# Ehrenfeucht's Game-Theoretic Characterization of First-Order Elementary Equivalence

John Peloquin

June 30, 2006

**Abstract**

In this paper, we provide a brief exposition of Ehrenfeucht's game-theoretic characterization of first-order elementary equivalence. Our treatment is intended to be an introduction, and we provide definitions for the various concepts involved. We borrow heavily from [1] in our treatment.

## 1 Preliminaries

Elementary equivalence is a central topic in mathematical logic, and Ehrenfeucht provides a simple and intuitive game-theoretical characterization. Before presenting it, we first review some basics of first-order logic, starting with the definition of a first-order language. Readers already familiar with this material may skip to the next section.

We begin by defining an alphabet:

**1.1 Definition.** Let $\mathscr{A}$ contain the following symbols:

1. $x_0, x_1, x_2, \ldots$ (*variable symbols*)
2. $\widehat{=}$ (*equality symbol*)
3. $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$ (*boolean symbols*)
4. $\forall, \exists$ (*quantifier symbols*)
5. ), ( (*parentheses*)

We call $\mathscr{A}$ an *alphabet* and the elements of $\mathscr{A}$ *logical symbols*.

The symbols in $\mathscr{A}$ are common to all of the languages we construct. In constructing a particular language, we must first specify a symbol set:

**1.2 Definition.** Let $S$ contain the following:

1. For each $n \geq 1$, zero or more *$n$-ary relation symbols*.
2. For each $n \geq 1$, zero or more *$n$-ary function symbols*.
3. Zero or more *constant symbols*.

With a given symbol set $S$, we can define the set of $S$-terms and then the set of $S$-formulas, the latter of which will form our language. We denote by $(\mathscr{A} \cup S)^*$ the set of all finite sequences over $\mathscr{A} \cup S$.

**1.3 Definition.** Let $S$ be a symbol set. We define the set $T^S$ of *$S$-terms* to be the intersection of all sets $T \subseteq (\mathscr{A} \cup S)^*$ which satisfy the following closure conditions:

1. If $x \in \mathscr{A}$ is a variable symbol, then $x \in T$.
2. If $c \in S$ is a constant symbol, then $c \in T$.
3. If $t_1, \ldots, t_n \in T$ and $f \in S$ is an $n$-ary function symbol, then $f\, t_1 \cdots t_n \in T$.

Note that $T^S$ is well-defined since there exist sets satisfying the closure conditions—for example, $(\mathscr{A} \cup S)^*$—and $T^S$ is the smallest set satisfying these conditions.

We similarly define the set of $S$-formulas:

**1.4 Definition.** Let $S$ be a symbol set and $T^S$ be the set of $S$-terms. We define the set $L^S$ of *$S$-formulas* to be the intersection of all sets $L \subseteq (\mathscr{A} \cup S)^*$ which satisfy the following closure conditions:

1. If $t_1, t_2 \in T^S$, then $t_1 \mathrel{\widehat{=}} t_2 \in L$.
2. If $t_1, \ldots, t_n \in T^S$ and $R \in S$ is an $n$-ary relation symbol, then $R\, t_1 \cdots t_n \in L$.
3. If $\varphi \in L$, then $\neg\varphi \in L$.
4. If $\varphi_1, \varphi_2 \in L$, then
$$\{(\varphi_1 \wedge \varphi_2), (\varphi_1 \vee \varphi_2), (\varphi_1 \to \varphi_2), (\varphi_1 \leftrightarrow \varphi_2)\} \subseteq L$$
5. If $\varphi \in L$ and $x \in \mathscr{A}$ is a variable symbol, then
$$\{\forall x\varphi, \exists x\varphi\} \subseteq L$$

As with the set of $S$-terms, the set of $S$-formulas is well-defined.

Based on Definitions 1.3 and 1.4, one can establish a number of important syntactic results including readability and unique readability for terms and formulas. The latter results state that each term and formula can be parsed syntactically in exactly one way, and these results are required to justify definitions and proofs by closure induction on terms and formulas. We omit their statement and proof here, as the details are not central to our investigation.

