Relative Interpretations and Substitutional Definitions of Logical Truth and Consequence

MIRKO ENGLER

Abstract: This paper proposes substitutional definitions of logical truth and consequence in terms of relative interpretations that are extensionally equivalent to the model-theoretic definitions for any relational first-order language. Our philosophical motivation to consider substitutional definitions is based on the hope to simplify the meta-theory of logical consequence. We discuss to what extent our definitions can contribute to that.

Keywords: substitutional definitions of logical consequence, Quine, meta-theory of logical consequence

1 Introduction

This paper investigates the applicability of relative interpretations in a substitutional account of logical truth and consequence. We introduce definitions of logical truth and consequence in terms of relative interpretations and demonstrate that they are extensionally equivalent to the model-theoretic definitions as developed in (Tarski & Vaught, 1956) for any first-order language (including equality). The benefit of such a definition is that it could be given in a meta-theoretic framework that only requires arithmetic as axiomatized by PA. Furthermore, we investigate how intensional constraints on logical truth and consequence force us to extend our framework to an arithmetical meta-theory that itself interprets set-theory. We will argue that such an arithmetical framework still might be in favor over a set-theoretical one.

The basic idea behind our definition is both to generate and evaluate substitution instances of a sentence φ by relative interpretations. A relative interpretation rests on a function f that translates all formulae φ of a language L into formulae $f(\varphi)$ of a language L' by mapping the primitive predicates P of φ to formulae ψ_P of L' while preserving the logical structure of φ

and relativizing its quantifiers by an L'-definable formula. In that sense, the application of translation functions boils down to the uniform substitution of predicates by complex formulae, which was the method used by Quine (1970) to give a substitutional definition of logical truth and consequence.

Substitutional definitions of logical truth and consequence play an important role in the history of logic. As it appears, a substitutional criterion has been traditionally the preferred choice for an analysis of what a logical truth is. However, since the seminal work of Tarski (1936) on the concept of logical truth and consequence, substitutional definitions were considered obsolete in the 20th Century by most of the logicians. An important exception to that is, as mentioned, Quine, who was interested in avoiding the ontological commitments of set-theory, which is the required meta-theory for a model-theoretic definition.[1] Still, it might seem perplexing to any modern logician that the most basic notion in logic, the notion of logical consequence, needs to be explained in the opulent realm of set-theory. At least, let it be our motivation to take up Quine's project to give a substitutional definition of logical consequence in order to avoid the use of set-theory as far as possible.

Another motivation for considering a substitutional definition of logical truth and consequence instead of a model-theoretic one stems from the alleged problem of truth preservation in a model-theoretic setting. In this context "truth preservation" means that from the logical truth of a sentence one can infer its truth simpliciter. It has been claimed that a model-theoretic setting fails to give a so-called "intended" interpretation of a language[2] by the notion of truth-in-a-model. Thus, an adequate explanation of "truth simpliciter" is missing in a model-theoretic setting. A discussion of that issue and a substitutional definition that trivially seems to fulfill truth preservation has been recently given by Halbach (2018). In the present paper, however, we will only focus on the aspect of avoiding a set-theoretic meta-theory for logical consequence.

The paper proceeds as follows: Section 2 gives a short outline of Tarski's reasons to reject substitutional definitions of logical truth and consequence as well as Quine's idea of how to improve substitutional definitions. By pointing out the limits of his definition in (Quine, 1970), we also try to motivate our framework of relative interpretations. Section 3 introduces

[1] For an outline of the development of Quine's conception of logical truth and consequence, see (Wagner, 2019).

[2] One example mentioned by Halbach (2018) is that in a model-theoretic framework the universal quantifier has to be evaluated as a set. For a set-theoretic language, however, this seems to be inadequate since the collection of all sets is a class and not a set.

a definition of logical truth and consequence and proves its extensional adequacy. The framework of this definition can be seen to only require PA if the antecedent-set of a logical consequence is assumed to be recursively enumerable. Nevertheless, we have to point out that this definition might be considered inadequate from an intensional perspective as compactness is used to define consequence and a corresponding notion of satisfaction is not available. Subsequently, Section 4 considers another definition that is extensionally adequate but also meets the mentioned intensional constraints. However, the required meta-theory for that definition itself interprets set-theory. Section 5 discusses in how far these results still can be considered as a method of avoiding set-theory in the meta-theory of logical consequence.

