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, KF + CONS, features a consistent but partial truth predicate. The second, KF + COMP, an inconsistent but complete truth predicate. It is 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 KF+CONS and KF+COMP raises a puzzle: the two theories can be proved to be strictly related, yet they appear to embody remarkably different conceptions of truth. We discuss the significance of the result for the broader debate on formal criteria of conceptual reducibility for theories of truth.

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).

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 axioms. These systems may have multiple aims: they may embody some conception of truth, including a solution to the difficulty posed by paradox; they may 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 this paper we contribute to 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. This question has already been investigated in other works (Halbach, 2000; Fujimoto, 2010; Nicolai, 2017); this paper considers a new case study. We focus in particular on the case 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 KF + CONS, whose truth predicate is consistent but partial (not every sentence is true or false). The second is the theory KF + COMP, whose truth predicate is inconsistent and complete (every sentence is either true or false). In the light of these differences, it would seem implausible to consider KF + CONS and KF + COMP as theoretically equivalent theories of truth.

Yet, in §4, we will show that KF + CONS and KF + COMP are intertranslatable, and therefore strong candidates for being indeed theoretically equivalent theories of truth. In §2, we will introduce Kripke-Feferman truth and some of the key properties of the truth predicates of KF + CONS and KF + COMP. In §3, we introduce the notions of theoretical reducibility an equivalence that will be employed in the paper. §4 contains the main technical observations: the well-known phenomenon of the duality between KF + CONS and KF + COMP will be recalled, and we will show that it can be lifted to the intertranslatability of the two theories. Finally, in §5, we provide a philosophical assessment of the formal result: we discuss what it means for a formal theory to capture a conception of truth, and isolate different conceptions of truth compatible with the Kripke-Feferman theories that support diverging verdicts on the intertranslatability of KF + CONS and KF + COMP.

2. Kripke-Feferman truth

The system KF is formulated in the language $\mathcal{L}_{\mathrm{Tr,F}}$ obtained by extending the language $\mathcal{L}_{\mathbb{N}}$ of arithmetic with unary predicates Tr and F for truth and falsity. It is convenient to formulate $\mathcal{L}_{\mathbb{N}}$ 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)). To formulate the truth axioms, we will make use of the expression $\dot{x}$ standing for the code of a constant symbol c_x associated with any x. This is justified by the existence of PA-definable primitive recursive injective functions sending each number x to a representative (the code of a constant symbol) c_x. In practice, although PA is formulated in a purely relational language $\mathscr{L}_{\mathrm{Tr,F}}$, the truth axioms are formulated for a "coded" language $\mathscr{L}^{+}_{\mathrm{Tr,F}}$ that contains constants $\dot{x}$ for each number x. We will not distinguish between $\mathscr{L}_{\mathrm{Tr,F}}$ and $\mathscr{L}^{+}_{\mathrm{Tr,F}}$ in what follows.

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

(KF0) $\quad \forall x((\mathrm{Tr}x \rightarrow \mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}(x)) \wedge (\neg\mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}(x) \rightarrow \mathrm{F}x))$

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

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

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

(KF4) $\quad \forall x(\mathrm{F}\ulcorner \mathrm{Tr}\dot{x}\urcorner \leftrightarrow \mathrm{F}x \leftrightarrow \mathrm{Tr}\ulcorner \mathrm{F}\dot{x}\urcorner)$

(KF5) $\quad \forall\varphi\forall\psi(\mathrm{Tr}(\varphi \dot{\wedge} \psi) \leftrightarrow (\mathrm{Tr}\varphi \wedge \mathrm{Tr}\psi))$

(KF6) $\quad \forall\varphi\forall\psi(\mathrm{F}(\varphi \dot{\wedge} \psi) \leftrightarrow (\mathrm{F}\varphi \vee \mathrm{F}\psi))$

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

(KF8) $\quad \forall v\forall\varphi(\mathrm{F}(\dot{\forall} v\varphi) \leftrightarrow \exists y\, \mathrm{F}\varphi(\dot{y}/v))$

(KF9) $\quad \forall\varphi((\mathrm{F}\varphi \leftrightarrow \mathrm{Tr}\dot{\neg}\varphi) \wedge (\mathrm{F}\dot{\neg}\varphi \leftrightarrow \mathrm{Tr}\varphi))$

In KF5-KF9, the quantification $\forall\varphi$... abbreviates $\forall x(\mathtt{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}(x) \rightarrow$ ... Other conventions follow Halbach (2014), including the under-dotting convention for syntactic functions. Notice that the formulation of KF9 with unrestricted quantification (not over sentences only) is inconsistent with KF1.

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

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

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

[1] In constrast to the other axioms, the schema I-$\mathtt{IND}(\mathscr{L}_{\mathrm{Tr,F}})$ is presented in non-abbreviated form (not quantifying over "sentences" directly). We preferred this presentation to make more explicit the restriction of unary formulae.

The theory $\mathtt{KF} \restriction \mathscr{L}_{\mathbb{N}}$ 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 to it, respectively, the axioms

(CONS) $\qquad \forall x(\mathrm{F}x \rightarrow \neg \mathrm{Tr}x),$

(COMP) $\qquad \forall x(\neg \mathrm{Tr}x \rightarrow \mathrm{F}x).$

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

KF can be seen as an axiomatization of Kripke's fixed point semantics Kripke (1975). Let $\Phi\colon \mathscr{P}(\omega)^2 \longrightarrow \mathscr{P}(\omega)^2$ be the arithmetical operator associated with the Kripke truth and falsity sets (Cantini, 1989, Def. 5.3).[2] A fixed point $X = (X^+, X^-)$ of Φ – with $X^+ \subseteq \mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}$ and $\omega \setminus X^- \subseteq \mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}$ – is *consistent* if $X^+ \cap X^- = \varnothing$; a *complete* fixed point is such that $X^+ \cup X^- = \omega$. Standard models of KF + CONS are precisely the consistent fixed points, and standard models of KF + COMP are the complete ones:

Fact 1 (Feferman).

