Notre Dame Journal of Formal Logic
Volume 29, Number 4, Fall 1988

Definability in Self-Referential Systems

J. ZIMBARG SOBRINHO

Introduction Self-referential systems are theories formulated in a (typed) first-order language $\mathfrak{L}$, in whose intended interpretation the predicates refer to objects which are themselves predicates of $\mathfrak{L}$. In order to avoid the Russell–Zermelo paradox, Hiller and Zimbarg Sobrinho [1] introduced *self-referential systems with involution*: these are theories whose intended models admit an elementary embedding into itself, denoted by '*', and called *involution map.* As a consequence, the universe of predicates inherits a structural hierarchy of objects, classified into countably many *types.*

A peculiar property of types refers to the size of the universe of their corresponding objects: the larger a type, the smaller its domain, and for this reason, types have been suggestively taken as negative integers: 0, -1, -2, and so on.

The main properties satisfied by self-referential systems with involution were outlined in [2], and are the following:

(a) $\mathfrak{L}$ possesses unrestricted (or universal) variables
(b) all predicates are extensional
(c) the Comprehension axiom for starred formulas is true
(d) (Definability condition) every element in the universe of a realization is definable by a one-free-variable formula of $\mathfrak{L}$.

The first three clauses above can be directly expressed in our (typed) first-order language $\mathfrak{L}$ without any further ado. With respect to the Definability condition, however, it is not altogether clear how it could be formulated in a first-order version of self-reference, due to its obvious higher-order character. The purpose of this article is to present first-order axioms which, added to $\mathfrak{W}$ (see [1]), produce the same effect as the apparently stronger 'Definability condition'.

1 Hiller's problem Realizations of self-reference in which the Definability condition holds have been referred to as *intended models.* It is well-known that

Received July 7, 1986

the upward Löwenheim–Skolem theorem precludes the characterization of intended models solely by means of first-order sentences. Nevertheless, as we are mainly interested in first-order statements, it is natural to ask whether there is a (typed) first-order extension yielding to those, and only those, sentences true in all intended models of a given theory. These considerations led Hiller to formulate the following general question:

Problem 1.1 (Hiller's problem) *Let T be a consistent extension of $\mathfrak{W}$ formulated in a given self-referential language $\mathfrak{L}$. Find a theory $\mathfrak{F}(T)$, formulated in $\mathfrak{L}$, such that:*

(i) *every model of $\mathfrak{F}(T)$ is elementarily equivalent to an intended model of T*
(ii) *every intended model of T is a model of $\mathfrak{F}(T)$*
(iii) *$\mathfrak{F}(T)$ is the weakest consistent extension in which (i) and (ii) hold.*

The general solution to Hiller's problem is complex and remains open. Nevertheless, it is possible to give a complete solution for it in case T is an extension of the theory $\mathfrak{W}+$ (regularity); from now on, we will denote this theory by '$\mathfrak{W}_r$'.

2 Ordinal definability Denote by '$V = OD$' the statement expressing that every element of the universe is *ordinal definable*. It is fairly well-known that ordinal definability is a second-order concept, which, in the presence of the axiom of regularity, is expressible by a first-order sentence in the language of ZF. Having that in mind, our endeavor is to prove that if T extends $\mathfrak{W}_r$, then $\mathfrak{F}(T) = T + (V = OD)$ is a solution to Hiller's problem. Before going into the details, let us review some of the ordinal-definability notions applied to $\mathfrak{W}$, in order to fix our terminology.

Definition 2.1 (Ordinal Definability for $\mathfrak{W}$) *Let $\langle M_i, E: i \in \omega\rangle$ be a model of $\mathfrak{W}$ and let $P \in M_0$. We say that P is ordinal definable in $\mathfrak{M}$ if there exists a formula of $\mathfrak{L}$, $A(v_0, v_1)$ say, and an element $\alpha \in M_0$ such that*

(a) $\mathfrak{M} \models OR[\alpha]$, *where the formula $OR(x)$ expresses in ZF that 'x is an ordinal'*
(b) $\mathfrak{M} \models (A \wedge \exists! v_0 A)[P, \alpha]$.

We say that $\mathfrak{M}$ satisfies the ordinal-definability condition if every member $P \in M_0$ is ordinal definable in $\mathfrak{M}$.

