JOURNAL OF COMPUTER AND SYSTEM SCIENCES 32, 393–406 (1986)

# First-Order Logic and Star-Free Sets

DOMINIQUE PERRIN AND JEAN-ERIC PIN

*L.I.T.P., 2 Place Jussieu, 75251 Paris Cedex 05 France*

Received April 9, 1985; revised July 17, 1985

It was the work of Büchi [2] that first showed how to use certain formulas of mathematical logic in order to describe properties of languages recognized by finite automata. These formulas (known as monadic second-order formulas) are built up from variables $x$, $y$,..., set variables $X$, $Y$,..., a binary relation symbol $<$ and a set

$$\pi = \{\pi_a \mid a \in A\}$$

of unary relation symbols in one-to-one correspondence with the alphabet $A$. Starting with atomic formulas of the type

$$x < y, \qquad \pi_a x, \qquad Xx, \qquad x = y,$$

formulas are built up in the usual way by means of the connectives $\neg$, $\vee$, $\wedge$ and the quantifiers $\exists$ and $\forall$ bounding up both types of variables.

Now, we say that a word $w$ on the alphabet $A$ satisfies such a sentence $\phi$ if $\phi$ is true when variables are interpreted as integers, set-variables are interpreted as set of integers and the formula $\pi_a x$ is interpreted as "the letter in position $x$ in $w$ is an $a$."

McNaughton [3] was the first to consider the case where the set of formulas is restricted to *first-order*, that is, when set-variables are ignored. He proved that the languages defined in this way are precisely the *star-free* languages, that is, all languages obtained from finite languages by boolean operations and concatenation product.

Later on, star-free languages have been considerably studied. First, a fundamental result of Schützenberger shows that star-free languages are exactly the languages recognized by an *aperiodic* finite monoid (i.e., a monoid all of whose groups are trivial). Further on, a great number of subclasses of star-free languages have been studied [6]. Among the most famous, let us quote the locally testable languages studied by McNaughton and Brzozowski and Simon and the piecewise testable languages, introduced by Simon.

Star-free languages are defined by two types of operations: boolean operations on one hand and concatenation product on the other hand. This naturally defines a hierarchy based on the alternative use of these operations. The hierarchy was originally introduced by Brzozowski who showed with Knast [1] that the inclusion was proper on each level. Furthermore the class of locally testable languages



0022-0000/86 $3.00

appears as the starting point of a natural subhierarchy inside the level 1. However many problems are still open and especially the decidability problem of each level, which is solved for the level 1 only.

Coming back to logical formulas, Thomas [9] showed that this hierarchy of languages corresponds in a very natural way with a classical hierarchy of first-order logic based on the alternation of existential and universal quantifiers.

The aim of this paper is to give a substantially different proof of the result of Thomas together with a generalization to the case of infinite words. In Sections 1–3, we give a proof of his theorem (Theorem 1.1) which does not rely on any previous knowledge in logic. This should make the proof more accessible to all those who feel more comfortable with automata and formal languages than with logical formulas. As a matter of fact, we use for technical reasons a hierarchy of aperiodic languages which is slightly different from the Brzozowski hierarchy. This hierarchy has been introduced by Straubing in [7]. Its level 0 is composed of the trivial Boolean algebra, and its level 1 is the family of piecewise testable languages. These will therefore correspond to Boolean combinations of existential formulas. In the last section, we prove an extension of the Thomas result to infinite words. The interpretation of logical formulas over infinite words instead of finite ones was, by the way, considered since the early paper of Büchi [2]. The correspondence between first-order formulas and star-free languages of infinite words was established by Thomas [8]. Among several possible equivalent definitions for star-free languages of infinite words (see [8, 5]) we choose the following one: it is the closure of the trivial family reduced to the empty set by Boolean operations and left product by star-free languages of finite words. We prove (Theorem 4.1) that, with these definitions, an exact analog of Theorem 1.1 holds for infinite words. We finally prove an extension to the case of two-sided infinite words (Theorem 4.5) also considered in [4].

## 1. The First-Order Logic on Words

Let $A$ be a fixed finite set called the alphabet. We denote by $A^*$ the set of words on the alphabet $A$. A subset of $A^*$ is called a language. For a word $w$ in $A^*$, we denote by $|w|$ its length. We define a set of logical formulas by considering the set

$$\mathscr{L} = \{<\} \cup \{\pi_a \mid a \in A\}$$

where $<$ denotes a symbol of binary relation and for each letter $a \in A$, $\pi_a$ is a symbol of unary relation.

Atomic formulas of $\mathscr{L}$ are formulas of the type $x < y$, $x = y$, and $\pi_a x$, where $a \in A$ and $x$, $y$ are variables.

Formulas of $\mathscr{L}$ are then constructed from atomic formulas by using connectives $\neg$, $\vee$, $\wedge$ and quantifiers $\forall$, $\exists$ bounding variables. For convenience, we also consider the formulas 0 (false) and 1 (true).

To each sentence (i.e., closed formula) $\phi$ of $\mathscr{L}$ is associated the set $L(\phi)$ of all words $w$ of $A^*$ such that $\phi$ is satisfied when variables are interpreted as integers on the set $\{1,..., |w|\}$, the relation $<$ is interpreted as the usual relation on integers and the formula $\pi_a x$ is interpreted as: "the letter of index $x$ in $w$ is an $a$."

EXAMPLE. Let $A = \{a, b\}$ if $\phi$ is the formula

$$\exists x\, \pi_a x$$

then $L(\phi)$ is composed of all words containing at least an occurrence of the letter $a$.