2 Substitutional Definitions of Logical Consequence: Tarski and Quine

A definition of logical truth and consequence on the basis of the substitution of syntactic particles in sentences is apparently close to an informal characterization of logical truth and consequence. Traditionally, the understanding of a logical consequence is reflected in criteria of truth preservation in virtue of form, wherein it is suggested to explain “form” in terms of substitution. As is well known, even Tarski considered a substitutional criterion for an adequate analysis of logical consequence in (Tarski, 1936). There he states a condition (F), which aims to characterize a logical consequence in the following way:

> (F) If, in the sentences of the class K and in the sentence X, the constants—apart from the purely logical constants—are replaced by any other constants (like signs are being everywhere replaced by like signs) and if we denote the class of sentences thus obtained from K by K', and the sentence obtained from X by X', then the sentence X' must be true provided only that all sentences of the class K' are true.[3]

Tarski considered condition (F) to be necessary for any definition of logical consequence. However, he proceeds with the claim:

> The condition (F) could be regarded as sufficient for the sentence X to follow from the class K only if the designations of all possible objects occurred in the language in question. This assumption, however, is fictitious and can never be realized.

[3] Translation of (Tarski, 1936) in (Tarski, 1956, pp. 409–420) by J.H.Woodger.

Here, Tarski indicates that a substitutional characterization of logical consequence cannot deal with a language that has not enough non-logical expressions in order to make a true but non-logical true conclusion false by a substitution. Whereas in a model-theoretic setting it would be possible to vary the evaluation of the non-logical constants over the domain of a certain model to eliminate this problem. The assumption that no substitutional method can compete with a set-theoretic method might seem plausible. Notably, the antinomies of Russell and Grelling show that there are formulae that don't determine a set as well as sets that cannot be determined by any formula. Nevertheless, it will turn out as a misconception that a substitutional method cannot be adequate for any language.

Quine's improvement of the substitutional method rests on the idea of a uniform substitution of predicates by complex formulae of a language, which provides an increased range of possible substitution instances. If a first-order relational language L (without equality) is strong enough[4] to express elementary arithmetic, then - Quine claimed - such a substitutional notion of logical consequence can be extensionally equivalent to the model-theoretic notion. An accessible presentation of that idea can be found in (Ebbs & Goldfarb, 2018).

Nonetheless, Quine's remarkable achievement is evidently still of minor generality compared to the model-theoretic treatment as it leaves out languages that cannot express elementary arithmetic or languages that consider equality as a logical constant. A crucial hurdle for any substitutional definition of logical truth and consequence is the treatment of the equality relation. For any language L that considers "=" as a logical constant we have (contingent) sentences without non-logical symbols (like "$\exists v_1 \exists v_2 (v_1 \neq v_2)$"). Consequentially, no symbol can be substituted since all symbols are logical ones. So the difficult question would be how to generate substitution instances that can come out false in the language L—especially if L is required to express elementary arithmetic.

Another problem for Quine but in general for any substitutional account that considers the preservation of truth under substitution arises from the notion of truth itself. Any definition of an adequate notion of truth for a language L that expresses elementary arithmetic cannot be given in the arithmetical language L. Any language L^+ that is capable of expressing the notion of truth for L must have expressive power exceeding that of L.

[4] This means that the language contains, inter alia, predicates that can be evaluated in a structure as the basic arithmetic relations of being the number zero, being the successor of a number, the sum of two numbers and the product of two numbers.

For that reason, a substitutional definition of logical consequence, where the notion of truth for L is defined in L^+, would not be applicable to L^+. As is well known, Tarski's original account of logical consequence in (Tarski, 1936) was also confronted with that problem, which led him to the model-theoretic framework of (Tarski & Vaught, 1956). While the definition of logical consequence is given in the language of set-theory, the definition is applicable to the language of set-theory itself. If "truth" is not taken as primitive, then the challenge for any substitutional definition will be not to fall back to the use of model-theory while preserving its universal applicability. Conclusively, we will critically have to review to what extent our definition is able to master this challenge.

3 Interpretations and Logical Truth and Consequence

The first improvement of our substitutional definition is based on the substitution of the predicate symbols of a language L by complex syntactic particles from *another* language. In addition to Quine's improvement of the substitution of predicates by *complex* formulae, we also allow these formulae to be taken (uniformly) from another language. We choose the arithmetical language of $L[\text{PA}]$. Thereby, we again extend the range of the available substitution instances for a language to a sufficient level and bypass the problem of excluding languages that are not rich enough to express arithmetic from the definition (in the sense explained above).

A convenient way of generating these substitution instances for any language is to use relative translation functions as they first appeared in (Tarski, Mostowski, & Robinson, 1953) to define relative interpretations.[5] To keep things simple, we only want to consider relational languages to be translated. Nevertheless, all results also apply to languages with constant and function symbols due to their relational eliminability and of course, the translating language may always contain constant and function symbols.

[5] A relative interpretation of a theory S in a theory T is a relative translation function f (as explained in Section 1) from $L[S]$ to $L[T]$ s.t. T proves all the translated theorems of S. For that, we shortly write $S \prec_f T$ and $S \prec T$ if there is a translation function f s.t. $S \prec_f T$.