(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.** The truth and falsity predicates of KF + CONS and KF + COMP diverge in some significant respects.[3] 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 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 Fl, for which the identity $l = \ulcorner \mathrm{F}l \urcorner$ is provable in PA. If λ, i.e. Fl, then also Trl by KF4. This would contradict CONS, therefore $\neg\lambda$. Similarly, if Tr$\dot{\neg}l$, then Fl by KF9, contradicting CONS again by KF4: therefore $\neg \mathrm{Tr}\dot{\neg}l$. We have then proved $\neg\lambda \wedge \neg \mathrm{Tr}\ulcorner\neg\lambda\urcorner$ in KF+CONS. This phenomenon extends to instances of some axioms of KF + CONS, such as the instance $\mathrm{F}l \rightarrow \neg \mathrm{Tr}l$. of CONS. They are declared not true by the theory: assuming $\mathrm{Tr}(\ulcorner\neg \mathrm{F}l \vee \neg \mathrm{Tr}l\urcorner)$ in KF+CONS, by distributing the truth predicate on the disjunction and applying KF9 one gets $\mathrm{FF}l \vee \mathrm{FTr}l$: however, each disjunct contradicts CONS by the previous reasoning. A noticeable feature of KF + CONS is that it derives a universally quantified version of

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

[3] For simplicity, we will often omit reference to falsity and talk only about truth in informal discussion.

the modal axiom K:

(K) $$\forall\varphi\forall\psi(\mathrm{Tr}\varphi \wedge \mathrm{Tr}(\varphi \dot{\rightarrow} \psi) \rightarrow \mathrm{Tr}\psi).$$

The proof of (K) in KF+CONS employs CONS and 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 $\mathrm{Tr}\psi$ by compositionality. The first option contradicts CONS, while the second gives us $\mathrm{Tr}\psi$. Either way, (K) follows. The sentence (K) may be seen – under some specific choices of the logical calulus – as formalizing the assertion that the (classical) rules of inference of KF + CONS are truth preserving. This is adequate, for instance, if 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 $\mathrm{Tr}l$, then also $\mathrm{F}l$ by the above reasoning; if $\neg\mathrm{Tr}l$, then $\mathrm{F}l$ by COMP and therefore $\mathrm{Tr}l$ as well. By classical logic, $\mathrm{Tr}l \wedge \mathrm{F}l$. By a straightforward induction on the complexity of the sentence A of $\mathscr{L}_{\mathrm{Tr,F}}$, all instances of the schema

(Tr-IN) $$A \rightarrow \mathrm{Tr}\ulcorner A \urcorner$$

are 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 $\mathrm{F}l \wedge \mathrm{Tr}l$ is provable in KF + COMP. By KF3, 4 and 9 this entails $\mathrm{F}\ulcorner\neg\neg\mathrm{Tr}l\urcorner \wedge \mathrm{F}\ulcorner\mathrm{F}l\urcorner$. By compositionality, we obtain

$$\mathrm{F}\ulcorner(\neg\neg\mathrm{Tr}l \vee \mathrm{F}l)\urcorner.$$

Finally, under the specific choice of the logical calculus considered above, KF + COMP may be seen to regard its rule of inference to fail to preserve truth. Since KF + COMP proves $\mathrm{Tr}l$, it also proves $\mathrm{Tr}\ulcorner\neg\mathrm{F}l\urcorner$ and therefore $\mathrm{Tr}\ulcorner\neg\lambda \vee 0 \neq 0\urcorner$. However, KF2 entails that $\neg\mathrm{Tr}\ulcorner 0 \neq 0\urcorner$. This is 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 (K).

3. Theoretical Equivalence

There is a rich variety of formal notions of inter-theoretic reduction which have been studied in logic and the philosophy of science: a comprehensive overview of such notions can be found for instance in Visser (2006) and Halvorson (2019). In what

follows we will apply some of these notions to KF + CONS and KF + COMP. This section contains the necessary background.

Given first-order theories T and W, a *relative translation* τ of $\mathscr{L}_T$ in $\mathscr{L}_W$ – 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;
- $(\forall x\ A(x))^\tau \ :\leftrightarrow\ \forall x\ (\delta(x) \to A^\tau)$.

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:*

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

For readability, we will often write δ_K for the domain of the interpretation K and φ^K for the translation of φ according to K. We will also write $K : T \to W$ for 'K is an interpretation of T in W'. Model-theoretically, a $K\colon T \to W$ 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 $K\colon T \to W$ and $L\colon W \to T$.

Given $\tau_0\colon \mathscr{L}_T \to \mathscr{L}_W$ and $\tau_1\colon \mathscr{L}_W \to \mathscr{L}_V$, 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 $K_0, K_1\colon T \to W$ 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. PA and PA + ¬Con(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 properties of theories such as κ-categoricity and finite axiomatizability.

Definition 2 (intertranslatability). *U and V are intertranslatable if and only if there are interpretations $K\colon U \to V$ and $L\colon V \to U$ 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.

By the definition of equality of interpretations and the completeness theorem, U and V are intertranslatable if and only if there are interpretations $K\colon U \to V$ and $L\colon V \to U$ such that: for any model $\mathscr{M} \vDash U$, we have (verifiably in $\mathscr{M}$), that $\mathscr{M} = \mathscr{M}^{K\circ L} = (\mathscr{M}^L)^K$, and for any $\mathcal{N} \vDash V$, $\mathcal{N}$ verifies that $\mathcal{N} = \mathcal{N}^{L\circ K} = (\mathcal{N}^L)^K$. We will occasionally speak of intertranslatability and bi-interpretability for pairs of models as well.

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 $\mathcal{L}_U$.[4]

Definitional equivalence is more rigid than intertranslatability in the following sense: whereas for U and V featuring disjoint signatures, the two notions coincide,[5] this is not so for 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.

4. Theoretical Equivalence for Kripke-Feferman Truth

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).[6] Informally, this duality consists in the fact that the truth predicate of KF + COMP (resp. KF + CONS) can be understood within KF + CONS

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

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

[6] 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 (Field, 2008, Ch. 7).