Consider the formulas

$$\forall x(x < y) \qquad \forall x \exists y(x < y)$$

In the formula on the left, $y$ does not appear within the context of a quantification $\forall y$ or $\exists y$, while in the formula on the right this is the case. We distinguish these two formulas on this basis by saying that $y$ is a *free variable* in the formula on the left, while $y$ is a *bound variable* in the formula on the right. Formally, we define the set of free variables for a formula recursively, starting with the set of variables in a term:

**1.5 Definition.** Let $S$ be a symbol set and $T^S$ be the set of $S$-terms. We define the function $\mathrm{var}(t)$ recursively on $T^S$ as follows:

1. If $t = x$ for some variable symbol $x \in \mathscr{A}$, then $\mathrm{var}(t) = \{x\}$.
2. If $t = c$ for some constant symbol $c \in S$, then $\mathrm{var}(t) = \emptyset$.
3. If $t = f\,t_1 \cdots t_n$, where $t_1, \ldots, t_n \in T^S$ and $f \in S$ is an $n$-ary function symbol, then

$$\mathrm{var}(t) = \mathrm{var}(t_1) \cup \cdots \cup \mathrm{var}(t_n)$$

By unique readability (mentioned above), $\mathrm{var}(t)$ is well-defined for all $t \in T^S$.

**1.6 Definition.** Let $S$ be a symbol set and $L^S$ be the set of $S$-formulas. We define the function $\mathrm{free}(\varphi)$ recursively on $L^S$ as follows:

1. If $\varphi = t_1 \mathrel{\widehat{=}} t_2$ for $t_1, t_2 \in T^S$, then $\mathrm{free}(\varphi) = \mathrm{var}(t_1) \cup \mathrm{var}(t_2)$.
2. If $\varphi = R\,t_1 \cdots t_n$ for $n$-ary $R \in S$ and $t_1, \ldots, t_n \in T^S$, then

$$\mathrm{free}(\varphi) = \mathrm{var}(t_1) \cup \cdots \cup \mathrm{var}(t_n)$$

3. If $\varphi = \neg\psi$, then $\mathrm{free}(\varphi) = \mathrm{free}(\psi)$.
4. If $\varphi = (\psi_1 * \psi_2)$, then $\mathrm{free}(\varphi) = \mathrm{free}(\psi_1) \cup \mathrm{free}(\psi_2)$, for $* \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$.
5. If $\varphi = Qx\psi$, then $\mathrm{free}(\varphi) = \mathrm{free}(\psi) \setminus \{x\}$, for $Q \in \{\forall, \exists\}$.

Again by unique readability, $\mathrm{free}(\varphi)$ is well-defined for all $\varphi \in L^S$. We can now formally define what it means for a variable to appear freely in a formula:

**1.7 Definition.** Let $\varphi \in L^S$ be an $S$-formula and $x \in \mathscr{A}$ be a variable symbol. We say that $x$ is a *free variable in* $\varphi$ if $x \in \mathrm{free}(\varphi)$. If $x$ appears in $\varphi$ but $x \notin \mathrm{free}(\varphi)$, we say that $x$ is a *bound variable in* $\varphi$.

This allows us to classify $S$-formulas:

**1.8 Definition.** Let $L^S$ be given. Set

$$L_0^S = \{\varphi \in L^S \mid \mathrm{free}(\varphi) = \emptyset\}$$

The elements of $L_0^S$ are precisely the $S$-formulas with no free variables, and we call them *$S$-sentences*.

While we have discussed the syntax of first-order languages, we have not discussed semantics. In particular, we have not discussed what it means for a formula to be 'true', and under what conditions this occurs. To formulate precisely the notion of truth for a formula, we must define several constructs:

**1.9 Definition.** Let $S$ be a symbol set. An *$S$-structure* is an ordered pair $\mathfrak{A} = (A, I)$ consisting of a nonempty *universe set* $A$ and a function $I$ with $S \subseteq \mathrm{dom}(I)$ such that

1. For each $n$-ary relation symbol $R \in S$, $I(R)$ is an $n$-ary relation on $A$.

2. For each $n$-ary function symbol $f \in S$, $I(f)$ is an $n$-ary function on $A$.
3. For each constant $c \in S$, $I(c) \in A$.