Definition 1 (Substitution Function) *Let L be a relational language of first-order logic with equality. Then a* substitution function *f for L is given by a function $I : L \to L[\mathrm{PA}]$ assigning a formula of $L[\mathrm{PA}]$ in n variables to every n-ary predicate symbol of L and a formula $\delta(v)$ of $L[\mathrm{PA}]$ with exactly one free variable v such that:*

1. $f(v_{i_n} = v_{i_m}) \doteq (v_{i_n} = v_{i_m})$[6]
2. $f(Pv_{i_1}...v_{i_n}) \doteq I(P)(v_{i_1}...v_{i_n})$[7]
3. $f(\neg\varphi) \doteq \neg f(\varphi)$ *and* $f(\varphi \to \psi) \doteq f(\varphi) \to f(\psi)$
4. $f(\forall v_i \varphi) \doteq \forall v_i(\delta(v_i) \to f(\varphi))$
5. $\mathrm{PA} \vdash \exists v \delta(v)$

A substitution function is simply a relative translation function where the translating language is $L[\mathrm{PA}]$ and the formula δ, which defines in $L[\mathrm{PA}]$ the domain for evaluating the quantifier, can be proven to be non-empty in PA (condition 5).[8] This condition ensures that a substitution function preserves logical truth in PA, which is the subject of Lemma 1 below.

The advantage of a translation function that can relativize quantifiers comes into effect when we require an adequate treatment of "=". As mentioned, a substitutional definition of logical truth and consequence for a language including "=" as a logical constant is confronted with the problem of defining substitution instances for (contingent) sentences without non-logical symbols (like "$\neg\forall v_1 \forall v_2(v_1 = v_2)$"). Since we cannot substitute any symbol in those sentences in order to generate a substitution instance that is unsatisfiable, the relativization of quantifiers offers a possibility to deal with that problem. For instance, a substitution function f where $\delta(v)$ equals "$v = \overline{0}$"[9] translates the sentence "$\neg\forall v_1 \forall v_2(v_1 = v_2)$" to "$\neg\forall v_1(v_1 = \overline{0} \to \forall v_2(v_2 = \overline{0} \to v_1 = v_2))$". Trivially, $\mathrm{PA} \vdash \exists v\delta(v)$,

[6] The symbol "$\doteq$" denotes equality in the meta-language.

[7] The substitution of the variables in $I(P)$ by $v_{i_1}, ..., v_{i_n}$ may cause a collision of variables, which means that variables which were free in the original formula are bound by quantifiers in $I(P)$. In such a case we rename the bounded variables of $I(P)$ by $v_{i_{m+1}}, ..., v_{i_{m+k}}$, where m is the maximum of the indices of the variables $v_{i_1}, ..., v_{i_n}$.

[8] Sometimes a similar condition can be found in definitions of relative interpretability, namely that the interpreting theory should prove the relativization δ of a translation f to be non-empty. However, if f is a relative interpretation, then the interpreting theory trivially proves "$\exists v\delta(v)$" for it has to interpret "$\exists vv = v$" as a theorem of logic.

[9] The symbol "$\overline{0}$" is an individual constant of $L[\mathrm{PA}]$ which is meant to denote 0.

so we gave a substitution function f such that the substitution instance of "$\neg\forall v_1 \forall v_2 (v_1 = v_2)$" is not provable in arithmetic. Moreover, its negation will be provable and we have got sufficient means to classify the sentence "$\neg\forall v_1 \forall v_2 (v_1 = v_2)$" as not logically true by our substitutional method.

The second improvement of our definition is based on using relative interpretations in order to omit the problem of explaining the notion of truth for any language. Instead of considering the invariance of truth under substitution, we consider the invariance of interpretability in PA under substitution for the classification of a logical truth. That this turns out to be sufficient is due to Lemma 2 below, which roughly states that for a consistent set of sentences Γ one can define in PA a model for Γ by assuming the formal consistency of Γ. Usually, one considers theories as the objects that are interpreted. However, we can extend the range of relative interpretations to sets of sentences if we take care of the fact that they are not always deductively closed. To define a logical consequence in this framework, we utilize that consequence can be classified in terms of logical truth due to the compactness of logical consequence. In that respect, our definition also resembles Quine's definition in (Quine, 1970).

Definition 2 (Logical Truth and Consequence) *Let φ be any sentence of a relational language L of first-order logic with equality, Γ a set of L-sentences;*

1. $LTr(\varphi) :\Leftrightarrow \mathbb{V} f(subst(f) \Rightarrow \mathrm{PA} \vdash f(\varphi))$[10]
2. $LConseq(\Gamma, \varphi) :\Leftrightarrow \exists\Gamma'(\Gamma'$ *is finite* $/\!\!\wedge\, \Gamma' \subseteq \Gamma \,/\!\!\wedge\, LTr(\bigwedge \Gamma' \to \varphi))$

Lemma 1 (Correctness) *Let φ be any sentence of a relational language L of first-order logic with equality, then $\models \varphi \Rightarrow \mathbb{V} f(subst(f) \Rightarrow \mathrm{PA} \vdash f(\varphi))$.*

