

KF_μ a theory of truth and of $L_{\omega_1^{CK}}$

Luca Castaldo¹, Mateusz Łełyk², and Konstantinos Papafilippou³

¹ Ludwig Maximilians University, Munich, Germany
castaldluca@gmail.com

² University of Warsaw, Warsaw, Poland
mlelyk@uw.edu.pl

³ University of Warsaw, Warsaw, Poland
k.papafilippou@uw.edu.pl

1 Introduction

The main character of this talk is the axiomatic theory of truth over arithmetic KF_μ , first introduced by Burgess [Bur14]. This is the theory whose ‘canonical’ (intended) model is the least fixed-point of the Kripke jump operator K and axiomatically it extends KF by asserting its minimality. The ‘canonicity’ of its model is expressed in that KF_μ is a solid theory [EL24] (a categoricity-like criterion) i.e. interpretations of its models do not induce non-trivial cycles with respect to definable isomorphism. However, as observed in [FHKS15] the theory is far from uniquely determining the least fixed-point of K , even in ω models, a fact that can be seen by complexity considerations. In the words of [FHKS15], KF_μ is not \aleph -categorical. This raises the question of what the non-standard models of KF_μ look like.

To this end, we show that KF_μ is synonymous with ID_1 (i.e. they have a common definitional expansion) and with a set theory which we dub KP_s (for KP ‘small’)—corresponding to the least admissible model of KP containing ω as a set. In effect, this ‘strengthens’ the (independently proven) mutual interpretability result for ID_1 and KP by Tapp and by Fujimoto [Tap99, Fuj18] while also being applicable to their corresponding fragments (restricted minimality and resp. restricted \in -induction). This correspondence gives an indirect proof of solidity of KP_s , making it a similarly ‘canonical’ theory of $L_{\omega_1^{CK}}$ —its intended model.¹ More importantly, with this synonymy result we can use tools from admissible set theory, like the Barwise completeness and compactness theorems [Bar17], to construct ω -standard models of KF_μ with their truth predicate being non standard. As a broad reference point, we know that the ordinals of such models \mathcal{M} will follow the structure of Harrison orders [Har68] and therefore the order type of $Ord_{\mathcal{M}}$ will be of the form $\omega_1^{CK} \times (1 + \mathbb{Q})$.

If time permits, we will also compare the Burgess KF_μ with its counterpart by Cantini in [Can89], denoted $KF^+ + GI$. We show that the latter is provable in the former but leave the problem of their equivalence as an open question.

2 Preliminaries

A well known theory of truth of substantial proof theoretic strength is the Kripke Feferman Burgess theory of truth or KF_μ . It is a “self-referential” theory of truth, meaning that it has a single truth predicate which can meaningfully refer to itself (e.g. “‘It’s true that the snow is white’ is true”). The idea behind the approach taken with KF_μ is that we construct the truth predicate of KF_μ step-by-step via iterative applications of a positive (and thus monotone) operator referred to as the Kripke Jump $K(X)$. The Kripke jump forms one step of an inductive definition of the truth predicate and makes for a positive formulation of it following the ideas presented in Kripke’s outline [Kri75]. By Knaster

¹A direct proof of solidity of KP_s was found independently by Łełyk and Gruza.

Tarski, it eventually reaches a fixed-point. We consider the least one in KF_μ which, starting from the empty set, is reached after ω_1^{CK} many steps (i.e. after exhausting all computable ordinals) and is Turing equivalent to a hyperjump. Applying the Kripke Jump on a class X lets us take one approximation of the truth predicate starting from X . Thus, treating X as already some approximation of truth, the Kripke jump gives us all true atomic sentences, allows us to take one compositional step of sentences in X (e.g. from $\varphi, \psi \in X$ we will have that $\varphi \wedge \psi \in K(X)$) and furthermore it lets us state that if a sentence is true (according to X), then it's true that it's true (i.e. if $\varphi \in X$ then $T(\varphi) \in K(X)$) and if it is false (according to X), then it's false that it's true according to $K(X)$ (i.e. if $\neg\varphi \in X$ then $\neg T(\varphi) \in K(X)$).

