta_K(x_i) \rightarrow \varphi^\tau.$$

We will write $K : T \longrightarrow W$ for 'K is an interpretation of T in W'. Model-theoretically, a $K\colon T \longrightarrow W$ provides a method for constructing, in any model $\mathscr{M} \vDash W$, an internal model $\mathscr{M}^K \vDash T$.

T and W are said to be *mutually interpretable* if there are interpretations $K\colon T \longrightarrow W$ and $L\colon W \longrightarrow T$.

Given $\tau_0\colon \mathscr{L}_T \longrightarrow \mathscr{L}_W$ and $\tau_1\colon \mathscr{L}_W \longrightarrow \mathscr{L}_V$, the composite of $K = (T, \tau_0, W)$ and $L = (W, \tau_1, V)$ is the interpretation $L \circ K = (T, \tau_1 \circ \tau_0, V)$, where $\delta_{L \circ K}(x) \;{:}{\leftrightarrow}\; \delta_K^L(x) \wedge \delta_L(x)$.

Two interpretations $K_0, K_1\colon T \longrightarrow W$ are *equal* if W, the target theory, proves this. In particular, one requires,

$$W \vdash \forall x \, (\delta_{K_0}(x) \leftrightarrow \delta_{K_1}(x)),$$
$$W \vdash \forall \vec{x} \, (R_{K_0}(\vec{x}) \leftrightarrow R_{K_1}(\vec{x})), \qquad \text{for any relation symbol } R \text{ of } \mathscr{L}_T.$$

Mutual interpretability is arguably a good measure of consistency strength, but it does not capture finer grained relations between theories. As it is well-known, it does not even differentiate between sound and unsound theories (e.g. $\mathtt{PA}$ and $\mathtt{PA}+\mathtt{Con}(\mathtt{PA})$ are mutually interpretable).

The notion of intertranslatability is a much stricter notion of theoretical equivalence. As shown in Visser (2006), it preserves many formal property of theories such as $\varkappa$-categoricity, finite axiomatizability.

Definition 2 (intertranslatability). *U and V are intertranslatable if and only if there are interpretations $K\colon U \longrightarrow V$ and $L\colon V \longrightarrow U$ such that V proves that $K \circ L$ and* id_V *– the identity interpretation on V – are equal and, symmetrically, U proves that $L \circ K$ is equal to* id_U.

In the philosophy of science the notion of *definitional equivalence*, a notion akin to intertranslatability, has played a prominent role (starting at least with Glymour (1970)). Two theories U and V – again for simplicity, we assume a finite relational signature – are definitionally equivalent if they have a common definitional extension. A definitional extension of a theory U is simply a theory in a new language featuring, alongside the axioms of U, explicit definitions of the new relation symbols not in $\mathscr{L}_U$.[5]

Definitional equivalence is more rigid than intertranslatability in the following sense: whereas for U and V featuring disjoint signatures, the two notions coincide,[6]

[5] For a precise definition, see (Halvorson, 2019, Def. 4.6.15).

[6] This is a folklore result. For a proof, see (Halvorson, 2019, Thm. 4.6.17, Thm. 6.6.21).

this is not so with theories sharing some part or all of their signature. An example is before our very eyes.

Observation 1. KF + COMP *and* KF + CONS *cannot be definitionally equivalent.*

In general, since a definitional extension of U and V includes both U and V, if U and V are mutually inconsistent, then they cannot have a common definitional extension. Intertranslatability is compatible with mutually inconsistent theories. We will in fact show that KF + CONS and KF + COMP are intertranslatable.

Some additional discussion of the relationship between intertranslatability and definitional equivalence is contained in the final section.

4. Duality Theorems and Theoretical Equivalence

That a relation of duality exists between KF + CONS and KF + COMP has been already noticed by Cantini (1989), which is the first systematic study of variations of the basic theory KF from Feferman (1991).[7] Informally, this duality consists in the fact that the truth predicate of KF + COMP (resp. KF + CONS) can be understood within KF + CONS (KF + COMP) as the predicate 'it's not false'. In semantic terms, given a consistent fixed point S, one can define a predicate $\neg\varphi \notin S$ – i.e. *φ is not determinately false* – to isolate a class of sentences satisfying the KF + COMP truth axioms. Symmetrically, given a complete fixed point R, the set $\{\varphi \mid \neg\varphi \notin R\}$ isolates the class of sentences of $\mathscr{L}_{\mathrm{Tr}}$ that are classically true (i.e. whose negation is not determinately true or glutty). As such, this set satisfies the truth axioms of KF + CONS.