Proof. Assume $\models \varphi$ and let f be a substitution function. Define a monotone operator Π on sets of L-sentences considering a usual list of logical axioms and rules of inference for first-order logic such that the logical truths are the smallest set which is closed under Π, i.e., $\models \varphi$ is equivalent to $\mathbb{V} X(\Pi(X) \Rightarrow \varphi \in X)$ and it holds that $\mathbb{V} X(\Pi(X) \Rightarrow (\models \varphi \Rightarrow \varphi \in X))$. Let $Y := \{\varphi \in Sent_L \mid \mathrm{PA} \vdash f(\varphi)\}$. It can be easily shown that $\Pi(Y)$ holds and so $\mathrm{PA} \vdash f(\varphi)$. □

[10] We use double-lined logical symbols like "$\mathbb{V}$" for the logical constant of the meta-language and "$subst(f)$" as a shorthand for "f is a substitution function".

Lemma 2 (Formalized Completeness) *Let T be a consistent extension of* PA *in* $L[\mathrm{PA}]$*,* Γ *be a set of sentences of a relational language L of first-order logic with equality s.t. its deductive closure is numerated*[11] *in T by an* $L[\mathrm{PA}]$*-formula γ, then* $\Gamma \prec T + \mathrm{Con}_\gamma$.[12]

Corollary 1 *Let* Γ *be a consistent set of sentences of a relational language L of first-order logic with equality s.t. its deductive closure is numerated in* PA *by an* $L[\mathrm{PA}]$*-formula γ, then* $\exists f(\mathit{subst}(f) \wedge \Gamma \prec_f \mathrm{PA} + \mathrm{Con}_\gamma)$.

Proof. By Lemma 2, it immediately follows that there is a relative translation function $f : L \to L[\mathrm{PA}]$ such that $\mathrm{PA} \vdash f(\psi)$ for all sentences ψ s.t. $\Gamma \vdash \psi$. The proof of Lemma 2 which can be found in (Lindström, 1997, §6) shows that already $\mathrm{PA} \vdash \exists v \delta(v)$, where δ is the relativization of a relative translation that interprets Γ in $\mathrm{PA} + \mathrm{Con}_\gamma$. □

Proposition 1 *Let φ be any sentence of a relational language L of first-order logic with equality and* Γ *a set of L-sentences, then;*

1. $\models \varphi \Leftrightarrow \mathit{LTr}(\varphi)$
2. $\Gamma \models \varphi \Leftrightarrow \mathit{LConseq}(\Gamma, \varphi)$

Proof. (1.) : Follows from (2.) for $\Gamma = \emptyset$.

(2. $\Rightarrow$) : Assume that $\Gamma \models \varphi$. Therefore, there is a finite subset Γ' of Γ such that $\models \bigwedge \Gamma' \to \varphi$. By Lemma 1, it follows that $\forall f(\mathit{subst}(f) \Rightarrow \mathrm{PA} \vdash f(\bigwedge \Gamma' \to \varphi))$. So $\mathit{LConseq}(\Gamma, \varphi)$.

(2. $\Leftarrow$) : Assume there is a finite subset Γ' of Γ s.t. $\forall f(\mathit{subst}(f) \Rightarrow \mathrm{PA} \vdash f(\bigwedge \Gamma' \to \varphi))$ and $\Gamma \not\models \varphi$. From the latter assumption it follows that $\Gamma \cup \{\neg\varphi\}$ is consistent. Furthermore, $\Gamma' \cup \{\neg\varphi\}$ is consistent. Since $\Gamma' \cup \{\neg\varphi\}$ is finite, its deductive closure can be numerated in PA by a formula γ^* in Σ^0_1. By Corollary 1, there is a substitution function f s.t. $\mathrm{PA} + \mathrm{Con}_{\gamma^*} \vdash f(\bigwedge \Gamma' \wedge \neg\varphi))$. By definition, $f(\bigwedge \Gamma' \wedge \neg\varphi)$ is equivalent to $\neg f(\bigwedge \Gamma' \to \varphi)$. By assumption, however, $\mathrm{PA} + \mathrm{Con}_{\gamma^*} \vdash f(\bigwedge \Gamma' \to \varphi)$, which means that $\mathrm{PA} + \mathrm{Con}_{\gamma^*}$ is inconsistent. As we assume PA to be consistent, it holds that $\mathrm{PA} \vdash \neg \mathrm{Con}_{\gamma^*}$. This is a contradiction as $\Gamma' \cup \{\neg\varphi\}$ is consistent and PA is assumed to be Σ^0_1-sound.[13] So $\Gamma \models \varphi$. □

[11] An $L[\mathrm{PA}]$-formula γ numerates a set Γ in T iff $\varphi \in \Gamma \Leftrightarrow T \vdash \gamma(\overline{\ulcorner \varphi \urcorner})$ for all φ, where $\overline{\ulcorner \varphi \urcorner}$ denotes the gödelnumeral of φ. γ bi-numerates Γ in T if additionally, it holds that $\varphi \notin \Gamma \Leftrightarrow T \vdash \neg\gamma(\overline{\ulcorner \varphi \urcorner})$ for all φ.