Observe that $L(0) = \varnothing$ and $L(1) = A^*$.

We define by induction two sequences $\mathscr{A}_n(A^*)$ and $\mathscr{B}_n(A^*)$ of sets of subsets of $A^*$ as follows

$$\mathscr{A}_0(A^*) = \mathscr{B}_0(A^*) = \{\varnothing, A^*\}.$$

Then, for $n \geqslant 0$, $\mathscr{B}_{n+1}(A^*)$ is the polynomial closure of $\mathscr{A}_n(A^*)$, that is the smallest class $C$ of subsets of $A^*$ such that

(i) $\mathscr{A}_n(A^*) \subset C$.

(ii) $C$ is closed under union and intersection.

(iii) For every $X, Y \in C$ and $a \in A$, $XaY \in C$.

Finally, $\mathscr{A}_{n+1}(A^*)$ is the boolean algebra generated by $\mathscr{B}_{n+1}(A^*)$ (with respect to the usual operations: intersection and complement).

Similarly, we define classes of formulas $\Sigma_n$ and $\Gamma_n$ by setting

$$\Sigma_0 = \Gamma_0 = \{\text{quantifier-free formulas of } \mathscr{L}\}.$$

Then for $n \geqslant 0$, $\Sigma_{n+1}$ is the smallest set $\Delta$ of formulas of $\mathscr{L}$ such that

(i) $\Gamma_n \subset \Delta$.

(ii) if $\varphi, \psi \in \Delta$, then $\varphi \vee \psi$, $\varphi \wedge \psi \in \Delta$.

(iii) If $\phi \in \Delta$ and $x$ is variable, then $\exists x\, \phi \in \Delta$.

Finally $\Gamma_{n+1}$ is the closure of $\Sigma_{n+1}$ by the connectives $\neg$, $\vee$, $\wedge$.

We can now state a slightly different version of a theorem of Thomas [9].

THEOREM 1.1. *For every $n \geqslant 0$ and for all $X \subset A^*$, we have $X \in \mathscr{A}_n(A^*)$ (resp. $\mathscr{B}_n(A^*)$) iff there exists a sentence $\phi \in \Gamma_n$ (resp. $\Sigma_n$) such that $X = L(\phi)$.*

It is easily proved that the set $\mathscr{A}(A^*) = \bigcup_{n \geqslant 0} \mathscr{A}_n(A^*)$ is the set of all star-free languages of $A^*$, that is, the smallest set of subsets of $A^*$ containing finite subsets and closed under boolean operations and concatenation product. Therefore we have

COROLLARY 1.2 [3]. *A language $X$ is star-free iff there exists a first-order sentence $\phi$ such that $X = L(\phi)$.*

In the next two sections, we shall give a proof of Theorem 1.1. In the sequel we shall use the notation $[k]$ for the set $\{1,..., k\}$.

## 2. FROM LANGUAGES TO FORMULAS

In this section, we prove by induction on $n \geqslant 0$ that for every language $X \in \mathscr{A}_n(A^*)$ (resp. $\mathscr{B}_n(A^*)$) there exists a sentence $\phi \in \Gamma_n$ (resp. $\Sigma_n$) such that $X = L(\phi)$. We first need a preliminary result.

PROPOSITION 2.1. *For every $n \geqslant 0$ and for each sentence $\phi \in \Sigma_n$, there exist two formulas $\phi_l(x)$, $\phi_r(x) \in \Sigma_n$ in which $x$ is the unique free variable and such that for every $k$, for every word $w = a_1 \cdots a_k$ of length $k$ and for every integer $s$ such that $1 \leqslant s \leqslant k$, we have*

(1) $w \in L(\phi_l(s)) \Leftrightarrow a_1 \cdots a_{s-1} \in L(\phi)$

(2) $w \in L(\phi_r(s)) \Leftrightarrow a_{s+1} \cdots a_k \in L(\phi)$.

*Proof.* We define $\phi_l$ and $\phi_r$ for every formula $\phi$. We treat the case of $\phi_l$ (the other case in dual). $\phi_l$ is constructed by induction as follows:
If $\phi$ is quantifier-free, then $\phi_l = \phi$. Otherwise, we set

$$(\exists z\, \phi)_l = \exists z((z < x) \wedge \phi_l)$$

$$(\forall z\, \phi)_l = \forall z((z < x) \rightarrow \phi_l)$$

$$(\phi \vee \psi)_l = \phi_l \vee \psi_l \qquad (\neg\phi)_l = \neg\phi_l \qquad (\phi \wedge \psi)_l = \phi_l \wedge \psi_l.$$

Then one can verify by induction on $n \geqslant 0$ that if $\varphi \in \Sigma_n$, then $\varphi_l \in \Sigma_n$. ∎

We are now ready to prove the property announced at the beginning of this section. If $n = 0$, then $\mathscr{A}_0(A^*) = \mathscr{B}_0(A^*) = \{\varnothing, A^*\}$. But we have $L(0) = \varnothing$ and $L(1) = A^*$, proving the result for $n = 0$.
Assume that the property is true for some $n \geqslant 0$. Denote by $C$ the class of all subsets $X$ of $A^*$ such that there exists a sentence $\phi \in \Sigma_{n+1}$ with $X = L(\phi)$. Then