**1.10 Definition.** Let $\mathfrak{A}$ be an $S$-structure. Then an $\mathfrak{A}$-*assignment* is a map

$$\alpha : \{x_i \in \mathscr{A} \mid i = 0, 1, 2, \ldots\} \to A$$

**1.11 Definition.** Let $S$ be a symbol set. Then an $S$-*interpretation* is an ordered pair $\mathfrak{I} = (\mathfrak{A}, \alpha)$ consisting of an $S$-structure $\mathfrak{A}$ and and $\mathfrak{A}$-assignment $\alpha$.

As its name suggests, an $S$-interpretation interprets all of the variable components of a given $S$-language. It assigns all relation, function, constant, and variable symbols to real relations, functions, constants, and objects on its domain. In other words, an $S$-interpretation can be seen as providing meaning to formulas in $L^S$ by associating real mathematical objects with the symbols in $L^S$.

With 'meaning' made precise for an $S$-formula, we can define 'truth' (in an $S$-structure) by means of a satisfaction relation '$\models$' defined recursively on $L^S$. First we need to extend the notion of an assignment:

**1.12 Definition.** Let $\mathfrak{A}$ be an $S$-structure and $\alpha$ be an $\mathfrak{A}$-assignment. We extend $\alpha$ to a function $\overline{\alpha}$ on $T^S$ recursively. For $t \in T^S$,

1. If $t = x$ for $x \in \mathscr{A}$, then $\overline{\alpha}(t) = \alpha(t)$.
2. If $t = c$ for $c \in S$, then $\overline{\alpha}(t) = I(c)$.
3. If $t = f\, t_1 \cdots t_n$ for $t_1, \ldots, t_n \in T^S$ and $n$-ary $f \in S$, then

$$\overline{\alpha}(t) = I(f)(\overline{\alpha}(t_1), \ldots, \overline{\alpha}(t_n))$$

By unique readability, $\overline{\alpha}$ is well-defined and is a unique extension of $\alpha$ (satisfying the conditions above). The function $\overline{\alpha}$ simply associates with each $S$-term its interpretation in $\mathfrak{A}$, as $\alpha$ does with variable symbols only.

For $a \in A$, we denote by $\alpha\frac{a}{x}$ the map agreeing with $\alpha$ everywhere except possibly $x$ and mapping $x$ to $a$.

**1.13 Definition.** Let $\mathfrak{I} = (\mathfrak{A}, \alpha)$ be an $S$-interpretation. For $\varphi \in L^S$, we define the *satisfaction relation*

$$\mathfrak{I} \models \varphi$$

recursively on $\varphi$:

| | | |
|---|---|---|
| $\mathfrak{I} \models t_1 \mathrel{\widehat{=}} t_2$ | iff | $\overline{\alpha}(t_1) = \overline{\alpha}(t_2)$. |
| $\mathfrak{I} \models R\, t_1 \cdots t_n$ | iff | $(\overline{\alpha}(t_1), \ldots, \overline{\alpha}(t_n)) \in I(R)$. |
| $\mathfrak{I} \models \neg\psi$ | iff | not $\mathfrak{I} \models \psi$. |
| $\mathfrak{I} \models (\psi_1 \wedge \psi_2)$ | iff | $\mathfrak{I} \models \psi_1$ and $\mathfrak{I} \models \psi_2$. |
| $\mathfrak{I} \models (\psi_1 \vee \psi_2)$ | iff | $\mathfrak{I} \models \psi_1$ or $\mathfrak{I} \models \psi_2$. |
| $\mathfrak{I} \models (\psi_1 \to \psi_2)$ | iff | not $\mathfrak{I} \models \psi_1$, or $\mathfrak{I} \models \psi_2$. |
| $\mathfrak{I} \models (\psi_1 \leftrightarrow \psi_2)$ | iff | $\mathfrak{I} \models \psi_1$ if and only if $\mathfrak{I} \models \psi_2$. |
| $\mathfrak{I} \models \forall x\psi$ | iff | for all $a \in A$, $(\mathfrak{A}, \alpha\frac{a}{x}) \models \psi$. |
| $\mathfrak{I} \models \exists x\psi$ | iff | there exists $a \in A$ such that $(\mathfrak{A}, \alpha\frac{a}{x}) \models \psi$. |

