The mathematical core of Tarski's truth definition

Wilfrid Hodges
Queen Mary, University of London

August 2007

www.maths.qmul.ac.uk/∼wilfrid/unilog07.pdf

Alfred Tarski gave a mathematical description
of the set of true sentences of a fully interpreted logic.

The basic work in 1929; published in Polish in 1933.

Model-theoretic version published with Vaught in 1956.

Forthcoming in *Alfred Tarski: Philosophical Background, Development, and Influence*, ed. Douglas Patterson,
I argue that Tarski

- had no programme for defining semantic notions,
- wouldn't have known what to try if he had such a programme,
- reached his truth definition by purely technical manipulations of other things in the Warsaw environment.

A general feature of mathematics:

> The underlying structure often comes to light *after* the results have been proved.
> Our first intuitions can be quite different from our later ones.

(A big problem for teaching mathematics.)

Up to the 1930s it was generally believed that
to recognise the meaningful constituents,
we need to know the meanings.

True even in Bloomfield's *Language* (1933).

For logicians in the early 20th century, syntax was concatenation of symbols;
e.g. in Quine's *Mathematical Logic* (1940).

But the sentences in the logics of Tarski's 1933 paper are all built up by a set of 'fundamental operations', viz.
truth-functional compounds and quantifications.
The resulting components can be recognised from the syntax alone,
allowing definitions and proofs by recursion on syntax
(new in the 1920s).

We can derive a general form of Tarski's truth definition
by formalising the facts just mentioned.

Tarski's setting (around 1930):

We have a language L, all of whose sentences are meaningful.
We do know, but only intuitively, what the sentences mean;
in easy cases we recognise which are true and which aren't.

We must analyse this intuition into a mathematical form.

A very old doctrine:

> Sentences are constructed by building up meaningful constituents, starting from atomic expressions with known meaning.

For example Al Fārābī (10th century):

> … the imitation of the composition of meanings
> by the composition of expressions
> is by [linguistic] convention …

Definition. By a *constituent structure* we mean an ordered pair of sets $(\mathbb{E}, \mathbb{F})$, where the elements of $\mathbb{E}$ are called the *expressions* and the elements of $\mathbb{F}$ are called the *frames*, such that the four conditions below hold.

(e, f etc. are expressions. F, $G(\xi)$ etc. are frames.)

1. $\mathbb{F}$ is a set of nonempty partial functions on $\mathbb{E}$.

('Nonempty' means their domains are not empty.)

2. (Nonempty Composition) If $F(\xi_1, \ldots, \xi_n)$ and $G(\eta_1, \ldots, \eta_m)$ are frames, $1 \leqslant i \leqslant n$ and there is an expression

$$F(e_1, \ldots, e_{i-1}, G(f_1, \ldots, f_m), e_{i+1}, \ldots, e_n),$$

then

$$F(\xi_1, \ldots, \xi_{i-1}, G(\eta_1, \ldots, \eta_m), \xi_{i+1}, \ldots, \xi_n)$$

is a frame.

Note: If $H(\xi)$ is $F(G(\xi))$ then the existence of an expression $H(f)$ implies the existence of an expression $G(f)$.

3. (Nonempty Substitution) If $F(e_1, \ldots, e_n)$ is an expression, $n > 1$ and $1 \leqslant i \leqslant n$, then

$$F(\xi_1, \ldots, \xi_{i-1}, e_i, \xi_{i+1}, \ldots, \xi_n)$$

is a frame.

4. (Identity) There is a frame $1(\xi)$ such that for each expression e, $1(e) = e$.

We say e is a *constituent* of f if
f is $G(e)$ for some frame G.

$F(e_1, f, e_3)$ is *the result of replacing the occurrence of* e_2 *in second place in* $F(e_1, e_2, e_3)$ *by* f.
(This notion depends on F, of course.)