$\mathscr{A}_n(A^*) \subset C$ by induction, since $\Gamma_n \subset \Sigma_{n+1}$. Next if $X, Y \in C$, there exist sentences $\phi, \psi \in \Sigma_{n+1}$ such that $X = L(\phi)$ and $Y = L(\psi)$. Now for each $a \in A$, the language $XaY$ is defined by the following sentence, which uses the previously defined formulas $\phi_r$ and $\phi_l$,

$$\exists x(\phi_l(x) \wedge \pi_a(x) \wedge \psi_r(x)).$$

Since this sentence is in $\Sigma_{n+1}$, we have $XaY \subset C$. Furthermore, we have for any sentences $\varphi$, $\psi$

$$L(\varphi \vee \psi) = L(\varphi) \cup L(\psi),\ L(\varphi \wedge \psi) = L(\varphi) \cap L(\psi),\ L(\neg \varphi) = A^* - L(\varphi).$$

Hence $C$ is closed under union and intersection. Thus $\mathscr{B}_{n+1}(A^*) \subset C$. Finally any element of $\mathscr{A}_{n+1}(A^*)$ is a boolean combination of languages in $\mathscr{B}_{n+1}(A^*)$ and therefore definable by a $\Gamma_{n+1}$-sentence. ∎

## 3. From Formulas to Languages

In this section we shall prove that for every sentence $\phi \in \Sigma_n$ (resp. $\Gamma_n$) and for every $n \geqslant 0$, $L(\phi) \in \mathscr{A}_n(A^*)$ (resp. $\mathscr{B}_n(A^*)$). We first need to define $L(\phi)$ for *every* formula and not only for a sentence.

Let $V$ be a finite set of variables of $\mathscr{L}$. We denote by $A^*_V$ the set of all pairs

$$m = (u, \sigma)$$

where $u \in A^*$ and $\sigma$ is a mapping from $V$ into $[|u|]$.

Therefore $A^*_V$ is a set of "marked words." Each variable of $V$ marks a position in the word. For $m = (u, \sigma) \in A^*_V$ and $n = (v, \tau) \in A^*_W$ such that $V \cap W = \varnothing$. We set

$$mn = (uv, \rho)$$

where $\rho$ is the application from $V \cup W$ into $[|uv|]$ defined by

$$\rho(z) = \begin{cases} \sigma(z) & \text{if } z \in V, \\ \tau(z) + |u| & \text{if } z \in W. \end{cases}$$

The set $A^*_\varnothing$ can be identified to $A^*$ without confusion. Observe, for further use, that every marked word $m \in A^*_V$ admits a unique decomposition of the form

$$(a_1, \sigma_1)(a_2, \sigma_2) \cdots (a_n, \sigma_n)$$

where $a_1, \ldots, a_n$ are letters of $A$. Let in fact $m = (u, \sigma)$, $u = a_1 a_2 \cdots a_n$ and $V_i = \sigma^{-1}(i)$. Then $\sigma_i$ maps $V_i$ to $[1]$.

We shall also denote by $A_V$ the set of all marked words $(a, \sigma) \in A_V^*$ such that $a \in A$. Finally for every set of variables $V$, we define a hierarchy of subsets of $A_V^*$ as follows

$$\mathscr{A}_0(A_V^*) = \mathscr{B}_0(A_V^*) = \{\varnothing, A_V^*\}.$$

Then, for $n \geqslant 0$, $\mathscr{B}_{n+1}(A_V^*)$ denotes the smallest class $C_V$ of subsets of $A_V^*$ such that

(i) $\mathscr{A}_n(A_V^*) \subset C_V$.

(ii) $C_V$ is closed under finite unions and intersections.

(iii) For every $X \in C_{V'}$, $Y \in C_{V''}$, and $(a, \sigma) \in A_W$ with $a \in A$ such that $V'$, $V''$, and $W$ are pairwise disjoint and $V$ is their union, we have

$$X(a, \sigma) Y \in C_V.$$

Finally $\mathscr{A}_{n+1}(A_V^*)$ denotes the boolean algebra generated by $\mathscr{B}_{n+1}(A_V^*)$.

We first establish some elementary properties of this hierarchy. First we have the obvious

LEMMA 3.1. *For every $n \geqslant 0$, $\mathscr{A}_n(A_\varnothing^*) = \mathscr{A}_n(A^*)$, and $\mathscr{B}_n(A_\varnothing^*) = \mathscr{B}_n(A^*)$.*

In the sequel we shall need the following notation, that extends to $A_V^*$ the usual notation for residuals. Let $W$ and $W'$ be two disjoint subsets of $V$ such that $W \cup W' = V$. For $m \in A_W^*$ and $X \subset A_V^*$, we set

$$m^{-1}X = \{m' \in A_{W'}^* \mid mm' \in X\}$$

$$Xm^{-1} = \{m' \in A_{W'}^* \mid m'm \in X\}.$$

The following lemma shows that the sets $\mathscr{B}_n(A_V^*)$ and $\mathscr{A}_n(A_V^*)$ are, in some sense, closed under taking residuals.