[12] Originally, see (Feferman, 1960, Theorem 6.2).

[13] This means $\mathrm{PA} \vdash \varphi$ implies $\mathfrak{N} \models \varphi$ if $\varphi \in \Sigma^0_1$. The Σ^0_1-soundness of PA is equivalent to its 1-consistency, a purely syntactical notion, which is ω-consistency restricted to p.r. formulas. For more details, see, e.g., (Smorynski, 1977, §4).

Proposition 1 shows that it is possible to give a substitutional definition of logical truth and consequence which is extensionally equivalent to the model-theoretic definition and only seems to require arithmetic as a meta-theory. Apart from that we will have to relativize this claim in the last section, it can be already pointed out that the definition still has some obvious flaws on the intensional level. For instance, if we expect the notion of logical truth defined in terms of relative interpretations to behave in the same way as its counterpart defined in terms of models, this might include the request that giving a structure which satisfies $\neg\varphi$ in order to show that φ is not a logical truth corresponds to giving a substitution function f such that PA $\vdash f(\neg\varphi)$. This is to request that $\neg Ltr(\varphi) \Leftrightarrow \exists f(subst(f) \mathbin{/\!\!\!\wedge}$ PA $\vdash f(\neg\varphi))$. As a reason for that, one could think of this equivalence as a feature of logical truth that ensures its semantical character. However, by definition of $\neg Ltr(\varphi)$, we only know that there is a substitution function f such that PA $\nvdash f(\varphi)$. But since PA is incomplete, this cannot be equivalent to PA $\vdash f(\neg\varphi)$ for any sentence φ of L[PA].[14]

This fact corresponds to the observation that we cannot have a definition of the notion of satisfaction in terms of relative interpretations in PA (or in any consistent, recursively enumerable extension of PA). Any adequate notion of satisfaction would have to fulfill that φ being not a logical consequence of Γ is equivalent to the satisfiability of $\Gamma + \neg\varphi$. For our substitutional definition of logical consequence, this also fails for reasons of incompleteness. The conclusion we can draw from this is that an adequate substitutional definition of logical truth and consequence should come with a corresponding notion of satisfaction.

Another point, which has been originally made by Boolos (1975), is that a substitutional definition of logical consequence is of minor generality compared to the model-theoretic one if the compactness of the consequence relation is built-in to its definition—like in our case. Moreover, it may turn out that if the consequence relation is defined without built-in compactness, compactness doesn't hold for a so defined consequence relation.[15] Both issues, that of a corresponding notion of satisfaction and that of compactness will be taken care of in the following section.

[14] For a counterexample, consider a sentence φ that expresses in L[PA] the inconsistency of PA. Trivially, φ is not a logical truth and $\neg\varphi$, which expresses the consistency of PA, is not interpretable in PA (see Feferman, 1960).

[15] See (Eder, 2016) for an outline of that discussion.

4 Interpretations, Satisfaction and Compactness

To meet the requirements which were brought up in the previous section, we restrict the following substitutional definition of satisfaction and logical consequence to arithmetically definable sets[16]of assumptions and extend our arithmetical background by the set $\mathrm{Tr}(\Pi^0_{n+1})$ for a given n, which characterizes in $L[\mathrm{PA}]$ the set of $L[\mathrm{PA}]$-sentences that are of complexity Π^0_{n+1} and true in the standard model. Thereby, we have got an arithmetical theory which proves the consistency of any consistent set of sentences that is arithmetically definable in Π^0_n. It is important to note that the set $\mathrm{Tr}(\Pi^0_{n+1})$ can be defined purely syntactically and we don't have to refer to the model-theoretic definition of satisfaction.[17]

Definition 3 *Let φ be a sentence of a relational language L of first-order logic with equality and Γ a set of L-sentences s.t. the deductive closure of Γ is arithmetically definable by an $L[\mathrm{PA}]$-formula in Π^0_n (if $n > 0$, Σ^0_1 otherwise), then*

1. $Sat^n(\Gamma) :\Leftrightarrow \exists f(subst(f) \bigwedge \Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}))$
2. $LConseq^n(\Gamma, \varphi) :\Leftrightarrow \forall f(subst(f) \Rightarrow (\Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \Rightarrow \varphi \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})))$
3. $LTr^*(\varphi) :\Leftrightarrow LConseq^0(\emptyset, \varphi)$

Corollary 2 *Let Γ be a set of sentences of a relational language L of first-order logic with equality s.t. the deductive closure of Γ is arithmetically definable by an $L[\mathrm{PA}]$-formula in Π^0_n (if $n > 0$, Σ^0_1 otherwise). If Γ is consistent, then $\exists f(subst(f) \bigwedge \Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}))$.*

The Corollary follows from Lemma 2, the observation that $\mathrm{PA} \vdash \exists v\delta(v)$, where δ is the relativization of a relative translation that interprets Γ in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$, and the fact that any arithmetically definable set Γ in Π^0_n can be numerated in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$, which is the subject of the following Lemma.

