VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

4
A categorical theory of truth[1]

Juhani Yli-Vakkuri & Zachary Goodsell

Introduction

Tarski's method for defining truth for languages of finite order is generally understood to be his most important contribution to semantics. Tarski sets a precise standard for a definition of truth to be 'adequate', and he proves that definitions constructed by his method meet it. The standard is *Convention T*:

Convention T. *A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an* adequate *definition of truth if it has the following consequences:*

(α) all sentences which are obtained from the expression 'x ∈ Tr if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and the symbol 'p' the expression which forms the translation [2] *of this sentence into the metalanguage; (β) the sentence 'for any x, if x ∈ Tr then x ∈ S' (in other words 'Tr ⊆ S').* (Tarski 1955[1933]: 188.)

A 'formally correct' definition of the symbol 'Tr' is a sentence of the form

[1] Thanks to Beau Madison Mount, Peter Fritz, Volker Halbach, John Hawthorne, Lavinia Picollo, Tim Williamson, Dan Waxman, and the audience at TimFest: A Conference in Honor of Timothy Williamson at Magdalen College, University of Oxford on August 30th–July 1st, 2023 for helpful comments.

[2] We depart from Tarski in assuming that the object language is included in the metalanguage, so the translation of any object language sentence in the metalanguage will be just that sentence itself.

Yli-Vakkuri, J. & Goodsell, Z. (2025). A categorical theory of truth. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 49–58). https://doi.org/10.61201/h5fbdk64

$$\mathrm{Tr} = \theta,$$

where θ is a predicate of the metalanguage in which 'Tr' does not occur. Here, the word 'definition' is reserved for such identity sentences. Conditions (α) and (β) concern the consequences of such an identification in the metatheory. (α) requires that every instance of the T-schema for the object language be derivable from the definition. So, if the object language includes, for example, the sentence '$1 + 1 = 2$' then the adequacy of a definition requires the sentence

$$\text{'}1 + 1 = 2\text{'} \quad \in \mathrm{Tr} \text{ if and only if } 1 + 1 = 2$$

or, in presently preferred notation, the sentence

$$\mathrm{Tr}\ \overline{1 + 1 = 2} \leftrightarrow 1 + 1 = 2$$

to be derivable from it in the metatheory. 'S' is Tarski's symbol for *sentence of the object language*, so condition (β) requires that in the metatheory we can prove by means of the definition that only sentences of the object language are true: 'in other words', as Tarski puts it,

$$\mathrm{Tr} \subseteq \mathrm{S} \qquad (1)$$

—a truism, since 'Tr' is simply an abbreviation for 'true sentence of the object language'. Let us call the class of metalanguage sentences comprising (1) together with all instances of the T-schema 'T'. A formally correct definition of truth, then, is deemed adequate by Convention T iff T is derivable from it in the metatheory.

Against convention T

Tarski's method for constructing definitions of truth is clearly a significant achievement, because the definitions constructed using his method are enormously fruitful, as he showed. However, the significance of Convention T is not as clear. By general agreement, T captures one important aspect of truth, but it is obvious that T does not include every generalization—not even every 'obvious' generalization, such as 'every sentence or its negation is true', that we expect to be able to prove in a good theory of truth. As Tarski himself was the first to point out, not even the theory that results from adding T to his preferred metatheory, which is (n+3)rd-order syntax formulated in a language that includes the object language, where n is the order of the object language, includes such obvious generalizations.

A policy of accepting definitions of truth based on whether they satisfy Convention T either fails to vindicate Tarski's definition or overgenerates. A policy of merely accepting *some* definition or other that can be proved adequate in the sense of Convention T fails to vindicate Tarski's definition, since in any ω-incomplete theory with an adequate definition, there is another adequate definition which is not provably equivalent. On the other hand, a policy of accepting *all* definitions that can

be proved adequate in the sense of Convention T does vindicate Tarski's definition, but it also requires going far beyond Tarski's definition, by requiring us to accept a theory that, if consistent, is not recursively enumerable.[3] Such theories cannot be presented by a recursive set of axioms and rules, so are of limited use.

The reason why we should accept Tarski's definition of 'Tr', if we have any reason to accept it at all at present, is the fruitfulness of that definition, not that the definition satisfies Convention T. The derivability of is only a criterion of *minimal* adequacy for definitions of truth: we can rule out definitions that don't satisfy it (in the sense of not accepting any such definition, not in the sense of accepting its negation), but adopting a policy of accepting any definition that satisfies it, or even every definition that can be proved to satisfy it, would be either unmotivated or impossible to follow (insofar as it is impossible to accept a theory that is not recursively enumerable).