The following is a precise statement corresponding to the informal picture above. We insert some detail of the proof mainly to adapt it to the abstract approach to translations and interpretations introduced in the previous section on theoretical equivalence. It should be understood as part of the proof of Proposition 1.

Lemma 1 (Cantini (1989)). KF + CONS *and* KF + COMP *are mutually interpretable.*

Proof. Let $\tau\colon \mathscr{L}_{\mathrm{Tr}} \longrightarrow \mathscr{L}_{\mathrm{Tr}}$ be such that it does not relativizes quantifiers, commutes with propositional connectives, leaves arithmetical vocabulary unchanged, and assignes $\neg\mathrm{Tr}\,\underset{\cdot}{\neg}x$ to $\mathrm{Tr}\,x$. In other words, τ is specified by:

$$\begin{array}{ll}
\delta_\tau := x = x & \mathrm{F}_\tau(\mathrm{Tr}) := \neg\mathrm{Tr}\,\underset{\cdot}{\neg}x \\
\mathrm{F}_\tau(R) := R(x_1,\dots,x_n) & \text{for each } R \in \mathscr{L}_{\mathbb{N}}. \\
\tau(\neg A) :\leftrightarrow \neg\tau(A) & \tau(A \wedge B) :\leftrightarrow \tau(A) \wedge \tau(B)
\end{array}$$

[7] Dates of publications of the references just given may be misleading: although Cantini's paper precedes Feferman's, the latter was circulating as a draft since the early 80's. Field also considers the duality phenomenon in a slightly different setting, without reference to these classic papers (Field, 2008, Ch. 7).

$$\tau(\forall x A) :\leftrightarrow \forall x \tau(A)$$

It is worth emphasizing that the translation does not act internally on codes as well, so that there is no need to employ more sophisticated tools such as Kleene's Recursion Theorem.

We let

$$\mathtt{K} = (\mathtt{KF} + \mathtt{CONS}, \tau, \mathtt{KF} + \mathtt{COMP}) \qquad \mathtt{L} = (\mathtt{KF} + \mathtt{COMP}, \tau, \mathtt{KF} + \mathtt{CONS}).$$

To verify that K and L are indeed interpretations, we treat the key cases of the axioms for truth ascriptions and CONS/COMP.

We start with K, and verify that the translation of KF4 is provable in KF + COMP. Reasoning in KF + COMP,

$$\begin{aligned}
(\mathrm{Tr}\ulcorner\neg\mathrm{Tr}\,\dot{x}\urcorner)^\tau &\leftrightarrow \neg\mathrm{Tr}\,\underset{\cdot}{\neg}\ulcorner\neg\mathrm{Tr}\,\dot{x}\urcorner \\
&\leftrightarrow \neg\mathrm{Tr}\ulcorner\mathrm{Tr}\,\dot{x}\urcorner && \text{by } \mathtt{KF9} \\
&\leftrightarrow \neg\mathrm{Tr}\,\underset{\cdot}{\neg}\underset{\cdot}{\neg}x && \text{by } \mathtt{KF3}, \mathtt{KF9} \\
&\leftrightarrow (\mathrm{Tr}\,\underset{\cdot}{\neg}x)^\tau && \text{def. of } \mathtt{K}
\end{aligned}$$

Similarly, for KF3 we have:

$$\begin{aligned}
(\mathrm{Tr}\ulcorner\mathrm{Tr}\,\dot{x}\urcorner)^\tau &\leftrightarrow \neg\mathrm{Tr}\,\underset{\cdot}{\neg}\ulcorner\mathrm{Tr}\,\dot{x}\urcorner \\
&\leftrightarrow \neg\mathrm{Tr}\,\underset{\cdot}{\neg}x && \text{by } \mathtt{KF4} \\
&\leftrightarrow (\mathrm{Tr}\,x)^\tau && \text{def. of } \mathtt{K}
\end{aligned}$$