(KF + COMP) as the predicate 'it's not false', and that the falsity predicate of KF + CONS (resp. KF + COMP) becomes in KF + COMP (KF + CONS) the predicate 'it's not true'. But even more is the case: if one replaces 'false' in 'it's not false' with 'not true', one obtains 'it's not not true', which leads us back to 'it's true' – and symmetrically with falsity.

The situation can be visualized in figures 1 and 2. In the former we are living in the

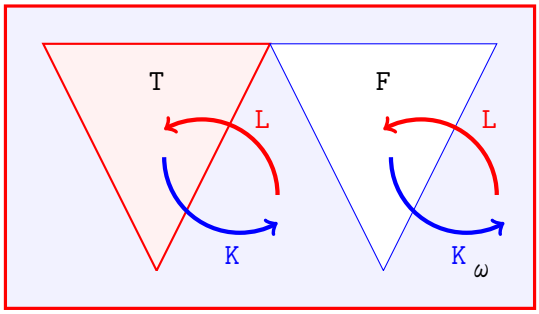

FIGURE 1

minimal fixed-point model of KF + CONS (i.e. the minimal fixed point of the operator Φ considered in §2). The light red triangle 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 (including the non-sentences). In the light blue space one finds all other sentences of $\mathscr{L}_{\mathrm{Tr,F}}$, including "ungrounded" sentences such as the Liar sentence λ. The model interprets Tr and F as T and F, respectively. We can define a mapping K – corresponding to a syntactic interpretation to be defined shortly – that behaves like the identity mapping on the denotations of primitives of $\mathscr{L}_{\mathbb{N}}$, and such that

$$\mathtt{T}^{\mathtt{K}} = \omega \setminus \mathtt{F}, \qquad\qquad \mathtt{F}^{\mathtt{K}} = \omega \setminus \mathtt{T}.$$

The "predicates" $\mathtt{T}^{\mathtt{K}}$ and $\mathtt{F}^{\mathtt{K}}$ satisfy the axioms of KF + COMP. However, one can interpret back the newly obtained predicates $\mathtt{T}^{\mathtt{K}}$ and $\mathtt{F}^{\mathtt{K}}$ via the mapping L, which behave exactly as K:[7]

$$\mathtt{T}^{\mathtt{L}\circ\mathtt{K}} = \omega \setminus \mathtt{F}^{\mathtt{K}} = \mathtt{T}, \qquad\qquad \mathtt{F}^{\mathtt{L}\circ\mathtt{K}} = \omega \setminus \mathtt{T}^{\mathtt{K}} = \mathtt{F}.$$

The situation with KF + COMP is symmetric and is represented in Figure 2. We are now working in what is often referred to as the the maximal fixed point of Φ over $\mathbb{N}$ – which can be defined as the model obtained by starting with the extensions of truth and falsity given by the pair $(\mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr,F}}}, \omega)$ over the standard model $\mathbb{N}$ and

[7]Notice that, given the definition of interpretation give earlier, two interpretations may be based on the same translation and yet differ.

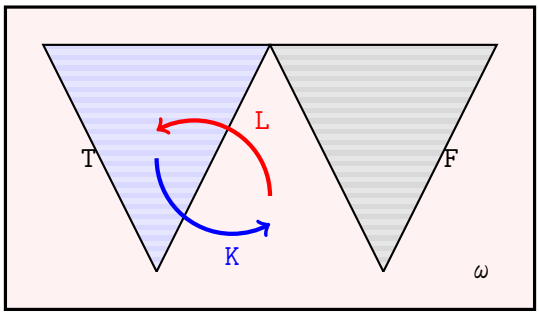

FIGURE 2

iterating the operator Φ while taking intersections at limit stages. The extensions $\mathtt{T}$ of the truth predicate corresponds now to the light red and blue spaces: everything but the determinately false sentences. The extension $\mathtt{F}$ of F coincides with the light red and gray spaces. The set $\mathtt{T}^{\mathtt{L}}$ now gives us the set of determinately true sentences (the complement of $\mathtt{F}$), but by applying $\mathtt{K}$ to $\mathrm{Tr}^{\mathtt{L}}$ we obtain the original extension $\mathtt{T}$, and symmetrically for $\mathtt{F}$.

These informal considerations essentially amount to the following.

OBSERVATION 2. *Let $\mathcal{J}$ be the minimal fixed point of Φ and $\mathcal{J}^*$ its dual. Then, the models $(\mathbb{N}, \mathcal{J})$ and $(\mathbb{N}, \mathcal{J}^*)$ are bi-interpretable, and in fact intertranslatable.*

The fact that standard models of theories are bi-interpretable, or even intertranslatable, does not suffice for establishing the bi-interpretability or intertranslatability of the corresponding theories. For instance, $\mathtt{V}_\omega$ and $\mathbb{N}$ are indeed bi-interpretable, but $\mathtt{ZF} \setminus \mathtt{Inf} + \neg\mathtt{Inf}$ – where $\mathtt{Inf}$ is the axiom of infinity – and $\mathtt{PA}$ are not bi-interpretable (Enayat et al., 2011, Thm. 5.1).

However, $\mathtt{KF}$ + $\mathtt{CONS}$ and $\mathtt{KF}$ + $\mathtt{COMP}$ are indeed intertranslatable. Cantini (1989) claims that $\mathtt{KF}$ + $\mathtt{CONS}$ and $\mathtt{KF}$ + $\mathtt{COMP}$ are mutually interpretable. The interpretations involved in the claim, however, do not relativize quantifiers, and the leave all vocabulary other than the truth predicate unchanged. These interpretations belong to a specific kind that has 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 theorems in $\mathcal{L}_{\mathbb{N}}$. Historically, identity of $\mathcal{L}_{\mathbb{N}}$-theorems has has played a central role in the study of inter-theoretic reduction between truth systems, especially in connection with Feferman's predicativist programme discussed below. 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 predicates. We can restate Cantini's claim in terms of truth-definitions.