Lemma 3 *Let Γ be a set of sentences of a relational language L of first-order logic with equality s.t. Γ is arithmetically definable by an $L[\mathrm{PA}]$-formula γ in Π^0_n, then γ bi-numerates Γ in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$.*

[16] An $L[\mathrm{PA}]$-formula γ arithmetically defines a set of sentences Γ iff $\varphi \in \Gamma \Leftrightarrow \mathfrak{N} \models \gamma(\overline{\ulcorner\varphi\urcorner})$ for all φ, where $\mathfrak{N}$ denotes the standard model for $L[\mathrm{PA}]$.

[17] For a definition, see, e.g., (Kaye, 1991, § 9.3).

Proof. Assume there is a γ in Π^0_n s.t. $\psi \in \Gamma \Leftrightarrow \mathfrak{N} \models \gamma(\overline{\ulcorner \psi \urcorner})$. Since $\gamma(\overline{\ulcorner \psi \urcorner})$ is in Π^0_n, if $\psi \in \Gamma$, then $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash \gamma(\overline{\ulcorner \psi \urcorner})$. Assume now $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash \gamma(\overline{\ulcorner \psi \urcorner})$ but $\psi \notin \Gamma$, which means that $\mathfrak{N} \models \neg\gamma(\overline{\ulcorner \psi \urcorner})$. Since $\neg\gamma(\overline{\ulcorner \psi \urcorner})$ is in Σ^0_n, we would have $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash \neg\gamma(\overline{\ulcorner \psi \urcorner})$, which leads to a contradiction since we assume that $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ is consistent. Analogously, it follows that γ is a bi-numeration. □

The compactness of Sat^n follows from the following Lemma, which generalizes Orey's Compactness Theorem in (Orey, 1961, Theorem 3.1) to arithmetically definable sets. As he shows in (Orey, 1961, Theorem 3.2), his Compactness Theorem fails for sets that are not recursively enumerable. However, this only applies if the interpreting theory is itself recursively enumerable. The proof is implicit in (Orey, 1961) and (Feferman, 1960).

Lemma 4 (Formalized Compactness) *Let T be a consistent extension of $\mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$ in $L[\mathrm{PA}]$ and Γ an arithmetically definable set of sentences of a relational language L of first-order logic with equality s.t. its deductive closure can be numerated in T. Let $\Gamma|k$ denote the set $\{\varphi \in \Gamma \mid \ulcorner \varphi \urcorner \leq k\}$, then $\forall k \in \mathbb{N} : \Gamma|k \prec T \Rightarrow \Gamma \prec T$.*

Proof. We assume that $\Gamma|k \prec T$ for every $k \in \mathbb{N}$ and that γ is a numeration of the deductive closure of Γ in T. As T is assumed to be consistent, $\Gamma|k$ is consistent for every $k \in \mathbb{N}$. Therefore, $T \vdash \mathrm{Con}_{\gamma_k}$ for every k (with γ_k in Σ^0_1 numerating the deductive closure of $\Gamma|k$ in T). Now define $\gamma^*(x) := \gamma(x) \wedge \mathrm{Con}_{\gamma_x}$ (x is free in Con_{γ_x}). Along the lines of the proof of (Lindström, 1997, Theorem 2.7) it follows that $T \vdash \mathrm{Con}_{\gamma^*}$. As γ numerates the deductive closure of Γ in T and $T \vdash \mathrm{Con}_{\gamma_k}$ for every $k \in \mathbb{N}$, also γ^* numerates the deductive closure of Γ in T. Finally, by $T \vdash \mathrm{Con}_{\gamma^*}$ and Lemma 2, we conclude that $\Gamma \prec T$. (Note that the Feferman-consistency statement Con_{γ^*} is equally eligible for an application of Lemma 2.) □

Proposition 2 *Let φ be a sentence of a relational language L of first-order logic with equality and Γ a set of L-sentences s.t. the deductive closure of Γ is arithmetically definable by an $L[\mathrm{PA}]$-formula in Π^0_n (if $n > 0$, Σ^0_1 otherwise), then;*

1. $\exists M(\mathit{structure}_L(M) \mathbin{/\!\!\!\wedge} M \models \Gamma) \Leftrightarrow Sat^n(\Gamma)$
2. $\Gamma \models \varphi \Leftrightarrow LConseq^n(\Gamma, \varphi)$
3. $\models \varphi \Leftrightarrow LTr^*(\varphi)$

Proof. (1. $\Rightarrow$): Follows directly from Corollary 2. (1 $\Leftarrow$): Assume that $\Gamma \prec \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$. Since $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ is assumed to be consistent, Γ is also consistent and therefore satisfiable in some L-structure.