The arguments above do not employ COMP, so the verification that KF3 and KF4 hold in KF + CONS via L is essentially the same modulo the notational differences. We turn to CONS. Reasoning again in KF + COMP: $(\mathrm{Tr}\,\underset{\cdot}{\neg}x)^\tau$ is $\neg\mathrm{Tr}\,\underset{\cdot}{\neg}\underset{\cdot}{\neg}x$. By KF9, this entails $\neg\mathrm{Tr}\,x$. By COMP, we obtain $\mathrm{Tr}\,\underset{\cdot}{\neg}x$ and therefore $\neg\neg\mathrm{Tr}\,\underset{\cdot}{\neg}x$, which is simply $(\neg\mathrm{Tr}\,x)^\tau$. We have thus shown

$$(\mathrm{Tr}\,\underset{\cdot}{\neg}x \to \neg\mathrm{Tr}\,x)^\tau.$$

Within KF + CONS, we assume $(\neg\mathrm{Tr}\,x)^\tau$, that is $\neg\neg\mathrm{Tr}\,\underset{\cdot}{\neg}x$. By logic and CONS, we obtain $\neg\mathrm{Tr}\,x$. KF9 gives us $\neg\mathrm{Tr}\,\underset{\cdot}{\neg}\underset{\cdot}{\neg}x$, which is simply $(\mathrm{Tr}\,\underset{\cdot}{\neg}x)^\tau$. Therefore:

$$(\neg\mathrm{Tr}\,x \to \mathrm{Tr}\,\underset{\cdot}{\neg}x)^\tau.$$

qed

By inspecting the proof above, one realizes that the argument is independent from the choice of the the non-logical schemata employed in the truth theories. This enables one to employ the same argument to obtain the next corollary.

Corollary 1.

(i) $\mathtt{KFI} + \mathtt{CONS}$ *and* $\mathtt{KFI} + \mathtt{COMP}$ *are mutually interpretable.*

(ii) $\mathtt{KF} \restriction \mathscr{L}_{\mathbb{N}} + \mathtt{CONS}$ *and* $\mathtt{KF} \restriction \mathscr{L}_{\mathbb{N}} + \mathtt{COMP}$ *are mutually interpretable.*

Lemma 1 and Corollary 1 are based on interpretations that do not relativize quantifiers, and the leave all vocabulary other than the truth predicate unchanged. These interpretations belong to a specific kind that have been recently dubbed *relative truth definitions* by Fujimoto (2010). Relative truth definitions preserve the arithmetical theorems. As such, a mutual truth definability between two theories entails the identity of their $\mathscr{L}_{\mathbb{N}}$ theorems. Identity of $\mathscr{L}_{\mathbb{N}}$-theorems has historically been considered to be the most important notion of reduction between truth systems, especially in connection with Feferman's programme discussed in the previous section. Such a measure is less relevant when a system of truth is studied in relation to a specific solution to the semantic paradoxes, or to a specific conception of truth. Relative truth definability goes beyond mere proof-theoretic equivalence in that it compares fine-grained properties of truth predicates by keeping the underlying syntax theory fixed, and it is certainly more suited for conceptual reductions of truth predicate.

We now turn to the main claim of the section. $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ are equivalent in a much stricter sense than the one given by truth definitions. The interpretations $\mathtt{K}$ and $\mathtt{L}$ given above are inverse to each other, provably in $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$. This witnesses the intertranslatability of the two theories. Since $\mathtt{K}$ and $\mathtt{L}$ are truth-definitions, the claim entails that $\mathtt{KF}+\mathtt{CONS}$ and $\mathtt{KF}+\mathtt{COMP}$ are mutually truth-definable. That intertranslatability given by truth-definitions is a properly stricter notion than mutual truth definability follows from results in Nicolai (2017): the theories $\mathtt{KF}$ and $\mathtt{PUTB}$ over a finitely axiomatizable theory such as $\mathtt{EA}$ or $\mathtt{I}\Sigma_1$ are mutually truth definable but not intertranslatable.

