# A CHARACTERIZATION OF SEMILINEAR SETS

RANI SIROMONEY

**Introduction.** The recent interest in the structure of programming languages has led to the study of their mathematical properties. Characterizations of bounded context-free languages (also called bounded ALGOL-like languages) [1] and bounded regular sets [3] have been given in terms of certain semilinear subsets of $N^n$. Semilinear sets have been extensively studied as subsets of lattice points in $n$-space which are finite unions of cosets of finitely generated subsemigroups of the set of all lattice points with nonnegative coordinates and which are also shown to be equivalent to the family of sets defined by modified Presburger formulas [2]. In this note we give a characterization and discuss decision procedures for semilinear sets of words (hereafter called semilinear sets) [4] which include bounded context-free languages and hence bounded regular sets.

1. **Preliminaries.** Let $\Sigma$ be a finite nonempty set and $\Sigma^*$ the free semigroup with identity $\epsilon$ generated by $\Sigma$. A subset $X$ of $\Sigma^*$ is said to be *bounded* if there exist words $w_1, \cdots, w_k$ in $\Sigma^*$ such that $X \subseteq w_1^* \cdots w_k^*$. For each $k$-tuple of words $w = \langle w_1, \cdots, w_k \rangle$ let $f_w$ denote the function defined on $N^k$ by $f_w(p) = w_1^{p(1)} \cdots w_k^{p(k)}$ where $p = (p(1), \cdots, p(k))$ is in $N^k$. Then $M \subseteq w_1^* \cdots w_k^*$ is said to be *semilinear in w* if $w = \langle w_1, \cdots, w_k \rangle$ and $f_w^{-1}(M)$ is a semilinear subset of $N^k$. A set $M$ is called *semilinear* if it is semilinear in some $k$-tuple $\langle w_1, \cdots, w_k \rangle$ [4].

An *equal matrix grammar* (abbreviated EMG) *of order k* [5] is a 4-tuple $G = (V, \Sigma, P, S)$ where (i) $V$ consists of the alphabet $\Sigma$, the initial symbol $S$, and the rest of the nonterminals $V_N$ in the form of ordered $k$-tuples $\langle A_1, \cdots, A_k \rangle$ where the $k$-tuples are distinct, $k$ being finite. In other words if $\langle A_1, \cdots, A_k \rangle$ and $\langle B_1, \cdots, B_k \rangle$ are any two $k$-tuples, $A_1, \cdots, A_k, B_1, \cdots, B_k$ are distinct. (ii) $P$ consists of the following types of matrix rules:

(a) A set of *initial matrix rules* (abbreviated initial rules) of the form $[S \to f_1 A_1 \cdots f_k A_k]$ where $f_1, \cdots, f_k$ are in $\Sigma^*$, $S$ the initial symbol and $\langle A_1, \cdots, A_k \rangle$ in $V_N$. (Note that $S \to f_1 A_1 \cdots f_k A_k$ is a context-free rule.)

(b) A set of *nonterminal equal matrix rules* (abbreviated nonterminal rules) of the form

Received by the editors August 12, 1968.

$$\begin{bmatrix} A_1 \to f_1A_1 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ A_k \to f_kA_k \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} A_1 \to f_1B_1 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ A_k \to f_kB_k \end{bmatrix}$$

where $f_1, \cdots, f_k$ are in $\Sigma^*$ and $\langle A_1, \cdots, A_k\rangle$, $\langle B_1, \cdots, B_k\rangle$ in $V_N$.

(c) A set of *terminal equal matrix rules* (abbreviated terminal rules) of the form

$$\begin{bmatrix} A_1 \to f_1 \\ \cdot\ \cdot\ \cdot\ \cdot \\ A_k \to f_k \end{bmatrix}$$

where $f_1, \cdots, f_k$ are in $\Sigma^*$, $\langle A_1, \cdots, A_k\rangle$ in $V_N$. An *equal matrix grammar* is an EMG of any finite order.