Lemma 1 (Cantini (1989)). KF + CONS *and* KF + COMP *are mutually truth-definable.*

Proof. Let $\tau\colon \mathscr{L}_{\mathrm{Tr,F}} \to \mathscr{L}_{\mathrm{Tr,F}}$ be specified by:[8]

$$\delta(x) := x = x \qquad (R(x_1,\dots,x_n))^\tau := R(x_1,\dots,x_n) \qquad \text{for each } R \in \mathscr{L}_{\mathbb{N}}.$$
$$(\mathrm{Tr}x)^\tau := \neg \mathrm{F}x \qquad (\mathrm{F})^\tau := \neg \mathrm{Tr}x$$
$$\tau(\neg A) :\leftrightarrow \neg\tau(A) \qquad \tau(A \wedge B) :\leftrightarrow \tau(A) \wedge \tau(B)$$
$$\tau(\forall x A) :\leftrightarrow \forall x \tau(A)$$

We let

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

The verification that K and L are interpretations requires a standard induction on the length of the proofs of the relevant theories.

qed

By inspecting the proof above, one realizes that the argument is independent from the choice of 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) KFI + CONS *and* KFI + COMP *are mutually truth-definable.*

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

We now turn to the main claim of the section. KF + CONS and KF + COMP are equivalent in a much stricter sense than the one given by truth definitions. The interpretations K and L given above are inverse to each other, provably in KF + CONS and KF + COMP. This witnesses the intertranslatability of the two theories. That intertranslatability given by truth-definitions is a properly stricter notion than mutual truth definability follows from results in Nicolai (2017): the theories KF and PUTB (cf. Halbach (2014) for a definition) over a finitely axiomatizable theory such as EA or $\mathtt{I}\Sigma_1$ are mutually truth definable but not intertranslatable.

[8] 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.

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

Proof. Given our definition of intertranslatability, it suffices to check that the interpretations $\mathtt{K}$ and $\mathtt{L}$ commute in the required sense for primitive predicates of $\mathscr{L}_{\mathrm{Tr,F}}$. 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 (including the trivial domains) is trivial in both directions and we omit it. We verify (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.

$(\mathrm{Tr}x)^{\mathtt{L}\circ\mathtt{K}}$ is $\neg \mathrm{F}^{\mathtt{L}}x$ by definition of $\mathtt{K}$; $\mathtt{L}$ then gives us $\neg\neg \mathrm{Tr}x$, which is obviously logically equivalent to $\mathrm{Tr}x$. Similarly, $\mathrm{F}^{\mathtt{L}\circ\mathtt{K}}x$ is just $\neg \mathrm{Tr}^{\mathtt{L}}$, that is $\neg\neg \mathrm{F}x$, which is equivalent to $\mathrm{F}x$. Since $\mathtt{K}$ and $\mathtt{L}$ are both based on τ, the same equivalence are obtained by inverting the roles of $\mathtt{K}$ and $\mathtt{L}$. Therefore we have:

$$\mathtt{KF} + \mathtt{CONS} \vdash (\mathrm{Tr}^{\mathtt{L}\circ\mathtt{K}} \leftrightarrow \mathrm{Tr}x) \wedge (\mathrm{F}^{\mathtt{L}\circ\mathtt{K}}x \leftrightarrow \mathrm{F}x)$$
$$\mathtt{KF} + \mathtt{COMP} \vdash (\mathrm{Tr}^{\mathtt{K}\circ\mathtt{L}}x \leftrightarrow \mathrm{Tr}x) \wedge (\mathrm{F}^{\mathtt{K}\circ\mathtt{L}}x \leftrightarrow \mathrm{F}x)$$

as desired. *qed*

Remark 1. The result is somewhat dependent on the chosen axiomatization of $\mathtt{KF}$. While the argument above goes through for Feferman's axiomatization of $\mathtt{KF}$ – i.e. with unrestricted $\mathtt{KF9}$ and without $\mathtt{KF0}$ –, it does not immediately go through for the axiomatization of $\mathtt{KF}$ dispensing with the falsity predicate (e.g. the one in Halbach (2014)). To give a sense of the issue, in defining $\mathrm{Tr}x$ as $\neg \mathrm{Tr}\dot{\neg}x$, an argument analogous to the one in Prop. 1 tells us that we would require the equivalence between $\mathrm{Tr}\dot{\neg}\dot{\neg}x$ and $\mathrm{Tr}x$. However, this would only hold with the contextual information that $\mathrm{Sent}_{\mathscr{L}_{\mathrm{Tr}}}(x)$. However, in order to enforce this contextual information, we would require an axiom stating that only sentences are true, which would then create problems to obtain Lemma 1.

By inspection of the proofs above, we notice that the induction schema $\mathtt{IND}(\mathscr{L}_{\mathrm{Tr,F}})$ 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 $\mathtt{IND}(\mathscr{L}_{\mathrm{Tr,F}})$. 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.*

Remark 2. Proposition 1 and Corollary 2 can be reformulated in stronger forms. The interpretations witnessing the intertranslatability of the theories are in fact $\mathscr{L}_{\mathbb{N}}$-preserving. Nicolai (2021) introduced the notions of $\mathscr{L}_{\mathbb{N}}$-bi-interpretability and $\mathscr{L}_{\mathbb{N}}$-intertranslatability, which replace relative interpretations with $\mathscr{L}_{\mathbb{N}}$-interpretations in the definitions of bi-interpretability and intertranslatability. Working on $\mathscr{L}_{\mathbb{N}}$-interpretations has the advantage that the notions become directly comparable with mutual truth-definability (in fact, properly stricter). By a theorem of Visser and Friedman (2014), $\mathscr{L}_{\mathbb{N}}$-bi-interpretability and $\mathscr{L}_{\mathbb{N}}$-interpretability coincide.

To summarize, the duality theorem for KF + COMP and KF + CONS tells us that each theory can reproduce the truth and falsity predicates of the other by means of a new predicate obtained by combining their primitive truth and falsity predicates with classical (external) 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. The intertranslatability of KF + CONS and KF + COMP reveals that the relationships between the two theories are in fact much stricter.

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, as we shall now see.