Proposition 1. $\mathtt{KF} + \mathtt{CONS}$ *and* $\mathtt{KF} + \mathtt{COMP}$ *are intertranslatable.*

Proof. The proof is strictly speaking by induction on the length of the proofs in $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ to prove, respectively, that

$$\mathtt{KF} + \mathtt{CONS} \vdash A \leftrightarrow A^{\mathtt{L}\circ\mathtt{K}},$$

$$\mathtt{KF} + \mathtt{COMP} \vdash A \leftrightarrow A^{\mathtt{K}\circ\mathtt{L}}.$$

However, to verify that $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ are intertranslatable, it suffices to check that the interpretations K and L commute in the required sense for primitive predicates of $\mathscr{L}_{\mathrm{Tr}}$. By abusing of notation for the sake of readability, I write $\mathtt{K}$ and $\mathtt{L}$ instead of τ for the translation as well.

The case of arithmetical relations is trivial in both directions and we omit it: both $\mathtt{K}$ and $\mathtt{L}$ behave like the identity interpretation on arithmetical vocabulary.

The interesting case concerns the verification (i) that the interpretation $\mathtt{L} \circ \mathtt{K}$ behaves like the identity intepretation in $\mathtt{KF}+\mathtt{CONS}$ on Tr, and (ii) that the interpretation $\mathtt{K} \circ \mathtt{L}$ behaves like the identity interpretation in $\mathtt{KF} + \mathtt{COMP}$ on Tr.

We start with (i):

$$\begin{aligned}
(\mathrm{Tr}\, x)^{\mathtt{L}\circ\mathtt{K}} &\leftrightarrow \neg\mathrm{Tr}^{\mathtt{L}} \underset{\cdot}{\neg} x && \text{By def. of } \mathtt{K} \\
&\leftrightarrow \neg\neg\mathrm{Tr}\,\neg\neg x && \text{def. of } \mathtt{L} \\
&\leftrightarrow \mathrm{Tr}\, x && \text{logic and } \mathtt{KF9}
\end{aligned}$$

In the third line, $\mathtt{KF9}$ is employed. By inverting the intepretations (i.e. starting with $\neg\mathrm{Tr}^{\mathtt{L}\circ\mathtt{K}} \underset{\cdot}{\neg} x$), we obtain (ii), that is the desired equivalence within $\mathtt{KF} + \mathtt{COMP}$ as well.

The induction step is also immediate by induction hypothesis, given that the composition of interpretations $\mathtt{L} \circ \mathtt{K}$ (resp. $\mathtt{K} \circ \mathtt{L}$) respects logical vocabulary in a uniform way. *qed*

It is useful take stock and paraphrase what goes on in intertranslatability result. The duality theorem for $\mathtt{KF} + \mathtt{COMP}$ and $\mathtt{KF} + \mathtt{CONS}$ tell us that each theory can reproduce the truth predicate of the other by means of a new predicate obtained by combining their primitive truth predicate *and a combination of classical (external) negation and nonclassical (internal) negation*. This is enough to guarantee the proof-theoretic equivalence of the two systems in several respects: the mutual interpretability result entails that the two systems have equal consistency strength; the fact that the translation τ is in fact a truth definition in the sense explained above entails that the two theories prove the same $\mathscr{L}_{\mathbb{N}}$-sentences. However, truth definability, let alone mutual interpretability, is not enough for theoretical equivalence. Mutually truth-definable theories may substantially diverge in a spectacular amount of theoretical properties (Nicolai, 2017).

The intertranslatability of $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ reveals that the relationships between the two theories are in fact much stricter. The situation can be visualized in figures 1 and 2. In the former we are living in an arbitrary fixed point model of $\mathtt{KF} + \mathtt{CONS}$ (for simplicity, one can think about it as the minimal fixed point of the operator Φ considered in §2), and consider the set of codes of sentences of $\mathscr{L}_{\mathrm{Tr}}$. The light red triangle labelled $\mathtt{T}$ represents the (consistent) extension of the truth predicate, i.e. the sentences that are determinately true. The white triangle represents the sentences that are determinately false. In the light blue space, all other sentences of

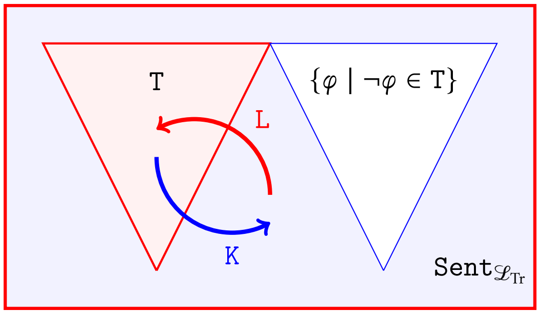
T
$\{\varphi \mid \neg\varphi \in \mathtt{T}\}$
L
K
$\mathtt{Sent}_{\mathscr{L}_{\mathrm{Tr}}}$

FIGURE 1. Intertranslatability in fixed point models of KF + CONS.