LEMMA 3.2. *Let $X \in \mathscr{B}_n(A_V^*)$ (resp. $\mathscr{A}_n(A_V^*)$) and let $V = W \cup W'$ with $W \cap W' = \varnothing$. The set of sets of the form $m^{-1}X$ or $Xm^{-1}$ (where $m \in A_W^*$) is a finite subset of $\mathscr{B}_n(A_{W'}^*)$ (resp. $\mathscr{A}_n(A_{W'}^*)$).*

*Proof.* By symmetry, it suffices to prove the property for the sets $m^{-1}X$. The property is clear for $n = 0$. Suppose now that the property is true for $\mathscr{B}_n(A_V^*)$ for some $n \geqslant 0$. That is, suppose that for every $X \in \mathscr{B}_n(A_V^*)$ the set

$$E_W(X) = \{m^{-1}X \mid m \in A_W^*\}$$

is a finite subset of $\mathscr{B}_n(A_{W'}^*)$. Now if $X \in \mathscr{A}_n(A_V^*)$, $X$ is, by definition, a boolean combination of a finite family $(X_i)_{1 \leqslant i \leqslant k}$ of elements of $\mathscr{B}_n(A_V^*)$. Since the operation $X \to m^{-1}X$ commutes with boolean operations, the set $E_W(X)$ is contained in the boolean algebra generated by the union $\bigcup_{1 \leqslant i \leqslant k} E_W(X_i)$. It follows that $E_W(X)$ is a finite set contained in $\mathscr{A}_n(A_{W'}^*)$ and this proves the property for $\mathscr{A}_n(A_V^*)$.

To conclude the induction we have to show that if the property holds for $\mathscr{A}_n(A_V^*)$, then it also holds for $\mathscr{B}_{n+1}(A_V^*)$. Let $Y \in \mathscr{B}_{n+1}(A_V^*)$. Then $Y$ is a boolean combination of sets $Y_i$, where each $Y_i$ has the form

$$X_0(a_1, \sigma_1) X_1 \cdots (a_k, \sigma_k) X_k \tag{$*$}$$

with $X_0 \in \mathscr{A}_n(A_{V_0}^*)$, $X_1 \in \mathscr{A}_n(A_{V_1}^*)$,..., $X_k \in \mathscr{A}_n(A_{V_k}^*)$, $(a_1, \sigma_1) \in A_{W_1}$,..., $(a_k, \sigma_k) \in A_{W_k}$, and where $V$ is the disjoint union of the sets $V_0, V_1,..., V_k, W_1,..., W_k$.

It suffices to show that for each $i$, the set $E_W(Y_i)$ is a finite subset of $\mathscr{B}_{n+1}(A_{W'}^*)$. Therefore we may assume that $Y$ is a set of the form $(*)$ above.

Denote by $F_W$ the set of all finite unions of subsets of the form

$$X_i'(a_{i+1}, \sigma_{i+1}) X_{i+1} \cdots (a_k, \sigma_k) X_k \tag{$**$}$$

with $0 \leqslant i \leqslant k$, $X_i' \in E_{W_i'}(X_i)$ and where $W$ is the disjoint union of the sets $V_0,..., V_{i-1}, W_1,..., W_{i-1}$, and $W_i'$.

By the induction hypothesis, each of the sets $X_i'$ is in $\mathscr{A}_n(A_{V_i \setminus W_i'}^*)$. Hence $F_W$ is a finite subset of $\mathscr{B}_{n+1}(A_{W'}^*)$. We claim that for every set $W$, $E_W(Y) \subset F_W$, and this claim will prove the proposition. To prove the claim, we show by induction on $t$ that if $m = (a_1, \sigma_1) \cdots (a_t, \sigma_t) \in A_W^*$ then $m^{-1}Y \in F_W$. If $t = 0$, then $m = 1$, $W = \varnothing$, and $m^{-1}Y = Y \in F_\varnothing$. Suppose that the result holds for all $m$ of length $\leqslant t$ and let $(a, \sigma) \in A_T$. Then we have

$$[m(a, \sigma)]^{-1}\, Y = (a, \sigma)^{-1}(m^{-1}Y)$$

and by induction hypothesis, $m^{-1}Y \in F_W$. Therefore $m^{-1}Y$ is a finite union of subsets of the form $(**)$. Now we have with the previous notations

$$\begin{aligned}(a, \sigma)^{-1}&[X_i'(a_{i+1}, \sigma_{i+1}) X_{i+1} \cdots (a_k, \sigma_k) X_k] \\ &= ((a, \sigma)^{-1} X_i')(a_{i+1}, \sigma_{i+1}) X_{i+1} \cdots X_k \cup R\end{aligned}$$

with

$$R = \begin{cases} X_{i+1}(a_{i+2}, \sigma_{i+2}) \cdots (a_k, \sigma_k) X_k & \text{if } W_i' = V_i \text{ (and hence } X_i' \subset A_\varnothing^*) \\ & \quad\text{if } 1 \in X_i' \text{ and if} \\ & \quad (a, \sigma) = (a_{i+1}, \sigma_{i+1}), \\ \varnothing & \text{otherwise.} \end{cases}$$

It follows that $(a, \sigma)^{-1}(m^{-1}y) \in F_{W \cup T}$ and this concludes the induction. ▌

Let us come back to formulas. Let $\phi$ be a formula of $L$ and let $V$ be a set of variables containing all free variables of $\phi$. We denote by $L_V(\phi)$ the set of all $(u, \sigma) \in A_V^*$ such that $\phi$ is true when one substitutes to each free variable $x$ of $\phi$ the value $\sigma(x)$ and when the sentence obtained in this way is interpreted as usual in $[|u|]$.

The following result concludes the proof of Theorem 1.1.

PROPOSITION 3.3. *Let $\phi$ be a formula of $L$ and let $V$ be a set of variables containing all free variables of $\phi$. Then we have for every $n \geqslant 1$:*

(1) *If $\phi \in \Sigma_n$ then $L_V(\phi) \in \mathscr{B}_n(A_V^*)$.*

(2) *If $\phi \in \Gamma_n$ then $L_V(\phi) \in \mathscr{A}_n(A_V^*)$.*

We first need two lemmas to prove this proposition.