Formal Theories

Definition 2.1 (Kripke jump). *Let $\mathcal{L}_{KF} := \mathcal{L}_{PA} + \{T\}$, where T is a unary predicate symbol, and let $\text{Sent}_{\mathcal{L}_{KF}}$ to be the set of codes of sentences of \mathcal{L}_{KF} as natural numbers. For any $X \subseteq \text{Sent}_{\mathcal{L}_{KF}}$,*

$$\begin{aligned} K(x, X) : \leftrightarrow & x \in \text{Sent}_{\mathcal{L}_{KF}} \wedge \\ & (\exists s \exists t (x = (s = t) \wedge \text{val}(s) = \text{val}(t)) \vee \exists s \exists t (x = (s \neq t) \wedge \text{val}(s) \neq \text{val}(t)) \vee \\ & \exists t (x = (Tt) \wedge \text{val}(t) \in X) \vee \exists t (x = (\neg Tt) \wedge \neg \text{val}(t) \in X) \vee \\ & \exists \varphi, \psi (x = (\varphi \wedge \psi) \wedge \varphi \in X \wedge \psi \in X) \vee \exists \varphi, \psi (x = \neg(\varphi \wedge \psi) \wedge \neg \varphi \in X \vee \neg \psi \in X) \vee \\ & \exists \varphi(v) (x = (\forall v \varphi) \wedge \forall t \varphi(t/v) \in X) \vee \exists \varphi(v) (x = \neg \forall v \varphi \wedge \exists t \neg \varphi(t/v) \in X) \vee \\ & \exists \varphi (x = (\neg \neg \varphi) \wedge \varphi \in X)) \end{aligned}$$

Observe that $K(x, X)$ is written as an arithmetical formula with a second order variable X and a first order variable x . Hence, given a formula $\psi(x)$ it makes sense to consider a substitution $K(x, X)[\psi(t)/t \in X]$ which replaces every occurrence of an atomic formula $t \in X$ with $\psi(t)$ (renaming bounded variables to avoid clashes). We shall abbreviate $K(x, X)[\psi(t)/t \in X]$ with $K(x, \psi(\hat{x}))$. Note that $K(x, \psi(\hat{x}))$ is a formula with free variable x .

Definition 2.2 (KF—Kripke-Feferman, cf. [Hal14]). *KF is the \mathcal{L}_{KF} -system extending PA (with induction on the extended language) as well as the following axiom:*

$$KF^0 \quad \forall x. K(x, T(\hat{x})) \leftrightarrow T(x).$$

The system KF_{cs} is obtained by extending KF with the axiom

$$\text{Cons} \quad \forall \varphi (T\varphi \rightarrow \neg T(\neg\varphi))$$

We recall some well-known properties of KF:

Lemma 2.3 ([Can89]).

1. *Positive T-schema: For all \mathcal{L}_{KF} -formulae $\varphi(z)$ where T occurs only positively, KF derives*

$$\forall z (\varphi(z) \leftrightarrow T^r \varphi(\dot{z})^r).$$

2. *T-Out schema: For all \mathcal{L}_{KF} -formulae $\varphi(z)$, KF_{cs} derives*

$$\forall z (T^r \varphi(\dot{z})^r \rightarrow \varphi(z))$$

where \dot{z} corresponds to the numeral coding the term $S \dots S0$ with z instances of the successor term S .

Definition 2.4 (KF_μ —Kripke-Feferman-Burgess, cf. [Hal14]). KF_μ denotes the extension of KF with the minimality scheme for T , i.e. all sentences of the form

$$\forall x (K(x, \psi(\hat{x})) \rightarrow \psi(x)) \rightarrow \forall x (T(x) \rightarrow \psi(x)),$$

for $\psi(x) \in \mathcal{L}_{KF}$. For $\Gamma \in \{\Sigma_n, \Pi_n, \Delta_n\}$, $KF_\mu \upharpoonright \Gamma$ denotes the extension of KF with the substitutional instances of the minimality scheme only with formulae from class Γ . In other words, $KF_\mu \upharpoonright \Gamma$ is a subtheory of KF_μ in which restrict the minimality scheme to Γ -formulae only.