$\mathscr{L}_{\mathrm{Tr}}$, including "ungrounded" sentences such as the Liar sentence λ. The interpretation K shifts the extension of the truth predicate to

$$\mathrm{Tr}^{\mathtt{K}} = \mathtt{T} \cup \{\psi \mid \psi \notin \mathtt{T} \cup \{\varphi \mid \neg\varphi \in \mathtt{T}\}\},$$

that is to the sentences that are determinately true and not determinately false. The predicate $\mathrm{Tr}^{\mathtt{K}}$ satisfies the axioms of KF + COMP. However, the key to the intertranslatability result is contained in the next step, when one interprets back the newly obtained truth predicate $\mathrm{Tr}^{\mathtt{K}}$ via L. In fact, L returns the predicate:

$$\mathrm{Tr}^{\mathtt{L}\circ\mathtt{K}} = \{\varphi \mid \neg\varphi \notin \mathrm{Tr}^{\mathtt{K}}\},$$

that is the set of sentences whose negation is neither determinately true nor indeterminate. Since we are reasoning about a truth predicate governed by the KF + CONS axioms, the set $\mathrm{Tr}^{\mathtt{L}\circ\mathtt{K}}$ is simply the set T.

The situation with KF + COMP is symmetric and is represented in Figure 2. We are

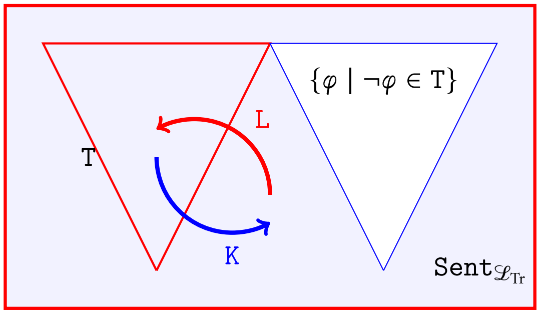
T
$\{\varphi \mid \neg\varphi \in \mathtt{T}\}$
L
K
$\mathtt{Sent}_{\mathscr{L}_{\mathrm{Tr}}}$

FIGURE 2. Intertranslatability in fixed point models of KF + COMP.

now working on the set of codes of sentences of $\mathscr{L}_{\mathrm{Tr}}$ within a fixed-point model of KF + COMP (it's useful to think about the model obtained by starting with the set of all sentences over the standard model $\mathbb{N}$ and excluding sentences by iterating the operator Φ and taking intersections at limit stages). The extension $\mathtt{T}$ of the truth predicate is now the entire light blue space: everything but the determinately false sentences. The set $\mathrm{Tr}^{\mathtt{L}}$ now gives us the set of determinately true sentences, but by applying $\mathtt{K}$ to $\mathrm{Tr}^{\mathtt{L}}$ we obtain the original extension $\mathtt{T}$. All this is carried out within $\mathtt{KF} + \mathtt{COMP}$.

By inspection of the proofs above, we notice that the induction schema $\mathtt{IND}(\mathscr{L}_{\mathrm{Tr}})$ does not play a key role: the proof only rests on the fact that $\mathtt{KF}+\mathtt{CONS}$ and $\mathtt{KF}+\mathtt{COMP}$ both feature $\mathtt{IND}(\mathscr{L}_{\mathrm{Tr}})$. Therefore, we have:

Corollary 2.