As criteria of minimal adequacy go, it is unclear why such a thing is needed, and it is also unclear why, supposing that such a thing is needed, Convention T should be it. Why should we not also require the derivability of compositional principles such as the principle

$$\forall x \in \mathrm{S}.\ (\mathrm{Tr}(\ulcorner \neg \frown x) \leftrightarrow \neg \mathrm{Tr}x)$$

which says that every sentence or its negation is true? A definition of truth that does not satisfy this condition would not be deemed minimally adequate, so, it seems, our criterion of minimal adequacy should be at least this strong. But it is easy to come up with further desirable theorems. A survey of the literature on axiomatic theories of truth[4] will turn up a large number of attractive combinations, and it is unclear why the derivability of any of them should be designated *the* condition of minimal adequacy for a definition of truth, if such a thing is needed at all.

Categoricity

It would be preferable to avoid Convention T altogether and to formulate an acceptable theory of truth in which Tarski's definition can simply be proved (rather than proved "adequate", whatever adequacy might be). Such a theory would be, in the terminology of Tarski's 1933 paper (1956: 257), a *categorical* theory of truth.[5]

---

[3] *Proof sketch.* Let M be any recursively axiomatizable metatheory such that 'Tr = ' is proved to satisfy Convention T, such that we can also prove in M

For all $x \in S$, either θx or $\theta(\ulcorner \neg \frown x)$.

Then let θ^M be the sentence:

$\lambda x \in S.\ \exists n$ (the length of x is n and and a contradiction cannot be derived in M in fewer than n steps).

The adequacy of 'Tr = θ' can be proved in M. So by Convention T we should accept '$\theta = \theta^M$', from which the consistency of M is derivable. The same will go for every recursively enumerable extension of M. □

[4] See Halbach 2011 for review.

[5] 'Categoricity' is also commonly used for a semantic rather than the present proof-theoretic property of theories. In contemporary model theory, a theory is called categorical if all of its models—including those with deviant interpretations of the quantifiers—are isomorphic. Tarski's theory of truth is not categorical in this sense (only negation-complete theories are). The term 'categorical' originates with Veblen 1904, where

Definition 1 (Categoricity)**.** A theory Γ is *categorical* with respect to the class of constants Δ= {c_1, c_2, . . .} if and only if, for any theory Γ′ obtained by replacing the constants in Δ by previously unused constants Δ′ = {c'_1, c'_2, . . .},

$$\Gamma, \Gamma' \vdash \overline{c_i = c'_i}$$

for each $c_i \in \Delta$.

Of interest here is the case where $\Delta = \overline{\{\mathrm{Tr}\}}$.

Tarski agreed that a categorical theory of truth would be preferable to Convention T, but he resorted to Convention T because he thought it would not be possible to formulate an acceptable such theory:

> [. . .] it seems natural to require that the axioms of the theory of truth, together with the original axioms of the metatheory, should constitute a categorical system. It can be shown that this postulate coincides in the present case with another postulate, according to which the axiom system of the theory of truth should unambiguously determine the extension of the symbol 'Tr' which occurs in it, and in the following sense: if we introduce into the metatheory, alongside this symbol, another primitive sign, e.g. the symbol 'Tr′' and set up analogous axioms for it, then the statement 'Tr = Tr′' must be provable. But this postulate cannot be satisfied. For it is not difficult to prove that in the contrary case the concept of truth could be defined exclusively by means of terms belonging to the morphology of language [i.e., syntax], which would be in palpable contradiction to Th. I.[6] (Tarski 1956[1933]: 257)

As the authors interpret this passage, Tarski is making a mistake. Tarski seems to be saying that any categorical set of axioms for 'Tr' from which T is derivable in the metatheory is inconsistent in the metatheory. But this simply cannot be the case. For consider the theory whose sole novel axiom is Tarski's own definition,

$$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}, \quad (2)$$

where '$\mathrm{Tr}_{\mathrm{Tarski}}$' abbreviates the complex truth predicate that Tarski showed us how to construct out of the object language, the constants of syntax, and quantifiers and variables of orders higher than those found in the object language. (2) is certainly categorical in Tarski's sense, and, as Tarski showed, T is derivable from (2) in the metatheory. And adding this one axiom cannot make the metatheory inconsistent

it is used for a geometrical theory which has only one model up to isomorphism where the non-geometrical constants are given the intended interpretation. Generalizing Veblen's geometric notion to semantics, we find that Tarski's original truth theory consisting of the sentences described in conditions (*α*) and (*β*) *is* categorical in the sense that among interpretations that agree with the intended one on all constants besides 'Tr', there is only one on which the theory comes out true. However, this fact could not possibly serve to justify Tarski's definition, since it takes Tarskian semantics for the metalanguage for granted.