A theory similar to KF_μ is the theory of inductive definitions which instead of having only one unique fixed-point, it instead does so for any arithmetic formula that induces a monotone operator on sets.

Definition 2.5 (ID_1 —Inductive Definitions, cf. [Poh09]). For every class of formulas Γ , let $\Gamma(\vec{X})$ be the class of formulas φ in Γ such that formulas of the form $\vec{x} \in X$ can occur as atoms for $X \in \vec{X}$. Formally, it is the least class of formulas \mathcal{Y} such that $\Gamma \subseteq \mathcal{Y}$ and for any $\varphi \in \mathcal{Y}$, any n , any n -ary $X \in \vec{X}$ and any $n + 1$ terms x, \vec{t} , the formula $\varphi[\vec{t} \in X/x = x] \in \mathcal{Y}$. We will define $\Gamma^+(\vec{X})$ to be the class of all formulas φ in $\Gamma(\vec{X})$ such that each atom $\vec{x} \in X$ occurs only positively in φ for all $X \in \vec{X}$. The language of \mathcal{L}_{ID_1} extends the language of PA with the addition of an n -ary predicate symbol I_φ for each formula $\varphi(\vec{x}, X) \in \mathcal{L}_{PA}^+(X)$, where X is an n -ary predicate variable. The theory ID_1 is axiomatised by the axioms of PA in the extended language as well as the following two axiom schemata for formulae $\varphi \in \mathcal{L}_{PA}^+(X)$ and $\psi \in \mathcal{L}_{ID_1}$:

$$ID_1^0 \quad \forall \vec{x}. \varphi(\vec{x}, I_\varphi) \leftrightarrow \vec{x} \in I_\varphi;$$

$$ID_1^1 \quad \forall \vec{x} (\varphi(\vec{x}, \psi) \rightarrow \psi(\vec{x})) \rightarrow \forall \vec{x} (\vec{x} \in I_\varphi \rightarrow \psi(\vec{x})).$$

For $\Gamma \in \{\Sigma_n, \Pi_n, \Delta_n\}$, $ID_1 \upharpoonright \Gamma$ denotes the the subtheory of ID_1 obtained by restricting the Axiom Schema ID_1^1 to formulae $\psi \in \Gamma$.

Definition 2.6. We will call a class X inductive if it is equal to some $I_\varphi^n := \{x \mid \langle x, n \rangle \in I_\varphi\}$ for some element n . Similarly we will call it coinductive if it is equal to the complement of some inductive class. Finally we say that X is hyperelementary, if it is both inductive and coinductive.

Definition 2.7 (KP). Kripke Platek set theory (for which we will implicitly assume infinity) is a subsystem of ZF consisting of the axioms Extensionality, Pair, Union, infinity and the axiom schemata of Δ_0 -Separation, Δ_0 -Collection, ω -induction² and ϵ -induction, the latter two we present below:

ϵ -induction: $\forall x (\forall y \in x \varphi(y) \rightarrow \varphi(x)) \rightarrow \forall x \varphi(x)$, for all formulae $\varphi(x)$ in which y does not occur free;

ω -induction: $\forall x \in \omega (\forall y \in x \varphi(y) \rightarrow \varphi(x)) \rightarrow \forall x \in \omega \varphi(x)$, for all formulae $\varphi(x)$ in which y does not occur free.

We write $KP \upharpoonright \Gamma$ for the theory axiomatised by the axioms of KP with ϵ -induction restricted to formulae in Γ , while ω -induction will remain unrestricted. For the purposes of this presentation, we will assume that $\Sigma_1 \subseteq \Gamma$ as it is the fragment sufficient to utilise all the usual framework developed by Barwise.