(i) $\mathtt{KFI} + \mathtt{CONS}$ *and* $\mathtt{KFI} + \mathtt{COMP}$ *are intertranslatable.*

(ii) $\mathtt{KF} \restriction \mathscr{L}_{\mathbb{N}} + \mathtt{CONS}$ *and* $\mathtt{KF} \restriction \mathscr{L}_{\mathbb{N}} + \mathtt{COMP}$ *are intertranslatable.*

The combination of Observation 1 and Proposition 1 provide us with another example of a pair of theories sharing part of their signature that are intertranslatable but not definitionally equivalent. Other, simpler examples are known. For instance, one can consider the theories in classical predicate logic $\{\forall x Px\}$ and $\{\forall x \neg Px\}$ in the signature $\{P\}$. By interpreting P as $\neg P$, one obtains a mutual interpretability result *and* the intertranslatability of the two theories. However, since the two theories are mutually inconsistent, they cannot be definitionally equivalent. Unlike those simple examples, Proposition 1 involves rich, non ad hoc theories that have been employed in several theoretical contexts (cf. §2.2).

5. A Dilemma?

The observations contained in the previous section are prima facie puzzling. As explained in §2.1, the two theories $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ formalize two different conceptions of truth. Yet, if one follows standard practice in the philosophy of science, and considers intertranslatability as a good (albeit rather strict) measure of the theoretical equivalence of the two theories, one should arrive at the rather surprising conclusion that $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ are equivalent for all theoretical purposes. This appears to be incorrect: after all, $\mathtt{KF} + \mathtt{CONS}$ states there are no truth value gluts, but that there may be sentences that are neither true nor false. $\mathtt{KF} + \mathtt{COMP}$ drastically disagrees, and states that any sentence is either true or false, and that occasionally it can be both.

One option is to reject $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ as viable theoretical alternatives. According to this line of thought, what the theoretical equivalence of $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$ shows is that the two theories are *failed* attempts to capture, respectively, a partial or inconsistent truth predicate in classical logic. This would be by no means the first critical assessment of KF + CONS and KF + COMP. Several authors attribute to KF and its variants a form of incoherence – for instance Field (2008), chapter 6, and Horsten (2012), chapter 9. It is incoherent to assert (prove) a sentence, and assert (prove) that it isn'true; it is incoherent to assert the negation of a sentence, a yet to assert its truth; it is incoherent to assert (prove) a disjunction, whose disjunct are both incoherent. As discussed in §2.1, this is what happens in KF + CONS, KF + COMP, and KF respectively.

Many of these critics, including Field and Horsten, are happy to give up classical logic to overcome this incoherence. In fact, by realigning the internal logic of the truth predicate and the external logic of the theory of truth, the asymmetries between provability and truth disappear. For instance, one can construct axiomatizations of fixed point semantics in the style of Halbach and Horsten's PKF (Halbach and Horsten, 2006) in which A and $\mathrm{Tr}\ulcorner A \urcorner$ are interderivable, and whose logic is either the internal logic of KF + CONS, Strong Kleene Logic, or the internal logic LP of KF + COMP. It is clear that, for quite trivial reasons, an analogue of Proposition 1 is not immediately available for such nonclassical systems. The very notion of relative interpretation is not devised to compare theories in different logics. Surely the truth systems would have the same $\mathscr{L}_{\mathbb{N}}$-consequences, but nothing like the strict correspondence given by intertranslatability would be available.[8]

The purported incoherence of KF + CONS and KF + COMP, however, should be weighed against the cost of giving up well-established logical principles. The adoption of a nonclassical logic impacts directly on contexts in which classical logic is traditionally undisputed, such as mathematics and its applicability to scientific theorizing (Williamson, 2018). To mention a familiar example, we would like to apply mathematical induction to properties involving the notion of truth itself. This obvious task is severely impeded if, say, we move from a classical theory such as KF+CONS to its nonclassical version in Strong Kleene logic. A significant amount of inductive reasoning is lost by adopting a nonclassical logic (Halbach and Horsten, 2006; Halbach and Nicolai, 2018).

We favour a different option. We can accept KF + CONS and KF + COMP as legitimate theoretical options, and claim that there isn't anything deeply problematic in embracing the theoretical equivalence of KF+CONS and KF+COMP, once the notion of theoretical equivalence at stake is clarified. However, the strange case of KF + CONS and KF + COMP tells us something about the scope and limits of formal notions of theoretical equivalence for theories of logical concepts such as truth.