If $\mathfrak{I} \models \varphi$, we say that $\mathfrak{I}$ *satisfies* $\varphi$ or $\mathfrak{I}$ *models* $\varphi$; we may also say $\varphi$ *is true under* $\mathfrak{I}$.

From Definition 1.13, one can derive many important semantic results. One of the most significant for our purposes states (roughly) that all that matters in determining whether a formula is true under an interpretation (aside from the interpretation of the nonlogical symbols) is the assignment of the free variables in the formula by the interpretation. Formally,

**1.14 Theorem.** *Let* $\mathfrak{A}$ *be an* $S$*-structure,* $\alpha_1$ *and* $\alpha_2$ *be* $\mathfrak{A}$*-assignments, and* $\varphi \in L^S$*. If* $\alpha_1$ *and* $\alpha_2$ *agree on* $\text{free}(\varphi)$*, then*

$$(\mathfrak{A}, \alpha_1) \models \varphi \quad \textit{iff} \quad (\mathfrak{A}, \alpha_2) \models \varphi$$

The proof uses induction on terms and then formulas, and we omit it. Note that for $\varphi \in L_0^S$, *any* two $\mathfrak{A}$-assignments agree on $\text{free}(\varphi)$, hence we may just write

$$\mathfrak{A} \models \varphi$$

if $(\mathfrak{A}, \alpha) \models \varphi$ for some $\mathfrak{A}$-assignment $\alpha$. This leads us to two important definitions:

**1.15 Definition.** Let $\mathfrak{A}$ be an $S$-structure. We define the (first-order) *theory* $\text{Th}(\mathfrak{A})$ of $\mathfrak{A}$ as follows:

$$\text{Th}(\mathfrak{A}) = \{\varphi \in L_0^S \mid \mathfrak{A} \models \varphi\}$$

The theory of a structure is thus the set of all true sentences in the structure. This set is important because it tells us about the nature of the structure apart from the specifics of particular elements. We note when two structures have the same theory:

**1.16 Definition.** Let $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures. Then we say $\mathfrak{A}$ *and* $\mathfrak{B}$ *are elementarily equivalent*, and we write $\mathfrak{A} \equiv \mathfrak{B}$, if $\text{Th}(\mathfrak{A}) = \text{Th}(\mathfrak{B})$.

# 2 Finite Isomorphisms and Fraïssé's Theorem

Ehrenfeucht's game-theoretic characterization of elementary equivalence is based on Fraïssé's characterization by means of finite isomorphisms between structures. In this section, we state Fraïssé's Theorem without proof.

First we define a partial isomorphism:

**2.1 Definition.** Let $\mathfrak{A} = (A, I_A)$ and $\mathfrak{B} = (B, I_B)$ be $S$-structures. We call $\pi$ a *partial isomorphism from* $\mathfrak{A}$ *to* $\mathfrak{B}$ if $\pi$ is an injective map with $\text{dom}(\pi) \subseteq A$ and $\text{ran}(\pi) \subseteq B$ such that the following properties hold:

1. For all $n$-ary relation symbols $R \in S$ and $a_0, \ldots, a_{n-1} \in \text{dom}(\pi)$,

$$(a_0, \ldots, a_{n-1}) \in I_A(R) \quad \text{iff} \quad (\pi(a_0), \ldots, \pi(a_{n-1})) \in I_B(R)$$