Models of KP are called admissible as are called the ordinals of the height of those models. The least admissible model including a transitive set a is denoted by $HYP(a)$ with $HYP = HYP(\omega) = L_{\omega_1^{CK}}$, where ω_1^{CK} is the smallest non-hyperarithmetical ordinal.

Definition 2.8 (KP_Σ). We consider KP_Σ as the extension of KP expressing that we live in the least model of KP , namely $L_{\omega_1^{CK}}$. Formally, $KP_\Sigma \upharpoonright \Gamma$ is just $KP \upharpoonright \Gamma$ as well as the axioms $V = L$ and $\neg \exists a (a \models KP \upharpoonright \Gamma)$.

²In the literature, ω -induction is usually considered as an added axiom when considering the fragments of KP .

Categoricity-like notions

When working with sufficiently strong theories, the usual categoricity of model theory (as well as its relevant machinery) cannot help us analyse theories like PA or ZF. This image however changes when we look at this from the scope of interpretability. In particular, we know by Visser [Vis06] that PA has the property that any two bi-interpretable extensions of it in the same language are deductively equivalent. This has led to the generation and study of a whole array of such notions (see eg. [EŁ24]) of which we will restrict ourselves here to just that of solidity.

Definition 2.9 (Interpretation, cf. [EŁ24]). *Given an \mathcal{L}_V -structure \mathcal{M} and an \mathcal{L}_U -structure \mathcal{N} , we say that \mathcal{M} (parametrically) interprets \mathcal{N} , written $\mathcal{M} \triangleright \mathcal{N}$ ($\mathcal{M} \triangleright_{par} \mathcal{N}$) iff there is a (parametric) translation $\sigma : \mathcal{L}_U \rightarrow \mathcal{L}_V$, such that $\mathcal{N} = \sigma(\mathcal{M})$.*

Interpretations between theories $V \triangleright U$ induce parameter free interpretations between their models $\mathcal{M} \triangleright \mathcal{N}$ for $\mathcal{M} \models U$ and $\mathcal{N} \models V$ and vice versa. Two theories are mutually interpretable when they interpret each other. This is contrasted by the stronger notion of bi-interpretability where we additionally require both compositions of those interpretations to be isomorphic to the identity interpretation. A notion that is essentially half of bi-interpretability is that of retraction where we only require that one of the compositions of the interpretations to be definably isomorphic to its respective identity interpretation.

Definition 2.10 (Retract, semantically, cf. [EŁ24]). *A structure \mathcal{M} is a (parametric) retract of a structure \mathcal{N} if there is an isomorphic copy \mathcal{M}^* of \mathcal{M} such that the following conditions jointly hold:*

- (i) $\mathcal{M} \triangleright_{par} \mathcal{N} \triangleright_{par} \mathcal{M}^*$;
- (ii) *there is an \mathcal{M} -definable isomorphism $f : \mathcal{M} \rightarrow \mathcal{M}^*$.*

Definition 2.11 (Synonymy). *Two theories U and V of disjoint signature are synonymous iff there is a theory W that is a definitional extension of both U and V .*

Though a more useful characterisation, in particular for the theories we are interested in, is given by Friedman & Visser in the following.

Theorem 2.12 (Friedman–Visser, see [FV25]). *Any two theories which are sequential and bi-interpretible via identity preserving interpretations are synonymous.*

Definition 2.13 (Solidity, see [Ena16, EŁ24]). *A first-order theory T is solid iff the following holds for all models $\mathcal{M}, \mathcal{M}^*$, and \mathcal{N} of T :*

If $\mathcal{M} \triangleright_{par} \mathcal{N} \triangleright_{par} \mathcal{M}^$, and if there is a parametrically \mathcal{M} -definable isomorphism $f : \mathcal{M} \rightarrow \mathcal{M}^*$, then there is a parametrically \mathcal{M} -definable isomorphism $g : \mathcal{M} \rightarrow \mathcal{N}$.*

Essentially a theory is solid if it is categorical modulo self retractions, which makes solidity a form of internal categoricity. Some solid theories are PA, ZF, Z_2 (see [Ena16]) and so is KF_μ [EŁ24]. On the contrary, the fragments of PA (eg. $I\Sigma_n$) (see [EŁ24]) are not solid and neither is Zermelo set theory Z nor KP (see [RFH21]) nor $\Pi_n^1\text{-CA}_0$ (see [RFW25]).