[8] Although, given the duality of the consequence relation between LP and K3, some nonstandard notion of theoretical equivalence for nonclassical logics may not be difficult to devise.

Natural notions of theoretical equivalence can be linearly arranged on the basis of their strictness. On one side, we find the strictest notion of equivalence, logical equivalence, followed by the looser notion of definitional equivalence (which, we have seen, coincide with our notion of intertranslatability under some plausible assumptions). On the looser end, we find mere consistency (i.e. consistent theories are all equivalent, and true[9]), arguably followed by mutual interpretability. There is much in between, and we refer to systematic studies on the topic for a comprehesive overview (Halvorson, 2019; Visser, 2006); these intermediate notions are not immediately relevant to our discussion. Halvorson convincingly argues that the choice of the right notion of theoretical equivalence is highly purpose relative, and it should be the outcome of suitable philosophical work (Halvorson, 2019, Ch. 8). The philosopher's job is precisely the one of finding good reasons to choose among the sophisticated formal alternatives provided by logical and mathematical work on reductions between formal theories.

Is intertranslatability the right notion to deem KF + CONS and KF + COMP (and variants thereof) theoretically equivalent? One one sense of 'theoretical equivalent', the answer is positive. The KF + CONS-theorist can define a predicate, $\mathrm{Tr}^{\mathtt{K}}$ (cf. §4), satisfying the axioms of KF + COMP, *and* verify that the definition of its own truth predicate by the KF + COMP-theorist given by the intepretation L returns *precisely* its own truth predicate. The KF + COMP-theorist can do the same by inverting the roles of the interpretations. In other words, each theorist not only can define in a natural way the other's truth predicate, but they can also see that the other's truth definition is a faithful one, returning their own concept of truth. Whatever theoretical purpose one is pursuing within KF + CONS (resp. KF + COMP), the intertranslatability of the two theories guarantees that this can be achieved in KF+COMP (KF+CONS) by a trustworthy definition of truth. This applies to any argument given in the two theories, including the derivations witnessing the peculiar behaviour of the Liar sentence in the two theories: for instance, KF + CONS can reproduce the proof of $\mathrm{Tr}^{\mathtt{K}}\ulcorner\lambda\urcorner$ in KF + COMP, and understand that this derivation is nothing else than its own derivation of λ – i.e. of $\neg\mathrm{Tr}^{\mathtt{L}\circ\mathtt{K}}\ulcorner\lambda\urcorner$.

However, if theoretical equivalence should mean that the two theories feature an equivalent truth predicate, things change. We have seen that KF + CONS is about a consistent, partial truth predicate, and KF + COMP is about an inconsistent, complete truth predicate. If the exact nature of the truth predicate of the two theories is at stake,

[9] This view is attributed to Putnam by Halvorson (Halvorson, 2019, p. 274), and called Zenonian equivalence.

even the strict notion of intertranslatability is bound to fail to deliver the required equivalence.

An analogy with standard mathematical theories may help. It is known that Peano Arithmetic and Finite Set Theory – more precisely, ZF minus infinity plus its negation and the sentence 'every set has a transitive closure' – are intertranslatable.[10] That each theory can (mutually) faithfully reproduce the inferential structure of the other is guaranteed by their intertranslatability. This is not to say, though, that the two theories are about the same subject matter, or that the concept of (finite) set is the same as the one natural number. This may sound trivial, but it isn't. Mathematicians are usually happy to identify isomorphic structures, and intertranslatability gives us isomorphism of models and much more. For instance, intertranslatability entails that in any structure satisfying the axioms of PA, there lives a universe of finite sets which contains, definably in PA, exactly the natural number structure we started with (similarly for Finite Set Theory). However, finer-grained considerations involving, for instance, the aboutness relation connecting a formal theory to its subject matter, or the nature of the basic concepts underlying some formal construction, are not captured by the intertranslatability relation as they are not usually relevant for mathematical theorizing – so that PA and Finite Set Theory may be considered to be equivalent for mathematical purposes. By constrast, such considerations are often central in philosophical debates.