The ordinal-definability condition is not directly expressible in $\mathfrak{L}$, since in order to do it, we need the notion of satisfaction. In $\mathfrak{W}_r$, however, this concept possesses a first-order counterpart, analogously to what happens in ZF. To see it, we prove the following:

Lemma 2.2 *Let $\mathfrak{M}$ be a model of $\mathfrak{W}_r$ satisfying the ordinal-definability condition. Then, for every $P \in M_0$ there is an ordinal $\alpha \in M_0$ and a ZF formula $B(v_0, v_1)$ such that*

$$\mathfrak{M} \models (OR(v_1) \wedge B \wedge \exists! v_0 B)[P, \alpha].$$

Proof: Let $V_i^{\mathfrak{M}}$ denote the internal object of $\mathfrak{M}$ satisfying the formula $\forall x(x \in y \leftrightarrow \exists x^i(x^i = x))$, for all $i < 0$, and let Ω_i be the least ordinal of $\mathfrak{M}$ not belonging to $V_i^{\mathfrak{M}}$. Pick any element $P \in M_0$. By the ordinal-definability condition, there exists a two-free-variable formula $A(v_0, v_1)$, and an ordinal β in $\mathfrak{M}$ such that

$$\mathfrak{M} \models (A \wedge \exists! v_0 A)[P, \beta].$$

The number of types in A is finite; so let us suppose that the lowest type in A is $-n$, for $n > 0$.

Our next claim is that there exists a formula $B_1(v_0, v_1, u_1, \ldots, u_n)$, having only variables of type 0, whose quantifiers are bounded, and such that

$$\mathfrak{M} \models \forall v_0 \forall v_1 (A \leftrightarrow B_1)[V_{-1}^{\mathfrak{M}}, \ldots, V_{-n}^{\mathfrak{M}}].$$

To construct B_1, it suffices to replace each variable v_m^i in A by a new variable x, and the quantifications $\exists v_m^i$ and $\forall v_m^i$ by the bounded quantifiers $(\exists x \in u_{-i})$ and $(\forall x \in u_{-i})$ respectively. It is being supposed that eventual collisions of variables have been carefully avoided, of course.

Next, since the axiom of regularity holds in $\mathfrak{M}$, each one of the $V_i^{\mathfrak{M}}$ coincides with V_{Ω_i} (recall that V_η is recursively defined by the following clauses: $V_0 = \varnothing$, and $V_\eta = \mathcal{P} \cup \{V_\xi : \xi \in \eta\}$, where $\mathcal{P}x$ is the power set of x). Denote by $R(u, v)$ the set-theoretical formula defining the V_η. Then, for $0 < -i \leq n$, the following holds:

$$\mathfrak{M} \models (R \wedge \exists! u R)[V_i^{\mathfrak{M}}, \Omega_i].$$

Now, consider the formula $B_2(v_0, v_1, w_1, \ldots, w_n)$ given by

$$B_2 \simeq \forall u_1 \ldots \forall u_n (R[u_1, w_1] \wedge \ldots R[u_n, w_n] \rightarrow B_1).$$

Then it is standard to prove that B_2 has only variables of type 0 and that

$$\mathfrak{M} \models \forall v_0 \forall v_1 (A \leftrightarrow B_2)[\Omega_{-1}, \ldots, \Omega_{-n}].$$

Finally, it is possible to collapse the n-tuple $\langle \beta, \Omega_{-1}, \ldots, \Omega_{-n} \rangle$ into a single ordinal α by means of ZF-definable pairing functions: so, let $J_n(v_1, w_1, \ldots, w_n, v_2)$ be a formula resulting from the combination of those pairing functions encoding the n-tuple of ordinals $\langle v_1, w_1, \ldots, w_{n-1} \rangle$ into a single ordinal v_2. Let $B(v_0, v_2)$ be the formula $\exists v_1 \exists w_1 \ldots \exists w_n (J_n \wedge B_2)$. Then, if α encodes $\langle \beta, \Omega_{-1}, \ldots \Omega_{-n} \rangle$, it follows that the ZF-formula B satisfies

$$\mathfrak{M} \models (B \wedge \exists! v_0 B)[P, \alpha].$$

In what follows, we present a few classical well-known results on ordinal definability for ZF Set Theory; they can be found, for instance, in Myhill and Scott [3].