NOTATION. Let $G=(V, \Sigma, P, S)$ be an EMG. We write $S\Rightarrow f_1A_1 \cdots f_kA_k$ if $[S\to f_1A_1 \cdots f_kA_k]$ is an initial rule in $P$, and $w_1\Rightarrow w_2$ if $w_1=x_1A_1 \cdots x_kA_k$, $w_2=x_1v_1 \cdots x_kv_k$, $x_i$ in $\Sigma^*$, $\langle A_1, \cdots, A_k\rangle$ in $V_N$ and

$$\begin{bmatrix} A_1 \to v_1 \\ \cdot\ \cdot\ \cdot\ \cdot \\ A_k \to v_k \end{bmatrix}$$

is in $P$. We write $w\overset{*}{\Rightarrow}y$ if either $w=y$ or there exist $w_0=w, w_1, \cdots, w_n=y$ such that $w_i\Rightarrow w_{i+1}$ for each $i$. A sequence of words $w_0, \cdots, w_n$ such that $w_i\Rightarrow w_{i+1}$ for each $i$, is called a *derivation* or *generation of* $w_n$ (from $w_0$) and is denoted by $w_0\Rightarrow \cdots \Rightarrow w_n$. $L\subseteq\Sigma^*$ is an *equal matrix language* (abbreviated EML) if there is an EMG $G=(V, \Sigma, P, S)$ such that $L=L(G)$ where $L(G)=\{w \text{ in } \Sigma^*/S\overset{*}{\Rightarrow}w\}$. $L(G)$ is said to be the language *generated* by $G$.

## 2. Characterization.

We now present a characterization of semilinear sets, which is related to Theorem 2.1 of [1] and Theorem 1.3 of [3].

THEOREM 2.1. *$X\subseteq\Sigma^*$ is semilinear if and only if $X$ is a bounded EML.*

PROOF. Let $X$ be semilinear. Then there is a $w=\langle w_1, \cdots, w_k\rangle$ such that $X$ is semilinear in $w$, i.e. $L=\{(i(1), \cdots, i(k))/w_1^{i(1)} \cdots w_k^{i(k)}$ in $X\}$ is a semilinear subset of $N^k$. Let $a_1, \cdots, a_k$ be $k$ distinct symbols not in $\Sigma$ and $h$ the homomorphism which maps each $a_i$ into $w_i$. Then by Theorem 2.2 of [5], $Y=\{a_1^{i(1)} \cdots a_k^{i(k)}/w_1^{i(1)} \cdots w_k^{i(k)}$ in $X\}$ is an EML. By the corollary to Theorem 3.2 of [6] homomorphism preserves EML. Hence $X$ is a bounded EML.

Now suppose $X$ be a bounded EML. $Y=h^{-1}(X)\cap a_1^* \cdots a_k^*$. By the corollary to Theorem 3.5 of [6] inverse homomorphism preserves EML and by Theorem 3.1 of [6] the intersection of an EML and a regular set is an EML. Hence $Y$ is an EML since $a_1^* \cdots a_k^*$ is regular. Again by Theorem 2.1 of [5], $L$ is a semilinear subset of $N^k$. Thus $X$ is semilinear.

Therefore the class of bounded EML is equivalent to the class of semilinear sets and includes the bounded context-free languages and hence the bounded regular sets.

Notation. Let $Z$ be a bounded set $\subseteq x_1^* \cdots x_k^*$, i.e. every $z$ in $Z$ is of the form $x_1^{i(1)} \cdots x_k^{i(k)}$, $x_1, \cdots, x_k$ being words in $\Sigma^*$. Then we write $Z\langle y_1, \cdots, y_k\rangle^* = \bigcup_{i\geq 0} z_1 y_1^i z_2 y_2^i \cdots z_k y_k^i$ where $y_1, \cdots, y_k$ are words in $x_1^*, \cdots, x_k^*$ respectively, and $z_1=x_1^{i(1)}$, $z_2=x_2^{i(2)}, \cdots, z_k=x_k^{i(k)}$ where $z=z_1\cdots z_k=x_1^{i(1)}\cdots x_k^{i(k)}$ is in $Z$. Inductively we write $Z\langle y_{11}, \cdots, y_{k1}\rangle \cdots \langle y_{1n}, \cdots, y_{kn}\rangle^* = Z\langle y_{11}, \cdots, y_{k1}\rangle \cdots \langle y_{1n-1}, \cdots, y_{kn-1}\rangle^* \langle y_{1n}, \cdots, y_{kn}\rangle^*$ where $y_{11}, \cdots, y_{1n}$ are words in $x_1^*, \cdots, y_{k1}, \cdots, y_{kn}$ are words in $x_k^*$.