LEMMA 3.4. *Let $\phi \in \Gamma_0$ and let $V$ be a set of variables containing the free variables of $\phi$. Then $L_V(\phi) \in \mathscr{B}_1(A_V^*)$.*

*Proof.* Every formula $\phi \in \Gamma_0$ is equivalent to a conjunction–disjunction of formulas of the form

$$\pi_a z \qquad \text{or} \qquad y < z$$

(for $y, z \in V$ and $a \in A$) or their negation. For instance $y = z$ is equivalent to $(\neg(y<z)) \wedge (\neg(z<y))$. But

$$L_V(\pi_a z) = \bigcup A_{V'}^*(a, \sigma)\, A_{V''}^*$$

where $(a, \sigma) \in A_W$ and the union runs over all triples $V'$, $V''$, $W$ such that $z \in W$ and $V$ is the disjoint union of $V'$, $V''$, and $W$. Similarly

$$L_V(\neg \pi_a z) = \bigcup A_{V'}^*(b, \sigma)\, A_{V''}^*$$

where $(b, \sigma) \in A_W$, and where the union runs over all triples $(b, \sigma)$, $V'$, $V''$ such that $b \neq a$ and such that $V$ is the disjoint union of $V'$, $V''$, and $W$. We also have

$$L_V(y<z) = \bigcup A_{V_0}^*(a, \sigma)\, A_{V_1}^*(b, \tau)\, A_{V_2}^*$$

with $(a, \sigma) \in A_{W_1}$, $(b, \tau) \in A_{W_2}$, and where the union runs over all quintuples $(a, \sigma)$, $(b, \tau)$, $V_0$, $V_1$, $V_2$ such that $y \in W_1$, $z \in W_2$, and $V$ is the disjoint union of $V_0$, $W_1$, $V_1$, $W_2$, $V_2$. Finally,

$$L_V(\neg(y<z)) = \left[\bigcup A_{V'}^*(a, \sigma)\, A_{V''}^*\right] \cup L_V(z<y)$$

with $(a, \sigma) \in A_W^*$, $a \in A$ and where the union runs over all triples $(a, \sigma)$, $V'$, $V''$ such that $y, z \in W$ and such that $V$ is the disjoint union of $V'$, $V''$, and $W$.

Since $L_V(\phi \vee \psi) = L_V(\phi) \cup L_V(\psi)$ and $L_V(\phi \wedge \psi) = L_V(\phi) \cap L_V(\psi)$ and since $\mathscr{B}_1(A_V^*)$ is closed under union and intersection, this concludes the proof of the lemma. ∎

LEMMA 3.5. *Let $\phi$ be a formula and let $n \geqslant 1$ be an integer such that $L_V(\phi) \in \mathscr{B}_n(A_V^*)$, where $V$ contains all free variables of $\phi$. Then for every free variable $z$ of $\phi$, $L_{V\setminus\{z\}}(\exists z\, \phi) \in \mathscr{B}_n(A_{V\setminus\{z\}}^*)$.*

*Proof.* Let $X = L_V(\phi)$. Then we have

$$X = \bigcup Y(u)(a, \sigma)\, Z(u)$$

where the union runs over all $u \in A_{V'}^*$ and all $(a, \sigma) \in A_W$ such that $V' \subset V\setminus\{z\}$, $\{z\} \subset W \subset V\setminus V'$ and with

$$Z(u) = [u(a, \sigma)]^{-1}X,\ Y(u) = \bigcap_{v \in Z} X[(a, \sigma)v]^{-1}.$$

By Lemma 3.2 the union above is finite and each term $Y(a, \sigma)\, Z$ satisfies $Y \in \mathscr{B}_n(A_{V'}^*)$, $Z \in \mathscr{B}_n(A_{V''}^*)$. We then have

$$L_{V\setminus\{z\}}(\exists z\, \phi) = \bigcup Y(a, \sigma')\, Z$$

where $\sigma'$ is the restriction of $\sigma$ to the set $W\setminus\{z\}$ and this proves the lemma. ∎

We are now ready to complete the proof of Proposition 3.3. For every $n \geqslant 1$, set $F_n = \{\phi \mid \text{for every } V \text{ containing all free variables of } \phi,\ L_V(\phi) \in \mathscr{B}_n(A_V^*)\}$. We shall prove by induction on $n \geqslant 1$ that

$$\Sigma_n \subset F_n.$$

For $n = 1$, we have $\Gamma_0 \subset F_1$ by Lemma 3.4. Next if $\phi \in F_1$, we have, by Lemma 3.5, $L_{V\setminus\{z\}}(\exists z\, \phi) \in \mathscr{B}_1(A_{V\setminus\{z\}}^*)$ and thus $\exists z\, \phi \in F_1$. Thus $\Sigma_1 \subset F_1$.

Assume now that $\Sigma_n \subset F_n$ for some $n \geqslant 1$. Then for every formula $\phi \in \Gamma_n$, we have $L_V(\phi) \in \mathscr{A}_n(A_V^*)$ for every set $V$ containing all free variables of $\phi$. Since $\mathscr{A}_n(A_V^*)$ is contained in $\mathscr{B}_{n+1}(A_V^*)$, we have $\Gamma_n \subset F_{n+1}$. Finally if $\phi \in F_{n+1}$, then by Lemma 3.5

$$L_{V\setminus\{z\}}(\exists z\, \phi) \in \mathscr{B}_{n+1}(A_{V\setminus\{z\}}^*)$$

for every free variable $z$ of $\phi$ and hence $(\exists z\, \phi) \in F_{n+1}$. Consequently $\Sigma_{n+1} \subset F_{n+1}$ and Proposition 3.3 follows immediately. ∎

## 4. First-Order Logic on Infinite Words

In this section we shall extend the previous results to the case where the set $A^*$ is replaced by the set $A^\omega$ of infinite words on the alphabet $A$. We shall see that, up to natural modifications in the definitions, the same results hold.