3 Synonymies

KF_μ is synonymous with ID_1

Lemma 3.1 (Sato, see [Sat15]). *Over $ID_1 \upharpoonright \Delta_0$, for every formula $\varphi \in \mathcal{L}_{PA}^+(X)$ there is a $(\Pi_1 \vee \Sigma_1)^+(X)$ -formula φ' such that $I_\varphi = I_{\varphi'}$.*

Lemma 3.1 is necessary to drop the formula complexity enough so that the following theorem holds under restricted minimality.

Theorem 3.2. $ID_1 \upharpoonright \Gamma$ is a definitional extension of $KF_\mu \upharpoonright \Gamma$, for any $\Gamma \supseteq \Delta_0$.

ID_1 is synonymous with KP_s

The mutual interpretability of ID_1 with KP is a result first proved by Tapp [Tap99] and later independently by Fujimoto [Fuj18]. For this synonymy we make use of the same mappings that Fujimoto uses and we show that with this interpretation, you can further induce synonymy between ID_1 and KP_s —which also holds true in their restrictions. The interpretations are as follows: Starting from a model of $KP_s \upharpoonright \Gamma$ (where $\Sigma_1 \subseteq \Gamma$) we can follow Gandy’s theorem in the context of restricted \in -induction to get the inductive predicates as least-fixed-points induced by Σ formulas in the language of KP .

Theorem 3.3 (Gandy with restricted \in -induction, see [Bar17]). *Over $KP_s \upharpoonright \Gamma$, given an arithmetical formula $\varphi(x, X) \in \mathcal{L}_{PA}^+(X)$, there is a Σ -definable fixed-point which satisfies Γ -minimality.*

For the other direction, we start from a model of ID_1 and the model of KP_s we construct is the collection of all well-founded hyper elementary trees. To express this, we make use of a sufficiently nice universal inductive set (for example, one induced by the KF_μ truth predicate) as a means of quantifying over inductive and hyper elementary sets. We show that beyond KP , the composition of the interpretations is a retraction. This is used to prove that the induced model of KP further satisfies $V = L$ and because of the retraction, we have that ω_1^{CK} cannot exist as that would have made any universal inductive set hyper elementary, contradicting its universality by an easy diagonalisation. We summarise the conclusion with the following theorem.

Theorem 3.4. $KP_s \upharpoonright \Gamma$ is synonymous with $ID_1 \upharpoonright \Gamma$ for all Γ such that $\Sigma_1 \subseteq \Gamma$.

Since it is known that KF_μ is solid, we obtain as a naturally corollary that so are ID_1 and KP_s .

4 Models of KF_μ

The analysis of the non-standard models of KF_μ has been an open question, even present in the corresponding article in the Stanford encyclopedia of philosophy [HLL25]. The synonymy result opens up a whole array of tools from the world of admissible set theory.

Lemma 4.1. *Suppose that $\mathcal{M}, \mathcal{N} \models KP_s$ are ω -standard and \mathcal{N} is an end-extension of \mathcal{M} . Then $T_{\mathcal{M}} \subseteq T_{\mathcal{N}}$ —where $T_{\mathcal{M}}$ (resp. $T_{\mathcal{N}}$) is the truth predicate of the model $\mathcal{M}' \models KF_\mu$ (resp. $\mathcal{N}' \models KF_\mu$) induced by our synonymy.*

Theorem 4.2 (Formalised Barwise Completeness). *Let $\mathcal{M} \models KP_s$. Suppose that T is a $\Sigma_1(\mathcal{M})$ -definable theory such that $\mathcal{M} \models \text{Con}(T)$. Then there is an \mathcal{M} -definable model $\mathcal{N} \models T$.*

Using Barwise completeness, we can then derive non-standard ω -models of KF_μ as well as definable infinite chains of models each of lower non-standard consistency strength. None of these models can be Π_1^1 -hard as they would otherwise compute the least fixed-point and hence fail minimality.