Definition We write $e \equiv_\mu f$ if for every 1-ary frame $G(\xi)$,

(a) $G(e)$ is in X if and only $G(f)$ is in X;

(b) if $G(e)$ is in X then $\mu(G(e)) = \mu(G(f))$.

We say e, f have *the same $\equiv_\mu$-value*, or for short *the same fregean value*, if $e \equiv_\mu f$.
We write $|e|_\mu$ for this fregean value
(determined only up to $\equiv_\mu$).

Leaving out (b), we get a purely syntactic equivalence relation, viz. $e \sim_\mu f$ if for every 1-ary frame $G(\xi)$,

(a) $G(e)$ is in X if and only $G(f)$ is in X;

Immediately by the definitions,

$$e \equiv_\mu f \;\Rightarrow\; e \sim_\mu f.$$

The function μ is relevant to $\sim_\mu$ only through its domain X; hence not at all if X is syntactically definable.

The lifting lemma

Let X be a set of expressions (for example the sentences) and $\mu : X \to Y$ any function.

We will define a relation $\equiv_\mu$ so that

$$e \equiv_\mu f$$

says that expressions e and f make the same contribution to μ-values of expressions in X.

In Tarski's case X is the set of sentences,
$\mu(e)$ the truth value of sentence e,
and we are asking what each constituent
contributes to the truth value of a sentence containing it.

In general the fact that $\equiv_\mu$ must be an equivalence relation forces us to the following definition.

(a) By assumption $F(e)$ is an expression,
$H(F(e))$ is in X for some $H(\xi)$, and $e \sim_\mu f$.

By Nonempty Composition $H(F(\xi))$ is a frame $G(\xi)$.
Since $e \sim_\mu f$ and $G(e)$ is in X, $G(f)$ is in X.
But $G(f)$ is $H(F(f))$, so $F(f)$ is an expression.

(b) Let $G(\xi)$ be any 1-ary frame such that $G(F(e))$ is an expression in X.
By Nonempty Composition $G(F(\xi))$ is a frame $J(\xi)$.

Since $e \sim_\mu f$ and $J(e)$ is in X, $J(f)$ is in X.
[This proves the Lemma with $\sim$ for $\equiv$.]

Since $e \equiv_\mu f$ and $J(e)$ is in X, $\mu(J(e)) = \mu(J(f))$.
This proves $\mu(G(F(e)) = \mu(G(F(f))$ as required. $\square$

In Tarski's 1933 truth definition, we can take formulas as the constituents, and X the set of sentences (= formulas with no free variables).

Then $e \sim f$ if and only if e and f have the same free variables.

LEMMA. Suppose $F(e_1, \ldots, e_n)$ is a constituent of some expression in X, and for each i, $e_i \equiv_\mu f_i$. Then:

(a) $F(f_1, \ldots, f_n)$ is an expression.

(b) $F(e_1, \ldots, e_n) \equiv_\mu F(f_1, \ldots, f_n)$.

[The same holds with $\sim$ for $\equiv$.]

For the **proof**, by Nonempty Substitution we can make the replacements one expression at a time.
So it suffices to prove the lemma when $n = 1$.

$$|F(e_1,\ldots,e_n)|_\mu = h_F(|e_1|_\mu,\ldots,|e_n|_\mu).$$

We call h_F the *Hayyan function* of F.

Abu Ḥayyān al-Andalusī (Egypt, 14th c.)
argued that such functions must exist,
from the fact that we can create and use new sentences.

Definition Let ω be a function defined on expressions.
A definition of ω is called *compositional* if
for each expression $F(e_1,\ldots,e_n)$,

$$\omega(F(e_1,\ldots,e_n))$$

is determined by F and the values $\omega(e_i)$.

So fregean values are compositional.

We say that X is *cofinal* if every expression is
a constituent of an expression in X.

In Tarski's languages, the set of sentences is cofinal.