An infinite word is a sequence $\mathbb{N} \to A$ usually denoted $a_0 a_1 \cdots a_n \cdots$. Just as before, to each (first-order) sentence of $\mathscr{L}$ is associated the set $L(\phi)$ of all words $w$ of $A^\omega$ such that $\phi$ is true when variables are interpreted as integers, the $<$ relation is the usual relation on integers and the formula $\pi_a x$ is interpreted as "the letter of index $x$ in $w$ is an $a$."

Nevertheless, the two following examples will emphasize the difference between finite and infinite words. The formula

$$\phi = \exists x \, \forall y (\pi_a x \wedge \neg (x < y))$$

defines the language of all finite words whose last letter is an $a$. However, no infinite words satisfy this formula.

In contrast, consider the formula

$$\psi = \forall x \, \exists y ((x < y) \wedge \pi_b \, y).$$

No finite word satisfies this formula, but $\psi$ defines the set of infinite words $(a^*b)^\omega$. We define by induction two sequences $\mathscr{A}_n(A^\omega)$ and $\mathscr{B}_n(A^\omega)$ in the following way

$$\mathscr{A}_0(A^\omega) = \mathscr{B}_0(A^\omega) = \{\varnothing, A^\omega\}.$$

Next, for $n \geqslant 1$, we denote by $\mathscr{B}_{n+1}(A^\omega)$ the set of all finite unions and intersections of sets of the form $XaY$ where $X \in \mathscr{B}_{n+1}(A^*)$, $a \in A$ and $Y \in \mathscr{A}_n(A^\omega)$. Finally $\mathscr{A}_{n+1}(A^\omega)$ denotes the boolean closure of $\mathscr{B}_{n+1}(A^\omega)$. Then we can state

**Theorem 4.1.** *For every $n \geqslant 0$ and for every $X \subset A^\omega$, one has $X \in \mathscr{A}_{n+1}(A^\omega)$ (resp. $\mathscr{B}_n(A^\omega)$) iff there exists a sentence $\phi \in \Gamma_n$ (resp. $\Gamma_n$) such that $X = L(\phi)$.*

Notice that the set $\mathscr{A}(A^\omega) = \bigcup_{n \geqslant 0} \mathscr{A}_n(A^\omega)$ can be directly defined as the smallest set of subsets of $A^\omega$ containing the empty set and closed under boolean operations and left concatenation with a language of $\mathscr{A}(A^*)$.

**Corollary 4.2** [8]. *One has $X \in \mathscr{A}(A^\omega)$ iff there exists a sentence $\phi$ of $\mathscr{L}$ such that $X = L(\phi)$.*

The proof of Theorem 4.1 requires some modifications with respect to the proof of Theorem 1.1. To pass from languages to formulas, that is, to show that for every $X \in \mathscr{B}_n(A^\omega)$ (resp. $\mathscr{A}_n(A^\omega)$, there exists a formula $\phi \in \Sigma_n$ (resp. $\Gamma_n$) such that $X = L(\phi)$, we need to modify the property of $\phi_r$ (Proposition 2.1) as follows:

for every infinite word $w = a_0 a_1 \cdots$ and for every $s \geq 0$, $w \in L(\phi_r(s))$ iff $a_{s+1} a_{s+2} \cdots \in L(\phi)$.

Next we can mimic the proof given in Section 2 with the formula

$$XaY = L[\exists x(\phi_l(x) \wedge \pi_a x \wedge \psi_r(x))]$$

where $X$ (resp. $Y$) is the set of finite (resp. infinite) words defined by $\phi$ (resp. $\psi$).

To pass from formulas to infinite words, that is, to prove that for every sentence $\phi \in \Sigma_n$ (resp. $\Gamma_n$), $L(\phi) \in \mathscr{B}_n(A^\omega)$ (resp. $\mathscr{A}_n(A^\omega)$, we need first to introduce the set $A_V^\omega$ of all pairs $(w, \sigma)$, where $w$ is an infinite word of $A^\omega$ and $\sigma$ is a mapping from $V$ into $\mathbb{N}$. If $m = (u, \sigma) \in A_V^*$ and $n = (v, \tau) \in A_W^\omega$ with $V \cap W \neq \varnothing$, we set

$$mn = (uv, \rho) \in A_{V \cup W}^\omega$$

where $\rho$ is the mapping from $V \cup W$ into $\mathbb{N}$ defined by

$$\rho(x) = \begin{cases} \sigma(x) & \text{if} \quad x \in V, \\ \tau(x) + |u| & \text{if} \quad x \in W. \end{cases}$$

If $\phi$ is a formula of $L$ and if $V$ is a set of variables containing all free variables of $\phi$, we define again $L_V(\phi)$ as the set of all $(u, \sigma) \in A_V^\omega$ such that, if one substitutes to each free variable $x$ of $\phi$ the value $\sigma(x)$, the sentence obtained in this way is satisfied by $u$.

Next we define the sequences $(\mathscr{A}_n(A_V^\omega))_{n \geq 0}$ and $(\mathscr{B}_n(A_V^\omega))_{n \geq 0}$ by setting

(1) $\mathscr{B}_0(A_V^\omega) = \mathscr{A}_0(A_V^\omega) = \{\varnothing, A_V^\omega\}$.