2. For all $n$-ary function symbols $f \in S$ and $a_0, \ldots, a_n \in \text{dom}(\pi)$,

$$I_A(f)(a_0, \ldots, a_{n-1}) = a_n \quad \text{iff} \quad I_B(f)(\pi(a_0), \ldots, \pi(a_{n-1})) = \pi(a_n)$$

3. For all constant symbols $c \in S$ and $a \in \mathrm{dom}(\pi)$,

$$I_A(c) = a \quad \text{iff} \quad I_B(c) = \pi(a)$$

We denote by $\mathrm{Part}(\mathfrak{A}, \mathfrak{B})$ the set of all partial isomorphisms from $\mathfrak{A}$ to $\mathfrak{B}$.

Intuitively, a partial isomorphism preserves structure between subsets of the universes of two structures. In characterizing elementary equivalence, we require partial isomorphisms that admit of finite extension. We define a hierarchy of such partial isomorphisms:

**2.2 Definition.** Let $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures. We say that $\mathfrak{A}$ *and* $\mathfrak{B}$ *are finitely isomorphic* if there exists a sequence $(I_n)$, $n = 0, 1, 2, \ldots$ of nonempty sets of partial isomorphisms from $\mathfrak{A}$ to $\mathfrak{B}$ satisfying the following properties:

1. *Forth property*: If $p \in I_{n+1}$ and $a \in A$, then there exists a $q \in I_n$, $p \subseteq q$, such that $a \in \mathrm{dom}(q)$.
2. *Back property*: If $p \in I_{n+1}$ and $b \in B$, then there exists a $q \in I_n$, $p \subseteq q$, such that $b \in \mathrm{ran}(q)$.

We write $(I_n) : \mathfrak{A} \cong_f \mathfrak{B}$.

Note that a partial isomorphism $p \in I_n$ can be extended $n$ times using the back and forth properties. Since $I_n$ is nonempty for each $n = 0, 1, 2, \ldots$ we can find, for arbitrarily large $n$, partial isomorphisms admitting of extension $n$ times.

We can now state Fraïssé's Theorem:

**2.3 Theorem (Fraïssé).** *Let $S$ be a finite symbol set and $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures. Then*

$$\mathfrak{A} \equiv \mathfrak{B} \quad \textit{iff} \quad \mathfrak{A} \cong_f \mathfrak{B}$$

# 3 Ehrenfeucht's Game-Theoretic Characterization

Ehrenfeucht describes in a game-theoretic context the existence of a finite isomorphism between two structures. We obtain that two structures are finitely isomorphic if and only if a certain player wins the Ehrenfeucht game associated with the structures. By Fraïssé's Theorem, it follows that two structures are elementarily equivalent if and only if that player wins the game.

**3.1 Definition.** Let $S$ be a symbol set and $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures with $A$ and $B$ disjoint. The *Ehrenfeucht game* $G(\mathfrak{A}, \mathfrak{B})$ associated with $\mathfrak{A}$ and $\mathfrak{B}$ is played by two players, $P_1$ and $P_2$, according to the following rules: for each *play* of the game,

1. $P_1$ chooses a natural number $m \geq 1$ to be the number of moves that each of $P_1$ and $P_2$ must make in the play.
2. Moves in the game are made alternately by $P_1$ and $P_2$, starting with $P_1$.
3. Each move consists of a player choosing an element from $A \cup B$.

4. If $P_1$ chooses an element $a_i \in A$ during $P_1$'s $i$-th move, then $P_2$ chooses an element $b_i \in B$ during $P_2$'s $i$-th move.
5. If $P_1$ chooses an element $b_i \in B$ during $P_1$'s $i$-th move, then $P_2$ chooses an element $a_i \in A$ during $P_2$'s $i$-th move.

After completion of a play, elements $a_1, \ldots, a_m \in A$ and $b_1, \ldots, b_m \in B$ have been chosen. We say $P_2$ *wins the play* if the map $a_i \mapsto b_i$ establishes a partial isomorphism from $\mathfrak{A}$ to $\mathfrak{B}$. We say that $P_2$ *wins the game* $G(\mathfrak{A}, \mathfrak{B})$ if it is possible for $P_2$ to win each play of $G(\mathfrak{A}, \mathfrak{B})$.