5. KF-truth and the significance of theoretical equivalence

The observations contained in the previous section are prima facie puzzling. KF + CONS and KF + COMP appear to formalize different notions of truth: KF + CONS states there are no truth value gluts, but that there may be sentences that are neither true nor false. KF + COMP drastically disagrees, and states that any sentence is either true or false, and that occasionally it can be both. Yet, the theories are intertranslatable, and therefore formally equivalent in a strict sense. In the present section, we discuss this seemingly paradoxical situation in the context of a more general analysis of the role of notions of theoretical equivalence for theories of truth.

Natural notions of theoretical equivalence can be linearly arranged on the basis of their strictness. One one side, we find the strictest notion of equivalence, definitional equivalence (which, we have seen, coincide with our notion of intertranslatability under some plausible assumptions).[9] On the looser end, we find mere consistency (i.e. consistent theories are all equivalent, and true[10]), 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. A parallel but less structured hierarchy involves theories formulated in signatures that extend $\mathscr{L}_{\mathbb{N}}$ and in which the interpretation of the arithmetical vocabulary is kept fixed. To this hierarchy belong mutual truth-definability, and the properly stricter notion of $\mathscr{L}_{\mathbb{N}}$-intertranslatability (cf. Remark 2).

Is is difficult to determine the philosophical import of notions of theoretical equivalence. It is so in the case of formalization of theories from the natural sciences, and it's even more so in the case of logico-mathematical theories that lack something like the distinction between theoretical and empirical vocabulary.[11] The significance of results on inter-theoretic equivalence is bound to be purpose-relative.[12] In certain contexts, strong notions of theoretical isolate equivalence classes of theories / models which represent the right kind of objects for all theoretical purposes. In certain areas of mathematical logic, bi-interpretability appears to be sufficient: according to the model theorist Pillay, for instance, 'the business of "pure" model theory becomes the classification of first-order theories up to bi-interpretability' (Buss et al., 2001, p. 186). At the same time, it's also clear that identification of structures up to (definable) isomorphism may be too crude; for instance, despite the definitional equivalence of theories of finite sets and arithmetic, the view that natural numbers and finite sets are irreducible structures can be coherently defended (Tait, 2015).

5.1. **Conceptions of Truth and Reducibility.** Since we are mainly interested in discussing formal theoretical equivalence in the case of theories of truth such as KF + CONS and KF + COMP, one would ideally require that formal theoretical equivalence could preserve the *conception of truth* embodied by a theory. A conception of

[9]We omit logical equivalence in the general overview as it only applies to theories with the same signature.

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

[11]Although such a distinction may be hard to motivate (Van Fraassen, 1980).

[12]See for instance (Halvorson, 2019, Ch. 8), for a similar conclusion.

truth may be defined as a collection of conditions that the notion of truth should possess. These conditions, of course, may be of different kind;[13] some of them may be constitutive of the notion – without them, one would not even be talking about truth –, others may be compatible with someone disagreeing with them and yet reasonably possessing a notion of truth (conception-specific). The literature offers several such conditions. The following, non-exhaustive list of criteria, for instance, builds on Leitgeb (2007), Halbach and Horsten (2015), and Nicolai (2017): *material adequacy* (the theory of truth should entail the full Tarski-biconditionals for the truth-free language), *preservation of ontological commitments* (theories of truth should allow for standard interpretations of the objects of truth), *compositionality* (truth (and falsity) should commute with the logical connectives of the internal logic in fully quantified form), *generalizing power* (truth should enable one to establish desirable generalizations, such as the truth of the base theory), *consistency* (no sentence and its negation should be true), *completeness* (either a sentence or its negation should be true), *finite axiomatizability* (the truth axioms should be given by a finite list), *no type restrictions* (the theory of truth should prove genuine instances of self-application).

Nicolai (2017) employed such criteria to calibrate the formal notions of theoretical equivalence to the comparison of theories of truth. The project starts with the observation that some of the criteria just listed, such as compositionality and finite axiomatizability (over the base theory), are not prima facie preserved by some natural candidate for conceptual equivalence of truth predicates, such as mutual truth-definability. For instance, the theory of type-free disquotational truth PUTB – see (Halbach, 2014, §-19.3-19.5) – and KF are mutually-truth definable: KF interprets PUTB via the identity interpretation, whereas PUTB employs the diagonal lemma and the truth predicate to mimic the KF-truth axioms. However, PUTB is a disquotational, infinitely axiomatized theory, whereas KF is compositional and features a finite set of truth axioms over PA in $\mathcal{L}_{\mathrm{Tr,F}}$. In Nicolai (2017) it is shown that bi-interpretability, and therefore intertranslatability, are more sensitive than mutual truth-definability to the criteria above; PUTB and KF fail to be bi-interpretable over a finitely axiomatized base theory.[14] Proposition 1 shows that even intertranslatability is not sensitive to some of the criteria above; consistency and completeness of truth are not preserved by intertranslatability. From this one may be tempted to conclude that even the strictest notion of theoretical equivalence fails to preserve a conception of truth.

[13]Compare the discussion of the concept and conceptions of set in (Incurvati, 2020, Ch.1).

[14]To obtain a failure of over the infinitely axiomatized PA, one needs to employ what Nicolai (2021) calls $\mathcal{L}_{\mathbb{N}}$-bi-interpretability and $\mathcal{L}_{\mathbb{N}}$-intertranslatability, which imposes the usual requirements for bi-interpretability and intertranslatability, respectively, on relative truth-definitions (i.e. interpretations based in $\mathcal{L}_{\mathbb{N}}$-preserving translations) instead of simple relative interpretations. See Remark 2.

The scope of this conclusion, however, is limited. The problem is that truth-theorists may starkly disagree on which criteria amount to a conception, and which do not. Undoubtedly, material adequacy in the sense above is constitutive of truth, but already compositionality in the strong form considered in the list is not uncontroversially recognized as a constitutive feature. For instance, transparency is taken by some theorists to be the only constitutive feature of truth.[15] But compositionalilty in fully quantified form is not necessary for transparency; a transparent truth predicate guarantees compositionality for each, externally given sentence, but not for all sentences in the sense of object-language quantifiers. An obvious reaction would be to impose that formal theoretical inter-reducibility ought to preserve only what is uncontroversially constitutive of truth. However, it is an upshot of our discussion that this risks simply recreating the debate on acceptable criteria at this more fundamental level. For instance, preservation of material adequacy does not even discriminate between primitive and non-primitive truth predicates, and would allow for weak notions of inter-theoretic reducibility to capture theoretical equivalence. The Tarskian Tr-schema for $\mathcal{L}_{\mathbb{N}}$-sentences, for instance, is mutually interpretable with PA.