(2) For $n \geq 0$, $\mathscr{B}_{n+1}(A_V^\omega)$ is the set of all finite unions and intersections of sets of the form $X'(a, \sigma) X''$.

with $X' \in \mathscr{B}_{n+1}(A_{V'}^*)$, $(a, \sigma) \in A_W$, $X'' \in \mathscr{A}_n(A_{V''}^\omega)$, and where $V$ is the disjoint union of $V'$, $W$, and $V''$.

(3) For $n \geq 0$, $\mathscr{A}_{n+1}(A_V^\omega)$ is the boolean closure of $\mathscr{B}_{n+1}(A_V^\omega)$.

Observe that, for every $n \geq 0$, $\mathscr{A}_n(A_\varnothing^\omega) = \mathscr{A}_n(A^\omega)$ (resp. $\mathscr{B}_n(A_\varnothing^\omega) = \mathscr{B}_n(A^\omega)$).

Let $W$ and $W'$ be two disjoint subsets of $V$ such that $W \cup W' = V$ and let $X \subset A_V^\omega$. If $m \in A_W^*$, we set

$$m^{-1}X = \{m' \in A_{W'}^\omega \mid mm' \in X\}$$

and if $m \in A_{W'}^\omega$, we set

$$Xm^{-1} = \{m' \in A_{W'}^* \mid m'm \in X\}.$$

Lemma 3.2 can now be modified as follows.

LEMMA 4.3. *Let* $X \in \mathscr{B}_n(A_V^\omega)$ *(resp.* $\mathscr{A}_n(A_V^\omega)$*) and let* $V = W \cup W'$ *with* $W \cap W' = \varnothing$. *Then*

(1) *The set of all subsets of the form* $m^{-1}X$ *(where* $m \in A_W^*$*) is a finite subset of* $\mathscr{B}_n(A_{W'}^\omega)$ *(resp.* $\mathscr{A}_n(A_{W'}^\omega)$.

(2) *The set of all subsets of the form* $Xm^{-1}$ *(where* $m \in A_W^\omega$*) is a finite subset of* $\mathscr{B}_n(A_{W'}^*)$ *(resp.* $\mathscr{A}_n(A_{W'}^*)$.

*Proof.* For the first assertion it suffices to mimic the proof of Lemma 3.2. For the second assertion, one argues by induction on $n$ just like in Lemma 3.2, but a difficulty arises. Indeed if we assume the result for $X \in \mathscr{A}_n(A_V^*)$, one cannot prove the property for $\mathscr{B}_{n+1}(A_V^*)$ simply by induction on the length of $m$ as we did for finite words.

Thus, let $X \in \mathscr{B}_{n+1}(A_V^*)$. Then $X$ is a finite boolean combination of sets

$$X'(a, \sigma)\, X''$$

with $X' \in \mathscr{B}_{n+1}(A_{V'}^*)$, $(a, \sigma) \in A_U$, $X'' \in \mathscr{A}_n(A_{V''}^\omega)$, and where $V$ is the disjoint union of $V'$, $U$, and $V''$.

By an argument already used in the proof of Lemma 3.2, we may assume that $X = X'(a, \sigma)\, X''$. Now, we have

$$Xm^{-1} = X'(a, \sigma)(X''m^{-1}) \cup R$$

with $R = \bigcup X'n^{-1}$, where this last union runs over all $n \in A_{V''}^\omega$ such that there exists a factorization $m = n(a, \sigma)\, n'$ with $n' \in X''$ and $W$ is the disjoint union of $W'''$, $U$, and $V''$.

By induction, the set of all $X''m^{-1}$ is a finite subset of $\mathscr{A}_n(A_{V''\setminus W}^\omega)$. Furthermore, by Lemma 3.2, the set of all $X'n^{-1}$ is a finite subset of $\mathscr{B}_n(A_{V\setminus W}^*)$. It follows that $X$ satisfies property (2) of the statement. The rest of the proof is easily adapted from the proof of Lemma 3.2. ∎

The last part of the proof of Theorem 4.1 is the same as the proof given in Section 3. We first show, as in Lemma 3.4, that for every formula $\phi \in \Gamma_0$,

$$L_V(\phi) \in \mathscr{B}_1(A_V^\omega).$$

Next we show, as in Lemma 3.5, that if $L_V(\phi) \in \mathscr{B}_n(A_V^\omega)$ then for every free variable $z$ of $\phi$,

$$L_{V\setminus\{z\}}(\exists z\, \phi) \in \mathscr{B}_n(A_{V\setminus\{z\}}^*)$$

by using this time Lemma 4.3. Finally we deduce, as in Proposition 3.3 that if $\phi \in \Sigma_n$ then $L_V(\phi) \in \mathscr{B}_n(A_V^\omega)$ and if $\phi \in \Gamma_n$ then $L_V(\phi) \in \mathscr{A}_n(A_V^\omega)$. This concludes the proof of Theorem 4.1. ∎

Let us mention a last extension of Theorem 1.1. We consider the set $A^{\mathbb{Z}}$ of all applications from $\mathbb{Z}$ into $A$, also called "biinfinite words," in reference to the notation $u = \cdots a_{-1}a_0a_1\cdots$ (where $a_i \in A$ for all $i \in \mathbb{Z}$) used to represent these words. Then one can generalize the previous results as follows: To each sentence $\phi$ of $\mathscr{L}$ is associated the set $L(\phi)$ of all biinfinite words $u$ such that $\phi$ is true on $u$. Observe that since $\mathscr{L}$ does not contain any constant symbol, every subset of $A^{\mathbb{Z}}$ of the form $L(\phi)$ is shift-invariant.