We now obtain the following theorem:

**3.2 Theorem.** *Let $S$ be a symbol set and $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures. Then*

$$\mathfrak{A} \cong_f \mathfrak{B} \quad \textit{iff} \quad P_2 \textit{ wins } G(\mathfrak{A}, \mathfrak{B})$$

*Proof.* We may assume without loss of generality that $A$ and $B$ are disjoint. Suppose $(I_n) : \mathfrak{A} \cong_f \mathfrak{B}$. Set

$$I'_n = \{\, p \mid \text{there exists } q \in I_n \text{ such that } p \subseteq q \,\}$$

Then $(I'_n) : \mathfrak{A} \cong_f \mathfrak{B}$. Now, for an arbitrary play of $G(\mathfrak{A}, \mathfrak{B})$, if $P_1$ chooses move count $m \geq 1$, then for every $a_i \in A$ (or $b_i \in B$) that $P_1$ chooses on the $i$-th move, $P_2$ can choose on the $i$-th move a corresponding element $b_i \in B$ (respectively $a_i \in A$) such that there exists $p_i \in I'_{m-i}$ with

$$p_i : a_j \mapsto b_j \qquad (1 \leq j \leq i)$$

(Note that we have constructed $(I'_n)$ so that $\emptyset \in I'_m$ and so that $p_{i+1}$ may be obtained from $p_i$ by adjoining at most one pair.) After $m$ moves, it follows that $p_m$ is a partial isomorphism from $\{a_1, \ldots, a_m\}$ to $\{b_1, \ldots, b_m\}$, so $P_2$ wins the play. Since the play was arbitrary, it follows that $P_2$ wins $G(\mathfrak{A}, \mathfrak{B})$ as desired.

Conversely, suppose $P_2$ wins $G(\mathfrak{A}, \mathfrak{B})$. We construct a partial isomorphism hierarchy between $\mathfrak{A}$ and $\mathfrak{B}$. For each $n$, say that $p \in \mathrm{Part}(\mathfrak{A}, \mathfrak{B})$ is *$n$-playable by $P_2$* if the following holds:

1. $\mathrm{dom}(p) = \{a_1, \ldots, a_k\}$, and
2. It is possible for $P_2$ to win any play of $G(\mathfrak{A}, \mathfrak{B})$ where
   (a) $P_1$ chooses a move count $m + k$ with $m \geq n$, and
   (b) Elements $a_1, \ldots, a_k$ and $p(a_1), \ldots, p(a_k)$ are chosen within each player's first $k$ moves.

Now set

$$I_n = \{\, p \in \mathrm{Part}(\mathfrak{A}, \mathfrak{B}) \mid p \text{ is } n\text{-playable by } P_2 \,\}$$

Since $P_2$ wins $G(\mathfrak{A}, \mathfrak{B})$, we have $\emptyset \in I_n$ for each $n$, and it is immediate that $(I_n)$ also satisfies the back and forth properties. Thus $(I_n) : \mathfrak{A} \cong_f \mathfrak{B}$ as desired. □

From Theorem 3.2 and Fraïssé's Theorem, we obtain Ehrenfeucht's Theorem as an immediate corollary:

**3.3 Theorem (Ehrenfeucht).** *Let $S$ be a finite symbol set and $\mathfrak{A}$ and $\mathfrak{B}$ be $S$-structures. Then*

$$\mathfrak{A} \equiv \mathfrak{B} \quad \textit{iff} \quad P_2 \textit{ wins } G(\mathfrak{A}, \mathfrak{B})$$

# References


[1] Ebbinghaus, H.–D. and J. Flum and W. Thomas. *Mathematical Logic*, 2nd ed. New York: Springer, 1994.

[2] Slaman, Theodore A. and W. Hugh Woodin. *Mathematical Logic: The Berkeley Undergraduate Course*. Berkeley: 2006.