6 Th. I is Tarski's undefinability theorem.

(nor could adding any definition of an otherwise unused expression, since such additions yield conservative extensions of the theories to which they are added). Of course, in justifying Tarski's definition it would not do to adopt that same definition as an axiom, but the point is that there is no in-principle problem with searching for a theory of truth that is both categorical and consistent.

A categorical theory of truth

Preliminaries

We work within finite-order fragments of Henkin's 1950[7] extensional higher-order logic plus standard axioms of syntax (Tarski used finite-order fragments of what he called the *calculus of classes*, which is essentially Henkin's system formulated with relational rather than functional types, and which he obtained by simplifying the extensional fragment of the *Principia Mathematica* system). A range of systems would work equally well for present purposes. In particular, systems to which Tarski's original hierarchy of definitions of truth satisfying Convention T can be adapted, and which are *intensional* in the sense that provable coextensionality of properties, relations, and propositions (if propositional quantification is included, as it is in Henkin's system) suffices for identity,[8] will generally also permit a modification of the theory presented here that is also categorical and a conservative extension of the system in question. For definiteness, we will ignore possible variations and adhere to the system of *extensional n^{th}-order syntax*, S^n, which has the following features.

- Simple functional types (as in Church 1940) with base types e and t and functional types—$(\sigma\tau)$ for any types σ and τ—of order $n + 2$ or less, where the order of a type is defined recursively as follows:

  – $O(e) = O(t) = 1$ and
  – $O(\sigma\tau) = \max\{O(\sigma) + 1, O(\tau)\}$ (i.e., the order of a functional type is one plus the order of the type of its argument, or is the order of the type of its output, whichever is greater).

- Infinitely many variables of each type of order n or less.
- λ-abstraction with the usual axioms asserting the substitutivity of β-equivalent terms (Church 1940).

[7] Henkin's system is what we get when we add the axiom of Boolean extensionality to Church's 1940 system—an addition considered and rejected by Church on the c.

[8] In contrast with *hyperintensional* systems like those of *Principia Mathematica* and Church 1940, where provable coextensionality is not sufficient for identity (Church's system lacks the axiom of Boolean extensionality). Notice that intensionality does not require that coextensionality implies identity, but only the substitutivity of provably coextensional terms (as in typical modal logics).

- Boolean connectives with classical propositional logic.
- Universal and existential quantifier symbols '$\forall_\sigma$', '$\exists_\sigma$' of type $(\sigma t)t$ for each type σ of order n or less, with classical quantifier logic at each type.
- Identity symbols '$=_\sigma$' of type $(\sigma t)\,(\sigma t)t$ for each type σ of order n or less, obeying the reflexivity of identity and Leibniz' law at each type.
- The axiom of Boolean extensionality,

$$\forall p\, q.\, ((p \leftrightarrow q) \rightarrow (p \rightarrow q))$$

- Axioms of function extensionality,

$$\forall f g^{\sigma\tau}.\, \big(\forall x^{\sigma}.\ (fx = gx) \rightarrow (f = g)\big).$$

- Axioms of choice,

$$\exists f^{(\sigma t)\sigma}.\, \forall g^{\sigma t}.\, (\exists g \rightarrow g(fg)).$$

- Standard axioms of syntax, asserting roughly that the strings are the free semigroup generated by some alphabet of characters. For definiteness, the name of a string will have type e (so the name of the class of strings, 'String', has type et).

A *finite signature* will be a finite list of typed constants separated by commas, and Σ_S will be the signature of syntax (which contains names for each character of the alphabet and a constant for concatenation of strings). Our object language, like Tarski's, will be $\mathscr{L}^n_\Sigma$ for an arbitrary order n and arbitrary finite signature Σ, which is the language described above but with the constants from Σ of order n included instead of the constants of syntax.

We will employ standard notational abbreviations for logic and syntax (e.g., '$\forall x \in \alpha.(\ldots)$' abbreviates '$\forall \lambda x.(\alpha x \rightarrow \ldots$'), and will take for granted standard formalizations of complex syntactic notions like the class of sentences in $\mathscr{L}^n_\Sigma$ (symbolized '$\mathscr{L}^{n\prime}_\Sigma$') and provability in S^{n+3} (symbolized '$S^{n+3}\vdash$'). The symbol 'Q' is used for the function which maps a string to its structural-descriptive name, and '$\frown$' for the concatenation function.