(2. $\Rightarrow$): Assume that $\Gamma \models \varphi$. Therefore, there exists a finite subset Γ' of Γ s.t. $\models \bigwedge \Gamma' \to \varphi$ and so (Lemma 1) $\forall f(\mathit{subst}(f) \Rightarrow \mathrm{PA} \vdash f(\bigwedge \Gamma' \to \varphi))$. Let f be any substitution function and assume that $\Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$. Therefore, $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash f(\bigwedge \Gamma')$ and $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash f(\varphi)$. So $\mathit{LConseq}^n(\Gamma, \varphi)$.

(2. $\Leftarrow$): Assume that $\Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \Rightarrow \varphi \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ for any substitution function f but $\Gamma \not\models \varphi$. $\Gamma \cup \{\neg\varphi\}$ is therefore consistent. The deductive closure of Γ is assumed to be arithmetically definable by a formula γ in Π^0_n. By Lemma 3, the deductive closure of Γ is bi-numerated by γ in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$. Therefore, $\gamma(\overline{\ulcorner \neg\varphi \to \dot{x} \urcorner})$ (x is free in $\gamma(\overline{\ulcorner \neg\varphi \to \dot{x} \urcorner})$) numerates the deductive closure of $\Gamma \cup \{\neg\varphi\}$ in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$. By Corollary 2, there exists a substitution function f s.t. $\Gamma \cup \{\neg\varphi\} \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$. In particular, $\Gamma \prec_f \mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ so that it follows by assumption that $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash f(\varphi)$—but also $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash \neg f(\varphi)$, which is a contradiction as $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ is assumed to be consistent. So $\Gamma \models \varphi$.

(3.): Follows from (2.) for $\Gamma = \emptyset$. □

Corollary 3 *Let φ be any sentence of a relational language L of first-order logic with equality and Γ a set of L-sentences s.t. the deductive closure of Γ is arithmetically definable by an $L[\mathrm{PA}]$-formula in Π^0_n (if $n > 0$, Σ^0_1 otherwise), then $\neg \mathit{LConseq}^n(\Gamma, \varphi) \Leftrightarrow \mathit{Sat}^n(\Gamma + \neg\varphi)$.*

The Corollary follows directly from Proposition 2 and it demonstrates that giving a model which satisfies $\Gamma + \neg\varphi$ in order to show that φ is not a logical consequence of Γ is equivalent to giving a substitution function f such that $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash f(\Gamma)$ but $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1}) \vdash \neg f(\varphi)$. In that respect, our substitutional definition of logical consequence is able to emulate an important property of its model-theoretic counterpart.

5 Conclusion

We proposed a substitutional definition of logical truth and consequence in terms of relative interpretations (Definition 2) and demonstrated that it is extensionally equivalent to the model-theoretic definitions for any relational first-order language with equality. We also demonstrated that the same result can be obtained for a substitutional definition (Definition 3) that comes with a corresponding notion of satisfaction and compactness is not built-in to the notion of satisfaction—if we restrict the definition to arithmetically definable sets of assumptions. As the motivation of this undertaking was to eliminate the need for set-theory in the meta-theory of logical consequence, we will now have to review to what extent this has been successful.

First of all, the use of set-theory usually enters (also in our presentation) the meta-theory of logical consequence not first through the explanation of semantical notions but already in specifying syntactical terms like “language”, “sentence” and “proof”. However, it is well-known since Gödel that all of these syntactical notions can be adequately represented in a meta-theory not extending weak arithmetical theories.[18] The same holds for the notions of “translation function” and “relative interpretation”. Apart from that, we obviously used set-theoretic terms like “set of sentences” in defining logical consequence substitutionally. If we want to specify this term purely arithmetically, we necessarily have to restrict the scope of our definition to sets of a certain complexity, namely to those which are definable by a first-order arithmetical formula. However, the notion of arithmetical definability itself involves the notion of satisfaction in the standard model of arithmetic. To avoid the use of set-theory in that aspect, we can instead restrict the definition to sets which can be numerated in an arithmetical theory. As Lemma 3 shows, this could be done in principle for any arithmetically definable set, but it requires a correspondingly strong arithmetical theory.

In the case of Definition 2, we don’t need to accept, at first sight, a meta-theory that exceeds PA.[19] Leaving aside its discussed intensional flaws, if we try to not exceed PA in the framework of Definition 2, we have to admit that the class of sets which can be numerated in PA covers only the recursively enumerable ones. Without a doubt, this is a limitation compared to a model-theoretic setting. However, it could be argued that we don’t lose much of generality since a non-recursively enumerable set of assumptions and its set of logical consequences is quite a rare thing to consider.

If we are interested in a substitutional definition that also meets the discussed intensional constraints, we necessarily have to exceed PA in any case. Even if we sacrifice the constraint that compactness should not be built-

[18] For details, see, e.g., (Smorynski, 1977).