To avoid such shortcomings, one can resort to the requirement – in line with what we called 'preservation of ontological commitments' in the list above – that quantifiers over the objects to which truth applies should not be relativized.[16] This would certainly rule out mutual interpretability as a conception-preserving notion of theoretical equivalence, while suggesting that mutual truth-definability is much better suited to preserve constitutive features of a conception of truth. That, as we have just seen, may also be controversial without further qualification. Compositionality, completeness, and consistency are not preserved by mutual-truth definability, yet they are strong candidates for being constitutive features of truth. Analogously, the distinction between type-free and typed theories of truth is not preserved by mutual truth-definability: for instance, finite iterations of Tarksian (typed) truth are mutually truth-definable with a suitable subsystem of Friedman-Sheard (type-free) truth allowing only finitely many applications of the necessitation rule.[17] To fix this, one may ascend even higher in the hierarchy inter-theoretic reduction, and only require bi-interpretability or intertranslatability to preserve constitutive features of truth. That as well, we have seen, may be controversial, as paraconsistent and paracomplete theories of truth diverge precisely on matters of consistency and completeness that are, by Proposition 1, not discriminated by such notions.

[15] See, for instance, Field (2008); Beall (2009).

[16] Actually, it may be more suitable to allow for a bi-interpretation (i.e. provable isormorphism) of two base theories.

[17] See (Halbach, 2014, §14).

Fortunately, the discussion does not need to proceed only at such a general level. Instead of quantifying over all possible conceptions of truth, we can discuss what the intertranslatability of KF + CONS and KF + COMP can tell us about specific conceptions underlying the main applications of Kripke-Feferman truth. Before this, we will discuss a position that undermines the significance of Proposition 1 by rejecting KF + CONS and KF + COMP as viable theoretical options.

5.2. **Rejecting KF + CONS and KF + COMP.** The theoretical equivalence of KF + CONS and KF + COMP could be used as the conclusion of a reductio argument. One could accept that intertranslatability is a good notion of theoretical equivalence for theories of truth, and employ the intertranslatability of KF + CONS and KF + COMP as showing that the two theories are *failed* attempts to truly capture, respectively, a partial or inconsistent truth predicate in classical logic. This would be in line with other assessments 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, and 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.[18]

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

[18]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.

traditionally undisputed, such as mathematics and its applicability to scientific theorizing (Williamson, 2018). We would for instance 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 Nicolai, 2018; Nicolai, 2022).

For theorists who regard the failure of transparency as more palatable than abandoning classical logic, KF + CONS and KF + COMP represent valid theoretical options. In fact, Kripke-Feferman truth has been employed in several theoretical contexts. It is to those that we now turn.

5.3. **Scientific Truth.** As it is argued in Fischer et al. (2021), KF (and variants theoreof) can be seen as a theory of *scientific* truth. The theoretical status of KF, they argue, depends entirely on the success of its applications outside logic. This is unlike nonclassical theories that are fully characterized logical properties such as the intersubstitutivity A with $\mathrm{Tr}^{\ulcorner} A^{\urcorner}$; these theories have a different epistemological status because their constitutive principles are conceptually necessary. They see applications of KF to the foundations of mathematics as part of the scientific uses of Kripke-Feferman truth. Here we take a different stance: scientific applications of KF will be the ones that fall outside logics and mathematics.

Kripke-Feferman truth, for instance, can be used in semantics, more specifically to articulate 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.[19] As such, Glanzberg's programme is not an a-priori enterprise, as it depends on empirical data to calibrate which articulation of Kripke-Feferman truth is more suited for the project; semantic theorizing may well reveal that it is not KF + CONS, but rather KF + COMP the best theory to model truth ascriptions. An analogous programme, using a combination of the KF axioms with other logical resources, is carried out in McGee (1991).

[19] *Omissis*.

Kripke-Feferman truth has also been employed to provide a diagnosis of the Liar paradox. Maudlin (2004) argued that KF + CONS is the basis of a theory of 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 an investigation of the norms relating truth and assertability that are compatible with such formal properties.

Since these applications of KF rely in part on empirical data, the import of Proposition 1 on theoretical equivalence touches on the broader discussion of the import of formal reducibility on the equivalence of scientific theories. On a standard neopositivist picture – where a distinction between theoretical and empirical vocabulary is assumed to be available – equivalence of observational consequences is sufficient for theoretical equivalence. Of course, one needs to spell out what this equivalence is, even if at the empirical level only. According to a sensible position, equivalence of observational consequences cannot be a matter of mere translation.[20] Translation cannot discriminate between mere notational differences – e.g. between the theories whose only axioms are 'Orchideen sind Blumen' and 'Orchids are flowers' – and semantic differences – e.g. between the theories 'Orchids are flowers' and 'Lions are cats'. In the latter case, even if the theories are intertranslatable, the operational meaning of empirical vocabulary is lost:

> ...mere commonality of logical form, even of a total theory when compared with another total theory, is certainly not by itself sufficient for theoretical equivalence. The meanings of the terms in the theories, however construed, are crucial to questions of equivalence. (Sklar, 1982, p. 12)

In other words, the operational meaning of empirical vocabulary appears to be essential for theoretical equivalence, unless one wants to adopt very strong forms of semantic holism in which the meaning of specific terms can only be fixed once the overall theory underlying all the community's usage of these empirical terms is available. The stance on empirical vocabulary proper of the view should be contrasted with the one on theoretical – e.g. purely logico-mathematical – apparatus of the theories, for which formal equivalence ensures theoretical equivalence in a full sense.