Denote by ${}^{\omega}A$ the set of all "left-infinite" words $u = \cdots a_{-1}a_0$. To each sentence $\phi$ of $\mathscr{L}$, one associates in the same way the set of all left-infinite words $u \in {}^{\omega}A$ such that $\phi$ is true on $u$ (by interpreting $\phi$ in the set of all negative or nul integers). We define the hierarchies $\mathscr{A}_n({}^{\omega}A)$ and $\mathscr{B}_n({}^{\omega}A)$ $(n \geqslant 0)$ analogous to $\mathscr{A}_n(A^{\omega})$ and $\mathscr{B}_n(A^{\omega})$.

For $X \subset {}^{\omega}A$ and $Y \subset A^{\omega}$, we denote by $XY$ the set of all biinfinite words $u = \cdots a_{-1}a_0a_1\cdots$ such that there exists an integer $i \in \mathbb{Z}$ with

$$\cdots a_{i-1}a_i \in X, \qquad a_{i+1}a_{i+2}\cdots \in Y.$$

Next we define

$$\mathscr{A}_0(A^{\mathbb{Z}}) = \mathscr{B}_0(A^{\mathbb{Z}}) = \{\emptyset, A^{\mathbb{Z}}\}$$

and for $n \geqslant 0$, we define $\mathscr{B}_{n+1}(A^{\mathbb{Z}})$ as the set of finite unions and intersections of sets

$$XaY$$

with $X \in \mathscr{B}_n({}^{\omega}A)$, $a \in A$, and $Y \in \mathscr{B}_n(A^{\omega})$. Finally, we denote by $\mathscr{A}_{n+1}(A^{\mathbb{Z}})$ the boolean closure of $\mathscr{B}_{n+1}(A^{\mathbb{Z}})$. We can now state

THEOREM 4.4. *For every $n \geqslant 0$ and for all $X \subset A^{\mathbb{Z}}$, one has $X \in \mathscr{A}_n(A^{\mathbb{Z}})$ (resp. $\mathscr{B}_n(A^{\mathbb{Z}})$). Iff there exists a formula $\phi \in \Gamma_n$ (resp. $\Sigma_n$) such that $X = L(\phi)$.*

If we set $\mathscr{A}(A^{\mathbb{Z}}) = \bigcup \mathscr{A}_n(A^{\mathbb{Z}})$, we have, just as in the previous cases, the following corollary.

COROLLARY 4.5. *For every $X \subset A^{\mathbb{Z}}$ one has $X \in \mathscr{A}(A^{\mathbb{Z}})$ iff there exists a formula $\phi$ of $\mathscr{L}$ such that $X = L(\phi)$.*

There is no particular problem to show that for every $X \in \mathscr{A}_n(A^{\mathbb{Z}})$, there exists a formula $\phi \in \Gamma_n$ such that $X = L(\phi)$. Indeed let $Y$ (resp. $Z$) be the set of all left- (right-) infinite words satisfying a sentence $\psi$ (resp. $\tau$) of $\Sigma_n$. Then

$$XaY = L(\exists z(\psi_l(z) \wedge \pi_a(z) \wedge \tau_r(z))).$$

Conversely, to show that $\phi \in \Gamma_n$ implies $L(\phi) \in \mathscr{A}_n(A^{\mathbb{Z}})$, one needs of course to introduce the set $A_V^{\mathbb{Z}}$ and the families $(\mathscr{A}_n(A_V^{\mathbb{Z}}))_{n \geqslant 0}$ and $(\mathscr{B}_n(A_V^{\mathbb{Z}}))_{n \geqslant 0}$ whose definition mimics the previous definitions. The remainder of the proof is the same. ∎

## ACKNOWLEDGMENTS

The authors wish to thank M. Takahashi who has pointed out several incorrect points on a first version of this paper.

## REFERENCES

1. J. BRZOZOWSKI AND R. KNAST, The dot-depth hierarchy of star-free languages is infinite, *J. Comput. System Sci.* **16** (1978), 37–55.
2. J. R. BÜCHI, On a decision method in restricted second-order arithmetic, *in* "Logic Methodology and Philosophy of Science" (Proc. 1960 Internat. Congr.), pp. 1–11, Stanford Univ. Press, Stanford, Calif., 1960.
3. R. McNAUGHTON AND S. PAPERT, "Counter Free Automata," MIT Press, Cambridge, Mass., 1971.
4. M. NIVAT AND D. PERRIN, Ensembles reconnaissables de mots biinfinis, *Canad. J. Math.*, to appear.
5. D. PERRIN, Recent results on automata and infinite words, *in* "Mathematical Foundations of Computer Science," Lecture Notes in Comput. Sci. Vol. 176, pp. 134–148, Springer-Verlag, New York/Berlin, 1984.
6. J.-E. PIN, "Variétés de langages formels," Masson, Paris, 1984.
7. H. STRAUBING, Finite semigroup varieties of the form $V * D$, *J. Pure Appl. Algebra* **36** (1985), 53–94.
8. W. THOMAS, Star free regular sets of $\omega$-sequences, *Inform. Control* **42** (1979), 148–156.
9. W. THOMAS, Classifying regular events in symbolic logic, *J. Comput. System Sci.* **25** (1982), 360–376.