Theorem 2.3 *Let $\mathfrak{M} = \langle M, E \rangle$ be a model of ZF. The following conditions are equivalent:*

(a) *every element of M is definable in $\mathfrak{M}$ by a one-free-variable formula of ZF having an ordinal as parameter*

(b) *for every $x \in M$, there exists an ordinal $\alpha \in M$ and a (internal) formula $\ulcorner A(v_0) \urcorner$ of the language $\ulcorner ZF \urcorner$, defining x in $\langle V_\alpha^{\mathfrak{M}}, E | V_\alpha^{\mathfrak{M}} \rangle$.*

The sentence

$$\forall x \exists \alpha \exists [A(v_0)] (\langle V_\alpha, \in \rangle \models [(A \wedge \exists! v_0 A)][x])$$

will be abbreviated by the expression '$V = OD$'.

As a consequence, we obtain

Theorem 2.4 *Let $\mathfrak{M} = \langle M_i, E \rangle$ be an intended model of $\mathfrak{W}_r$. Then $\mathfrak{M} \models (V = OD)$.*

This theorem is a direct consequence of Lemma 2.2, Theorem 2.3 and the fact that $\mathfrak{M}$ is a model of ZF.

As a direct consequence of theory ZF + $(V = OD)$, the universe possesses a canonical well-ordering definable in the language of ZF. It is denoted by $\leq_D$. To define it, given x, let $\langle \alpha_x, n_x \rangle$ be the pair associated to x according to the following clauses:

(i) α_x is the least ordinal for which x is definable in $\langle V_{\alpha_x}, \in | V_{\alpha_x} \rangle$.
(ii) n_x is the smallest Gödel number of a formula which defines it in $\langle V_{\alpha_x}, \in | V_{\alpha_x} \rangle$.

Then, the well-ordering $\leq_D$ can be expressed in ZF by the following formula:

$$x \leq_D y \quad \text{iff} \quad \alpha_x \leq \alpha_y \vee (\alpha_x = \alpha_y \wedge n_x \leq n_y).$$

Definition 2.5 (Definable substructure of a model for $\mathfrak{W}_r$) *Let $\mathfrak{M} = \langle \mathfrak{M}_i, E: i \in \omega \rangle$ be a model of $\mathfrak{W}_r$. We denote by $Def(M_0)$ the set of elements of M_0 for which there exists a one-free-variable formula of $\mathfrak{L}$, without parameters, which defines it in $\mathfrak{M}$. In symbols,*

$$\alpha \in Def(M_0) \text{ iff } \alpha \in M_0 \;\&\; \mathfrak{M} \models (A \wedge \exists! x A)[\alpha]$$

for some formula $A(x)$ of $\mathfrak{L}$.

The definable substructure of $\mathfrak{M}$ is the substructure whose universe is $Def(M_0)$. We denote it by '$Def(\mathfrak{M})$'

$$Def(\mathfrak{M}) = \langle Def(M_0) \cap M_i, E | Def(M_0): i \in \omega \rangle.$$

Lemma 2.6 *Let $\mathfrak{M}$ be a model of $\mathfrak{W}_r + (V = OD)$. Then the definable substructure $Def(\mathfrak{M})$ is an elementary substructure of $\mathfrak{M}$, and, moreover, it is an intended model of $\mathfrak{W}_r$. In symbols,*

$$Def(\mathfrak{M}) \preceq \mathfrak{M}.$$

This result is an immediate consequence of Tarski's criterion for elementary substructures, and an analogous proof of a similar result can be found in Zimbarg Sobrinho [4]. We omit details.

Now, our main result:

Theorem 2.7 *Let T be a consistent extension of $\mathfrak{W}_r$. Then, the theory*

$$\mathfrak{F}(T) = T + (V = OD)$$

is a solution for Hiller's problem.

This result can be derived from Theorem 2.4 and Lemma 2.6. As a byproduct of the results presented above we simply mention that the axiom schema $A_6^{(n)}$ stated in [1] is derivable in the theory $\mathfrak{W}_r + (V = OD)$.

REFERENCES

[1] Hiller, A. P. and J. Zimbarg Sobrinho, "Self-reference with negative types," *The Journal of Symbolic Logic*, vol. 49 (1984), pp. 754–773.

[2] Hiller, A. P. and J. Zimbarg Sobrinho, "Fundamentals of self-reference," to appear.

[3] Myhill, J. and D. Scott, "Ordinal definability," pp. 271–278 in *Axiomatic Set Theory, Proceedings of Symposia in Pure Mathematics*, vol. XIII, Part 1, ed. D. Scott, American Mathematical Society, Providence, Rhode Island, 1971.

[4] Zimbarg Sobrinho, J., "On the consistency of self-referential systems," *The Journal of Symbolic Logic*, vol. 52 (1987), pp. 425–436.

Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo, Brazil