The F-schema

For the object language $\mathscr{L}^n_\Sigma$, the metalanguage Tarski uses to formulate his theories of truth is

$$\mathscr{L}^{n+3}_{\Sigma,\Sigma_s,\overline{\mathrm{Tr}}}$$

where as usual 'Tr' is the primitive truth predicate. Tarski's theory of truth is the list of sentences mentioned in conditions (α) and (β). Corresponding to condition (α) is an infinite list of sentences, one for each sentence φ of the object language:
T-schema$^{\varphi}$

$$\mathrm{Tr}\,\overline{\varphi} \leftrightarrow \varphi.$$

Condition (β) corresponds to a single additional sentence:

Sentential truth

$$\mathrm{Tr} \subseteq \mathscr{L}^n_\Sigma.$$

Call the class of all such sentences T^n_Σ (Convention T then says that a definition is adequate if every sentence of T^n_Σ can be proved from the definition).

Our metalanguage is, instead, $\mathscr{L}^{n+6}_{\Sigma,\Sigma_S,\overline{Tr}}$—the language of $(n{+}6)^{\text{th}}$-order logic plus the constants of Σ and of syntax and the primitive truth predicate—and our metatheory is S^{n+6}—$(n{+}6)^{\text{th}}$-order syntax. That is, our metalanguage and metatheory are what Tarski would use as metametalanguage and metametatheory for semantic theorising about the metalanguage. Our categorical theory of truth is given in its entirety by the following axiom schema where Φ may be replaced by any term of type *et* of $\mathscr{L}^{n+3}_{\Sigma,\Sigma_S}$:

Factivity of $S^{n+6} + \mathsf{T}^n_\Sigma(\Phi)$

$$\forall x \in \mathrm{String}.\left(\mathrm{S}^{n+3} + \mathsf{T}^n_\Sigma \vdash \left(\overline{\Phi} \frown \mathrm{Q}x\right)\right) \rightarrow \forall x \in \ \mathrm{String}.\,\Phi x.$$

Call the class of such sentences the *F-schema*, or F^n_Σ ('F' for 'Factivity'). The F-schema can be intuitively understood by way of example. One instance says that if the sentence 'if φ is a sentence then either it or its negation is true' can be derived in Tarski's minimal theory of truth for every string φ, then every string in fact has the property that if it is sentence then either it or its negation is true.

The F-schema is, in essence, a combination of highly plausible principles of closure and disquotation for *truth-in-*$\mathscr{L}^{n+3}_{\Sigma,\Sigma_S}$. The F-schema for a given predicate Φ can be decomposed into the following theses for a primitive notion of truth-in-$\mathscr{L}^{n+3}_{\Sigma,\Sigma_S,\overline{\mathrm{Tr}}}$ which we symbolize 'Tr^{n+3}':

Truth of $S^{n+3} + \mathsf{T}^n_\Sigma$ Every theorem of $S^{n+3} + \mathsf{T}^n_\Sigma$ is true-in-the-metalanguage (i.e., Tr^{n+3}).

$$\forall x.\ \left(\left(\mathrm{S}^{n+3} + \mathsf{T}^n_\Sigma \vdash x\right) \rightarrow \mathrm{Tr}^{n+3}x\right)$$

ω-Closure of truth$^\Phi$ If, for all sentences x, the application of the predicate Φ to the structural-descriptive name of x is true-in-the metalanguage, then the sentence '$\forall x \in \mathrm{String}.\Phi x$' is true-in-the-metalanguage.

$$\forall x \in \mathrm{String}.\ \mathrm{Tr}^{n+3}\left(\overline{\Phi} \frown Qx\right) \rightarrow \mathrm{Tr}^{n+3}\overline{\forall x \in \mathrm{String}.\,\Phi x}$$

T (out) schema for string quantification$^\Phi$ If the sentence '$\forall x \in \mathrm{String}.\Phi x$' is true-in-the-metalanguage, then $\forall x \in \mathrm{String}.\Phi x$.

$$\mathrm{Tr}^{n+3}\,\overline{\forall \mathrm{x} \in \mathrm{String}.\,\Phi \mathrm{x}} \rightarrow \forall \mathrm{x} \in \mathrm{String}.\,\Phi \mathrm{x}$$

Proposition 1. *Every instance of* $\mathrm{F}^{\mathrm{n}}_\Sigma$ *can be derived from Truth of* $\mathrm{S}^{\mathrm{n}+3} + \mathsf{T}^{\mathrm{n}}_\Sigma$*, instances of* ω*-Closure of Truth*$^\Phi$*, and instances of the T-schema for String Quantification.*

In addition to being very plausible, the F-schema is categorical, and indeed proves Tarski's definition.