For semantic applications of KF + CONS and KF + COMP, we can reasonably assume that the truth predicate(s) of the theories correspond to the "fixed", observational component of the vocabulary. What to do with the rest, the syntactic/mathematical

[20] This argument goes back at least to Sklar (1982). For a recent exposition, see Teitel (2021)

component of the theories, is a more delicate matter. As discussed in Nicolai (2015),[21] Peano arithmetic plays a double role in traditional axiomatizations of truth. The first is a structural role of providing the basic combinatorics presupposed by speech and writing in natural and formal languages. The second is the role of an object theory in a specific language, whose sentences are apt to semantic scrutiny in the same way as truth-ascriptions are.

The conception of truth that is relevant for semantic applications of Kripke-Feferman truth is then one in which the exact patterns of truth ascriptions uncovered by the theories of truth in question matter, and should be preserved by theoretical equivalence. Each of KF+CONS and KF+COMP provides with incompatible verdicts on some fundamental semantic features of truth, and theoretical equivalence becomes only relevant after (partially) fixing the meaning of truth via one axiomatization. Since the two axiomatizations are incompatible, so are the meanings they assign to their truth predicates. However, nothing prevents one from employing strong notions of intertheoretic reducibility to compare the what corresponds to the "theoretical vocabulary" of the theories. Still, intertranslatability does not seem to determine theoretical equivalence in the semantic sense.

We have mentioned that what is commonly known as the base theory of a theory of truth can be seen as providing double role. If the base theory is seen only as a theoretical machinery for formal syntax, formal theoretical equivalence can be obtained by formally equivalent choices of this formal syntax. For instance, one can consider the theory $\mathtt{ZF}_{\mathtt{fin}}$, Zermelo-Fraenkel set theory with the axiom of infinity replaced by its negation and the axiom stating the existence of the transitive closure of any set (Kaye and Wong, 2007), or the theory HF from Świerczkowski (2003), an inductive version of adjunctive set theory with terms; KF + CONS and KF + COMP are then considered like functors applying to different base theories. Although semantic use of Kripke Feferman truth just outlined would deem KF + CONS and KF + COMP *not* theoretically equivalent despite their intertranslatability, it would certainly be compatible with the full theoretical equivalence of KF+CONS[PA], KF+CONS[$\mathtt{ZF}_{\mathtt{fin}}$], and KF + CONS[HF] (and analogously for KF + COMP[$\cdot$]).

If instead the base theory is considered in its double role, there seem to be two options. One can extend the treatment of the truth predicate to the base language as well by considering the base theory in its role of "object language". In this case, not only semantic predicates need to preserved by translations, but also the vocabulary of

[21]The discussion is anticipated in Halbach (2014) and builds on technical results in Heck (2015) and Leigh and Nicolai (2013).

the base theory; in effect, on these assumptions the right notion of theoretical equivalence would arguably amount to what Halvorson calls Heraclitean equivalence: theories should be identified if the are logically equivalent. Again, Proposition 1 would not pose a particular problem to this conception, since theoretical equivalence cuts finer than intertranslatability. A second option is to be faithful to this double role, and to consider instead Kripke-Feferman theories as binary functors on theories operating on a purely structural support theory and an object theory. The formal implementation of the proposal can follow the blueprint of the theories discussed in Leigh and Nicolai (2013). Semantic investigation would then be concerned directly with the truth and falsity predicates and with the second argument of this functor. Not only the theories $\mathtt{KF} + \mathtt{CONS}[\mathtt{PA}, U]$ and $\mathtt{KF} + \mathtt{COMP}[\mathtt{PA}, U]$ would fail to be apt for theoretical equivalence, but the same would apply to theories such as $\mathtt{KF} + \mathtt{CONS}[\mathtt{PA}, U]$ and $\mathtt{KF} + \mathtt{CONS}[\mathtt{PA}, V]$, with U and V logically inequivalent theories. However, theories that only differ in their theoretical apparatus, such as $\mathtt{KF} + \mathtt{COMP}[T, U]$ and $\mathtt{KF} + \mathtt{COMP}[S, U]$, would be candidates for theoretical equivalence; in fact, theories such as $\mathtt{KF} + \mathtt{COMP}[\mathtt{PA}, U]$ and $\mathtt{KF} + \mathtt{COMP}[\mathtt{HF}, U]$ would turn out to be theoretically equivalent on this view. All in all, the semantic conception of Kripke-Feferman truth is compatible with intertranslatability only in a very limited sense; on this view the seemingly puzzling aspects of Proposition 1 are resolved by the realization that intertranslatability is not strict enough for full semantic equivalence.

5.4. **Truth as a logico-mathematical tool.** Truth is often employed in reasoning as a generalizing device, and Kripke-Feferman truth can be employed as a purely logic-mathematical tool. Deflationists about truth believe that this is the only purpose of the truth predicate. Independently of deflationary attitudes, the logico-mathematical character of truth is widely recognized, and so is the essential role played by compositional principles such as $\mathtt{KF5} - \mathtt{KF8}$ in the generalizing power of truth.[22]

KF in fact originated as a logico-mathematical tool, in the context of 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 arithmetically

[22]See, for instance, Fujimoto (2022) for an up to date discussion of the role of compositional principles in reasoning with truth.

equivalent to the union of $\alpha < \varepsilon_0$ 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.

Another use of Kripke-Feferman truth as a logico-mathematical tool can be found in Reinhardt (1986). Reinhardt advocated KF + CONS as a theoretical tool to uncover the truth-theoretic content of fixed-point semantics, which he recognized to be a successful resolution of the Liar paradox. 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.

Besides Feferman's study of the implicit commitment of formal theories, the phenomenon of incompleteness of mathematical theories prompted more daring philosophical questions, still related to the logico-mathematical nature of truth. 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).