Assume X is cofinal. Then by (b) of the Lemma,
if $e_i \equiv_\mu f_i$ for each i then $F(e_1,\ldots,e_n) \equiv_\mu F(f_1,\ldots,f_n)$
provided these expressions exist.
So F and the fregean values of the e_i
determine the fregean value of $F(e_1,\ldots,e_n)$.

Hence there is, for each n-ary frame F, an n-ary map
$h_F : V^n \to V$, where V is the class of $\sim_\mu$-values,
such that whenever $F(e_1,\ldots,e_n)$ is an expression,

$$|F(e_1,\ldots,e_n)|_\mu = h_F(|e_1|_\mu,\ldots,|e_n|_\mu).$$

We say an expression e is *atomic* if $e = F(f_1, \ldots, f_n)$ implies F is $1(\xi)$.

We say a frame F is *fundamental* if it is not 1 and is not the result of Composition or Substitution.

We say the language L is *well-founded* if every expression of L is got by applying fundamental frames (any number of times) to atomic expressions.

ABSTRACT TARSKI THEOREM Suppose L is a language with a constituent structure in which sentences are cofinal, L is well-founded and each sentence ϕ has a truth value $\mu(\phi)$. Let ν be the restriction of $|.|_\mu$ to atomic expressions. Then μ is definable by recursion on the complexity of constituents as follows:

- If e is atomic then $|e|_\mu = \nu(e)$.
- $|(F(e_1, \ldots, e_n)|_\mu = h_F(|e_1|_\mu, \ldots, |e_n|_\mu)$.
- If e is a sentence then $\mu(e) = p(|e|_\mu)$.

PROPOSITION. Suppose $e \equiv_\mu f$ and e is an expression in X. Then f is in X and $\mu(e) = \mu(f)$.

Proof. This is immediate from the definition,
by applying the identity frame $1(\xi)$. $\square$

So on X the relation $\equiv_\mu$ is a refinement of the relation $\mu(\xi_1) = \mu(\xi_2)$.

This guarantees there is a function p_μ (the *read-out function*) so that for each e in X,

$$\mu(e) = p_\mu(|e|_\mu).$$

One can easily give necessary and sufficient conditions for p to be the identity.
These conditions are met for Tarski's languages.

The abstract Tarski theorem shows that every ‘reasonable’ language has a Tarski truth definition.

Nothing is assumed about reference or satisfaction.
In fact Hintikka’s Independence-Friendly languages have Tarski truth definitions by this theorem, but provably these languages have no truth definition based on satisfaction.

For Tarski’s 1950s model-theoretic truth definition the value of sentence e is the class of structures in which e is true.

We can apply the same machinery to syllogistic logic.

The naturalness of the resulting truth definitions can be used to assess syntactic analyses of syllogistic sentences.

An old discredited theory
about ‘every man’ referring to the class of all men
turns out provably correct if we replace ‘reference’
by ‘fregean value’.

For Tarski’s truth definition the required value $|e|_\mu$ is

> the set of those assignments to free variables of e
> which satisfy e.

In 1933 Tarski defines instead the relation

> assignment α satisfies formula e.

The reason is technical:
it allows a more elementary truth definition.
In work with Kuratowski in 1930,
Tarski had defined the set of assignments.

In 1933 Tarski’s assignments are to all variables, not just those free in the formula.
The set of *these* assignments is not the fregean value, because it doesn’t determine $\sim_\mu$.
E.g.

$$P(x_1), \qquad P(x_1) \wedge (x_2 = x_2)$$

have the same ‘meaning’ but different fregean values.

Even in natural languages, most discrepancies between fregean value and intuitive ‘meaning’ seem to be of this kind; the meaning plus the $\sim_\mu$ class gives the fregean value.

In principle the theorem extends to natural languages too.

Three complications (among others):

- The choice of a constituent system.
- Truth need not be a suitable classifier of sentences.
- Finding fregean values with some cognitive content.