Theorem 4.3. *Suppose that $(\mathbb{N}, T) \models KF_\mu$. Then there is T' such that $(\mathbb{N}, T') \models KF_\mu$ but $T \subsetneq T'$.*

Theorem 4.4. *There exists an infinite sequence $X_0 \subsetneq X_1 \subsetneq X_2 \subsetneq \dots$ such that for each i , $(\mathbb{N}, X_i) \models KF_\mu$ and $X_{i+1} \in \text{Def}(\mathbb{N}, X_i)$.*

References

- [Bar17] Jon Barwise. *Admissible Sets and Structures*. Perspectives in Logic. Cambridge University Press, 2017.
- [Bur14] John P Burgess. Friedman and the axiomatization of Kripke’s theory of truth. In *Foundational adventures: Essays in honor of Harvey M. Friedman*, pages 125–148. College Publications, 2014.
- [Can89] Andrea Cantini. Notes on formal theories of truth. *Mathematical Logic Quarterly*, 35(2):97–130, 1989.
- [EŁ24] Ali Enayat and Mateusz Łelyk. Categoricity-like properties in the first order realm. *Journal for the Philosophy of Mathematics*, 1:63–98, Sep. 2024.
- [Ena16] Ali Enayat. *Variations on a Visserian Theme*, pages 99–110. 2016.
- [FHKS15] Martin Fischer, Volker Halbach, Jönne Kriener, and Johannes Stern. Axiomatizing semantic theories of truth? *The Review of Symbolic Logic*, 8(2):257–278, 2015.
- [Fuj18] Kentaro Fujimoto. Truths, inductive definitions, and Kripke-Platek systems over set theory. *The Journal of Symbolic Logic*, 83(3):868–898, 2018.
- [FV25] Harvey Friedman and Albert Visser. When bi-interpretability implies synonymy. *The Review of Symbolic Logic*, 18(4):971–990, 2025.
- [Hal14] Volker Halbach. *Axiomatic theories of truth*. Cambridge University Press, second edition, 2014.
- [Har68] Joseph Harrison. Recursive pseudo-well-orderings. *Transactions of the American Mathematical Society*, 131(1):526–543, 1968.
- [HLŁ25] Volker Halbach, Graham E. Leigh, and Mateusz Łelyk. Axiomatic Theories of Truth. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2025 edition, 2025.
- [Kri75] Saul Kripke. Outline of a theory of truth. *The Journal of Philosophy*, 72(19):690–716, 1975.
- [Poh09] Wolfram Pohlers. *Proof theory: The first step into impredicativity*. Springer Science & Business Media, 2009.
- [RFH21] Alfredo Roque Freire and Joel David Hamkins. Bi-interpretation in weak set theories. *The Journal of Symbolic Logic*, 86(2):609–634, 2021.
- [RFW25] Alfredo Roque Freire and Kameryn J. Williams. Non-tightness in class theory and second-order arithmetic. *The Journal of Symbolic Logic*, 90(2):627–654, 2025.
- [Sat15] Kentaro Sato. Full and hat inductive definitions are equivalent in nbg. *Archive for Mathematical Logic*, 54(1):75–112, Feb 2015.
- [Tap99] Christian Tapp. Eine direkte einbettung von KP_ω in ID_1 . Master’s thesis, Westfälische Wilhelms-Universität Münster, 1999.
- [Vis06] Albert Visser. *Categories of theories and interpretations*, page 284–341. Lecture Notes in Logic. Cambridge University Press, 2006.