In the logico-mathematical conception, the generalizing power of the truth predicate is a constitutive feature of truth. The truth predicate ought to enable one to achieve deductive or expressive strength, e.g. by proving reflection principles for the base theory, or by replacing set-existence assumptions. In this picture, the non-structural, semantic component of theoretical equivalence considered in empirical applications is less relevant. Therefore, the intertranslatability of KF + CONS and KF + COMP appears to be sufficient for theoretical equivalence. After all, Proposition 1 entails that the theories have equal consistency strength, and the specific nature of the interpretations K and L, which preserve $\mathscr{L}_{\mathbb{N}}$-vocabulary, entails that the theories will have the same consequences in $\mathscr{L}_{\mathbb{N}}$. More importantly, general claims are preserved in a strong way via the result. Consider again the generalization

$$\forall\varphi\forall\psi(\mathrm{Tr}(\varphi \dot{\rightarrow} \psi) \wedge \mathrm{Tr}\varphi \rightarrow \mathrm{Tr}\psi) \tag{K}$$

stating that Tr is closed under the material conditional. We have seen that (K) is a theorem of KF + CONS but not of KF + COMP. Yet, the required generality can be recaptured in KF + COMP via

$$\forall\varphi\forall\psi(\mathrm{Tr}^{\mathrm{K}}(\varphi\dot{\rightarrow}\psi) \wedge \mathrm{Tr}^{\mathrm{K}}\varphi \rightarrow \mathrm{Tr}^{\mathrm{K}}\psi) \tag{1}$$

K leaves the structure of the objects of truth unchanged, so that the generalization is performed over a given ontology, and the interpretation only operates on the specific generalizing role that the truth predicate of KF + CONS plays in (K). A symmetric case can be made for KF + COMP and L, if one starts with the theorem of KF + COMP

$$\forall\varphi\forall\psi((\mathrm{Tr}\varphi \rightarrow \mathrm{Tr}\psi) \rightarrow \mathrm{Tr}(\varphi\dot{\rightarrow}\psi)). \tag{2}$$

Analogously, we can consider the theorem of KF + COMP stating that all instances of excluded middle in $\mathscr{L}_{\mathrm{Tr,F}}$ are true:

$$\forall\varphi(\mathtt{LEM}(\varphi) \rightarrow \mathrm{Tr}\varphi). \tag{3}$$

Even without a complete primitive truth predicate, KF + CONS again can achieve this generalizing power via

$$\forall\varphi(\mathtt{LEM}(\varphi) \rightarrow \mathrm{Tr}^{\mathrm{L}}(\varphi)). \tag{4}$$

What we just said, however, leaves open whether intertranslatability is also necessary for theoretical equivalence in the logico-mathematical conception. To mimic generalizations such as (K), (2), and (3) one requires at least the existence of a truth-definition in the sense of §4. Surely, for the reasons discussed above, mutual interpretability would not be sufficient. However, also mutual-truth definability is dubious: we mentioned above that some type-free theories of truth are mutually truth-definable with typed theories of truth. Generalizations in one theory acquire a different logical role when translated in this way: a generalization on all sentences of $\mathscr{L}_{\mathrm{Tr,F}}$ may then become a generalization on sentences that only have a certain amount of Tarskian truth predicates. This seems to compromise the role of truth as a generalizing device, because some syntactic generalizations are in fact restricted by truth-definitions.

Intertranslatability is certainly a safer option. For one thing, in intertranslatable theories the generalizing roles of the two truth predicates are deeply intertwined, in a much more explicit way than in mutually-truth definable theories. Each theory can not only mimic truth-theoretic generalizations over a common syntactic base, but it has also the means to recognize how its truth predicate is used in the other theory, and what is required to revert back to its own truth predicate. The KF + CONS-theorist can define a predicate, Tr^{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 primitive generalizing device. This is unlike what happens in mutual truth-definitions, where there may be no duality of interpretations, and the equivalence between the two kinds of generalizing tools may only be available from an external point of view. For instance, we have seen that the theory PUTB and KF are mutually-truth definable. However, the relevant interpretations are not dual, since otheriwse PUTB would prove the full compositional axioms of KF, which is not possible.[23] In the logico-mathematical conception of truth, the intertranslatability of KF + CONS and KF + COMP can be fully embraced. The consistency or completeness of truth do not constitute the raison d'être of truth, and intertranslatability preserves unequivocally the generalizing nature of the truth predicate.

References

Beall, J. C. (2009). *Spandrels of Truth*. Oxford University Press.

Buss, S. R., Kechris, A. S., Pillay, A., and Shore, R. A. (2001). The prospects for mathematical logic in the twenty-first century. *Bulletin of Symbolic Logic*, 7(2):169–196.

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.

Enayat, A., Schmerl, J., and Visser, A. (2011). ω-models of finite set theory. In *Set Theory, Arithmetic, and Foundations of Mathematics: Theorems, Philosophies*, pages 43–65. 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.

[23] For these results (the mutual-truth definability, and the impossibility of compositionality in PUTB, see (Halbach, 2014, §-19.3-19.5).

Fujimoto, K. (2022). The function of truth and the conservativeness argument. *Mind*, 131(521):129–157.

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. (2000). Truth and reduction. *Erkenntnis*, 53(1):97–126.

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 Horsten, L. (2015). Norms for theories of reflexive truth. In *Unifying the philosophy of truth*, pages 263–280. Springer.

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.

Heck, R. G. (2015). Consistency and the theory of truth. *Review of Symbolic Logic*, 8(3):424–466.

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

Incurvati, L. (2020). *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press.

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.

Leigh, G. E. and Nicolai, C. (2013). Axiomatic truth, syntax and metatheoretic reasoning. *Review of Symbolic Logic*, 6(4):613–636.

Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, 2(2):276–290.

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

McGee, V. (1991). *Truth, vagueness, and paradox*. MIT University Press.

Nicolai, C. (2015). Deflationary truth and the ontology of expressions. *Synthese*, 192(12):4031–4055.

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.

Nicolai, C. (2022). The dream of recapture. *Analysis*. Forthcoming.

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.

Sklar, L. (1982). Saving the noumena. *Philosophical Topics*, 13(1):89–110.

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

Świerczkowski, S. (2003). *Finite sets and Gödel's incompleteness theorems*.

Tait, W. W. (2015). In defense of the ideal. http://home.uchicago.edu/ wwtx/In

Teitel, T. (2021). What theoretical equivalence could not be. *Philosophical Studies*, 178(12):4119–4149.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

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.

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