[19] We assume that accepting a theory as a meta-theory implies that also its consistency is assumed and in case of an arithmetical theory also its ω-consistency, which is a common practice in Metamathematics. As pointed out, the 1-consistency of PA has to be assumed for Proposition 1 and the consistency of $\mathrm{PA} + \mathrm{Tr}(\Pi^0_{n+1})$ for Proposition 2.

in to the definition of satisfaction, we still need a meta-theory that involves $\mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$ to allow for an adequate substitutional notion of satisfaction. It could be argued that this fact undermines our intend to eliminate the need for set-theory in the meta-theory of logical consequence as set-theory axiomatized by ZF is interpretable in $\mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$. Still, we want to reply that Definition 3 offers a philosophical improvement for the meta-theory of logical consequence.

Evidently, we were able to replace the use of set-theoretic *vocabulary* by an arithmetical one. For the latter, it could be argued that its intuitive understanding is much clearer compared to the former as set-theory is a rather modern invention of 20th Century Mathematical Logic. Consequently, any definition in arithmetical terms is to favor over an equivalent definition in terms of sets.

Another point can be made if we consider again Quine's interest in a substitutional definition. He aimed to relieve the meta-theory of logical consequence from its ontological commitment to sets. The question is to what extent the fact that $\mathrm{ZF} \prec \mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$ affects this elimination of ontological commitment to sets. It has to be admitted that the term "ontological commitment" is not an entirely clear notion. So the previous question will not have a clear-cut answer and our argumentation can only be based on a somehow intuitive persuasion. Having said this, the aspect that we want to bring to attention here is that it is doubtful whether a presupposed ontology of a theory—in whatever sense—is preserved by relative interpretability. For instance, $\mathrm{ZF} + \mathrm{CH} \prec \mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$ as well as $\mathrm{ZF} + \neg\,\mathrm{CH} \prec \mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$. Intuitively, it seems plausible that an assumed ontology for $\mathrm{ZF} + \mathrm{CH}$ is in conflict with an assumed ontology for $\mathrm{ZF} + \neg\,\mathrm{CH}$. Thus, both ontologies cannot be adopted equally in an assumed ontology for $\mathrm{PA} + \mathrm{Tr}(\Pi^0_1)$. But as there appears to be no criterion to decide which one may be adopted, it seems questionable whether any ontology is preserved under relative interpretation in general.

Taking all this into consideration, we may conclusively say that this paper established a substitutional approach to logical truth and consequence that is able to avoid set-theoretic *vocabulary* in its meta-theory. For a definition that is extensionally adequate for recursively enumerable sets of assumptions, it was shown that the meta-theory has not to exceed PA. Though we don't expect our framework to replace the use of model-theory in the practice of determining a logical consequence, we consider it instructive that it could be done without set-theory in principle.

References

Boolos, G. S. (1975). On second-order logic. *Journal of Philosophy*, *72*(16), 509–527.

Ebbs, G., & Goldfarb, W. (2018). First-order validity and the Hilbert-Bernays Theorem. *Philosophical Issues*, *28*(1), 159–175.

Eder, G. (2016). Boolos and the metamathematics of Quine's definitions of logical truth and consequence. *History and Philosophy of Logic*, *32*, 170–193.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, *49*(1), 35–92.

Halbach, V. (2018). The substitutional analysis of logical validity. *Nous*. doi: 10.1111/nous.12256.

Kaye, R. (1991). *Models of Peano Arithmetic*. Clarendon Press, Oxford.

Lindström, P. (1997). *Aspects of Incompleteness* (2nd (2003) ed.). Springer, Berlin.

Orey, S. (1961). Relative interpretations. *Mathematical Logic Quarterly*, *7*(10), 146–153.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press, Cambridge, Massachusetts.

Smorynski, C. (1977). The incompleteness theorems. In J. Barwise (Ed.), *Handbook of Mathematical Logic* (pp. 821–866). North-Holland, Amsterdam.

Tarski, A. (1936). Über den Begriff der logischen Folgerung. *Actes du congres international de philosophie scientifique*, *7*, 1–11.

Tarski, A. (1956). *Logic, Semantics, Metamathematics*. Clarendon Press, Oxford.

Tarski, A., Mostowski, A., & Robinson, R. M. (1953). *Undecidable theories*. North-Holland, Amsterdam.

Tarski, A., & Vaught, R. (1956). Arithmetical extensions of relational systems. *Compositio Mathematica*, *13*, 81–102.

Wagner, H. (2019). Quine's substitutional definition of logical truth and the philosophical significance of the Löwenheim-Hilbert-Bernays theorem. *History and Philosophy of Logic*, *40*(2), 182–199.

Mirko Engler
Humboldt University Berlin, Department of Philosophy
Germany
E-mail: mir.engler@gmail.com