Theorem 2 (Categoricity)**.** $\mathrm{F}^{\mathrm{n}}_\Sigma$ *is equivalent in* S^{n+6} *to* '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$'*, where* Tr_{Tarski} *abbreviates what Tarski defined truth (in* $\mathscr{L}^n_\Sigma$*) to be.*

Proof. To derive the definition from the F-schema, it will suffice to show that each instance of

$$\mathrm{Tr}\overline{\varphi} \leftrightarrow \mathrm{Tr}_{\mathrm{Tarski}}\overline{\varphi}$$

is a theorem of $S^{n+3} + \mathsf{T}^n_\Sigma$. This holds because each instance of the T-schema is assumed for 'Tr' and is provable for $\mathrm{Tr}_{\mathrm{Tarski}}$.

To derive the F-schema from the definition, let $\mathrm{Tr}^{n+3}_{Tarski}$ be the Tarskian defined truth-predicate for the object language $\mathscr{L}^{n+3}_{\Sigma,\Sigma_S\overline{Tr}}$. Tarski shows that we can derive all instances of ω-Closure of Truth and the T (Out) Schema for String Quantification in S^{n+6} when 'Tr^{n+3}' is replaced by $\mathrm{Tr}^{n+3}_{Tarski}$. It is also easy to show that $S^{n+3} + \mathsf{T}^n_\Sigma$. is satisfied when and only when 'Tr' is interpreted as $\mathrm{Tr}_{\mathrm{Tarski}}$. □

Remark 1. *$S^{n+6} + F^n_\Sigma$, since it follows from a definition, is a conservative extension of* S^{n+6}.

Remark 2. *The F-schema is axiomatized by the single instance where* Φ *is*

$$\lambda y.\big(\mathrm{Tr}\, y \leftrightarrow \mathrm{Tr}_{\mathrm{Tarski}}\ y\big),$$

since this instance suffices to prove '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$*' from which every other instance can be derived by Theorem 2.*

Relation to other categorical theories

The truth-definition '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$', being a definition, is a conservative extension of S^{n+3}. It is also categorical, as previously mentioned. By contrast, adding the F-schema to S^{n+3} results in a non-conservative extension of S^{n+3} if S^{n+3} is consistent, because the F-schema implies the consistency of S^{n+3}. However, although the F-schema is consistency-theoretically stronger than '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$', it is clearly unobjectionable from a Tarskian point of view. For anyone who adopts Tarski's unamended approach to truth also accepts S^{n+6}; they will regard S^{n+6} as the meta*meta*theory rather than as the metatheory, but they accept it just the same, and the result of adding the F-schema to S^{n+3} is a conservative extension of S^{n+6}. And the F-schema is not only unobjectionable on consistency-theoretic grounds from a Tarskian point of view; in a sense, accepting it already comes with the Tarskian approach: while those taking that approach do not make the F-schema part of their theory, in accepting $\mathrm{S}^{n+3} + \mathsf{T}^n_\Sigma$, they of course accept that whenever, 'For all strings s, 'Φs" is derivable from these in their metametatheory, then, for all strings s, Φs (they don't reject this, or suspend judgment on it).

Another categorical theory that has been discussed in the literature (first by Tarski himself immediately after he commits the mistake quoted above) is the closure of the Tarskian metatheory, S^{n+3}, under a syntactic ω-rule, which requires that when formulae

$$\Phi\overline{\gamma}$$

are provable for every string γ, then so is the formula

$$\forall x \in \mathrm{String}.\ (\Phi x).$$

As is widely known, theories closed under such a rule are either inconsistent or are not recursively enumerable, and here we are only considering formal (i.e., recursively enumerable) theories of truth.

References

Church, Alonzo (1940): 'A Formulation of the Simple Theory of Types', *Journal of Symbolic Logic* 5(2): 56–68. URL = https://doi.org/10.2307/2266170

Halbach, Volker (2011): *Axiomatic Theories of Truth*, Cambridge: Cambridge University Press.

Henkin, Leon (1950): 'Completeness in the Theory of Types', *Journal of Symbolic Logic* 15(2): 81–91. URL = https://doi.org/10.2307/2266967

Tarski, Alfred (1956 [1933]): *Logic, Semantics, Metamathematics*, Oxford: Clarendon Press.

Veblen, Oswald (1904): 'A System of Axioms for Geometry', *Transactions of the American Mathematical Society* 5(3): 343–384. URL = https://doi.org/10.2307/1986462