Tarski's Definition of Truth

In first-order logic with equality, we always have that $=^{\mathcal{A}} \;=\; \{\langle a, a\rangle \mid a \in |A|\}$ where $|A| = U_{\mathcal{A}}$ = the universe of $\mathcal{A}$. That is, the equality predicate symbol, "=", must always be interpreted as true equality:

$$\mathcal{A} \models t_1 = t_2 \quad \Leftrightarrow \quad t_1^{\mathcal{A}} = t_2^{\mathcal{A}} \; .$$

For $\mathcal{A} \in \text{STRUC}[\Sigma], t \in \text{term}(\Sigma), \varphi \in \mathcal{L}(\Sigma)$ we give the following inductive defintions of $t^{\mathcal{A}}$ and $\mathcal{A}(\varphi)$:

term base: for $x_i \in$ var, $x_i^{\mathcal{A}}$ is already given. (Each structure has a default value for each variable.)

term inductive: $(f_i(t_1, \ldots, t_{r_i}))^{\mathcal{A}} = f_i^{\mathcal{A}}(t_1^{\mathcal{A}}, \ldots, t_{r_i}^{\mathcal{A}})$

truth base: $\mathcal{A}(R_i(t_1, \ldots, t_{a_i})) =$ **if** $\langle t_1^{\mathcal{A}}, \ldots, t_{a_i}^{\mathcal{A}}\rangle \in R_i^{\mathcal{A}}$ **then** 1 **else** 0

truth inductive:

1. $\mathcal{A}(\neg\alpha) = 1 - \mathcal{A}(\alpha)$
2. $\mathcal{A}(\alpha \vee \beta) = \max(\mathcal{A}(\alpha), \mathcal{A}(\beta))$
3. $\mathcal{A}(\exists x_i(\alpha)) = \max_{a \in |\mathcal{A}|}((\mathcal{A}, x_i/a)(\alpha))$

$(\mathcal{A}, x_i/a)$ is the same structure as $\mathcal{A}$ with the single exception that $x_i^{(\mathcal{A}, x_i/a)} = a$, i.e., the default value of x_i in $(\mathcal{A}, x_i/a)$ is $a \in |\mathcal{A}|$.

Game-Theoretic Definition of Truth

The truth of a first-order formula corresponds to a two-person game: $\mathcal{G}(\mathcal{A}, \varphi)$ is the game on structure $\mathcal{A}$, formula φ. Assume that φ is in **negation normal form**, i.e., the quantifiers are $\forall, \exists$, the propositional connectives are $\wedge, \vee, \neg$ and all $\neg$'s have been pushed inside as far as possible using the de Morgan laws, so the only occurences of $\neg$'s are directly in front of atomic formulas. The truth game has two players named Dumbledore (**D**) and Gandalf (**G**). D is trying to prove that $\mathcal{A} \models \varphi$ and **G** is trying to prove that $\mathcal{A} \models \neg\varphi$.

In $\mathcal{G}(\mathcal{A}, \varphi)$,

game base: If φ is atomic, then if $\mathcal{A} \models \varphi$ then D wins, else **G** wins

game inductive:

1. If $\varphi = \alpha \vee \beta$, then D chooses one of the disjuncts: $\psi \in \{\alpha, \beta\}$ and the next position is $\mathcal{G}(\mathcal{A}, \psi)$.
2. If $\varphi = \alpha \wedge \beta$, then **G** chooses one of the conjuncts: $\psi \in \{\alpha, \beta\}$ and the next position is $\mathcal{G}(\mathcal{A}, \psi)$.
3. If $\varphi = \exists x_i(\psi)$, then D chooses an element $e \in |\mathcal{A}|$ and the next position is $\mathcal{G}((\mathcal{A}, x_i/e), \psi)$.
4. If $\varphi = \forall x_i(\psi)$, then **G** chooses an element $a \in |\mathcal{A}|$ and the next position is $\mathcal{G}((\mathcal{A}, x_i/a), \psi)$.

Theorem: For any vocabulary Σ, formula $\varphi \in \mathcal{L}(\Sigma)$ in negation normal form, and structure $\mathcal{A} \in \text{STRUC}[\Sigma]$, Tarski's definition of truth, and the game theoretic definition of truth are equivalent, i.e,

$$\begin{aligned} \mathcal{A} \models \varphi \quad &\Leftrightarrow \quad \text{D has a winning strategy for } \mathcal{G}(\mathcal{A}, \varphi) \text{ and,} \\ \mathcal{A} \not\models \varphi \quad &\Leftrightarrow \quad \textbf{G} \text{ has a winning strategy for } \mathcal{G}(\mathcal{A}, \varphi). \end{aligned}$$

Proof: This can be proved by induction on φ. It would be a good exercise for you to fill in the details. $\square$