If one seeks the formal counterpart of a *conception of truth*, the right notion of theoretical equivalence is closer to what Halvorson calls Heraclitean equivalence: theories should be identified if the are logically equivalent. Very strict notions of formal theoretical inter-reducibility (let alone structure-isomorphism) such as intertranslatability are not a sufficient criterion for conceptual equivalence. Otherwise, in light of Proposition 1, we would need to identify truth predicates that are based on clearly incompatible conceptions of truth. However, this does not entail that looser criteria of theoretical equivalence should not be employed, even in the analysis and comparison of the conceptions of truth (and other broadly logical concepts) captured by some formal theories. Such criteria proved already to be useful, especially to establish negative results about philosophical reductions of truth predicates to each other and to other logical notions such as higher-order quantifiers (Nicolai, 2017, 2021).

[10] Qualifications are in order here: Kaye and Wong (2007) show that Peano Arithmetic and ZF minus infinity plus its negation and the claim 'every set has a transitive closure' are bi-interpretable, or weakly intertranslatable in the terminology of Halvorson (2019). Since the interpretation is identity preserving, by a result of Albert Visser and Harvey Friedman, the two theories are intertranslatable (Visser and Friedman, 2014).

References

Cantini, A. (1989). Notes on formal theories of truth. *Zeitschrift für Logik un Grundlagen der Mathematik*, 35:97–130.

Cieśliński, C. (2017). *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge University Press.

Enderton, H. B. (2001). *A mathematical introduction to logic*. Elsevier.

Feferman, S. (1964). Systems of predicative analysis. *Journal of Symbolic Logic*, 29:1–30.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, 56: 1–49.

Field, H. (2008). *Saving truth from paradox*. Oxford University Press, Oxford.

Fischer, M., Horsten, L., and Nicolai, C. (2021). Hypatia's silence: Truth, justification, and entitlement. *Noûs*, 55(1):62–85.

Fujimoto, K. (2010). Relative truth definability of axiomatic truth theories. *Bulletin of symbolic logic*, 16(3):305–344.

Glanzberg, M. (2015). Complexity and hierarchy in truth predicates. In *Unifying the philosophy of truth*, pages 211–243. Springer.

Glymour, C. (1970). Theoretical realism and theoretical equivalence. In *PSA: Proceedings of the biennial meeting of the philosophy of science association*, volume 1970, pages 275–288. D. Reidel Publishing.

Hájek, P. and Pudlák, P. (2017). *Metamathematics of first-order arithmetic*, volume 3. Cambridge University Press.

Halbach, V. (2014). *Axiomatic theories of truth. Revised edition*. Cambridge University Press.

Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's theory of truth in partial logic. *Journal of Symbolic Logic*, 71: 677–712.

Halbach, V. and Nicolai, C. (2018). On the costs of nonclassical logic. *Journal of Philosophical Logic*, 47:227–257.

Halvorson, H. (2019). *The logic in philosophy of science*. Cambridge University Press.

Horsten, L. (2012). *The Tarskian Turn*. MIT University Press, Oxford.

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

Koellner, P. (2018). On the question of whether the mind can be mechanized, ii: Penrose's new argument. *The Journal of Philosophy*, 115(9):453–484.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72:690–712.

Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*. Oxford University Press.

Nicolai, C. (2017). Equivalences for truth predicates. *The Review of Symbolic Logic*, 10(2):322–356.

Nicolai, C. (2021). Fix, express, quantify: Disquotation after its logic. *Mind*, 130(519):727–757.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Reinhardt, W. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, 15:219–251.

Schütte, K. (1965). Predicative well-orderings. In *Studies in Logic and the Foundations of Mathematics*, volume 40, pages 280–303. Elsevier.

Stern, J. (2018). Proving that the mind is not a machine? *Thought: A Journal of Philosophy*, 7(2):81–90.

Visser, A. (2006). Categories of theories and interpretations. In *Logic in Tehran*, volume 26, pages 284–341. Assoc. Symbolic Logic La Jolla, Calif.

Visser, A. and Friedman, H. (2014). *Logic Group Preprint Series*, 320.

Weatherall, J. O. (2019). Part 1: theoretical equivalence in physics. *Philosophy Compass*, 14(5):e12592.

Williamson, T. (2018). Alternative logics and applied mathematics. *Philosophical Issues*, 28(1):399–424.