Corollary 1. *Let $w_1, \cdots, w_k$ be words in $\Sigma^*$. Each EML$\subseteq w_1^* \cdots w_k^*$ is the finite union of sets of the form*

$$x\langle y_{11}, \cdots, y_{k1}\rangle \cdots \langle y_{1n}, \cdots, y_{kn}\rangle^*$$

*where each $y_{rm}$ is in $w_r^*$, $r=1, \cdots, k$; $m=1, \cdots, n$ and $x=x_1\cdots x_k$ where $x_r$ is in $w_r^*$; and conversely each finite union of sets of the above form is an EML$\subseteq w_1^* \cdots w_k^*$.*

Corollary 2. *The family of bounded EML is the smallest family of sets containing all finite sets and closed with respect to the following operations:*

(a) *finite union,*

(b) *finite product,*

(c) *$Z\langle x_1, \cdots, x_k\rangle^*$ where $x_1, \cdots, x_k$ are words.*

This is related to Theorem 3.1 of [1]. In view of Theorem 3.2 of [4] we obtain the following

Corollary 3. *$S(L)$ is a bounded EML for each bounded EML $L$ and each gsm $S$.*

3. **Decidability.** In this section, we consider the problem of determining of an arbitrary EML whether or not it is semilinear. We shall show that there is a decision procedure. Also another simple characterization of semilinear sets is given.

Notation. For each EMG $G$ of order $k$ and for each $k$-tuple of nonterminals $\langle A_1, \cdots, A_k\rangle$ let

$$X_{A_1}(G) = \{u_1/u_1 \text{ in } \Sigma^*,\ A_1 \cdots A_k \overset{*}{\Rightarrow} u_1A_1 \cdots u_kA_k \quad \text{for some } u_2, \cdots, u_k \text{ in } \Sigma^*\},$$

$$X_{A_2}(G) = \{u_2/u_2 \text{ in } \Sigma^*,\ A_1 \cdots A_k \overset{*}{\Rightarrow} u_1A_1 \cdots u_kA_k \quad \text{for some } u_1, u_3, \cdots, u_k \text{ in } \Sigma^*\},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$X_{A_k}(G) = \{u_k/u_k \text{ in } \Sigma^*,\ A_1 \cdots A_k \overset{*}{\Rightarrow} u_1A_1 \cdots u_kA_k \quad \text{for some } u_1, \cdots, u_{k-1} \text{ in } \Sigma^*\}.$$

The results that follow are obtained by suitably modifying the methods of Ginsburg and Spanier [1].

LEMMA 3.1. *If $L(G)$ is nonempty and bounded where $G$ is of order $k$, then $X_{A_1}, \cdots, X_{A_k}$ are all commutative for each $k$-tuple $\langle A_1, \cdots, A_k\rangle$.*

PROOF. Let $G=(V, \Sigma, P, S)$ be the EMG generating $L$. Assume that $S$ depends on each $k$-tuple of nonterminals in $G$, and that $W_A = \{w_1 \cdots w_k/A_1 \cdots A_k \overset{*}{\Rightarrow} w_1 \cdots w_k, w_r \text{ in } \Sigma^*\}$ is nonempty for each $k$-tuple $\langle A_1, \cdots, A_k\rangle = A$ in $G$. Since $S$ depends on $A$, there exist $u_1, \cdots, u_k$ in $\Sigma^*$ such that $\{u_1w_1 \cdots u_kw_k/w_1 \cdots w_k \text{ in } W_A\} \subseteq L(G)$. Thus $W_A$ is nonempty and bounded. Let $x_1 \cdots x_k$ be a specific word in $W_A$.

Consider the set $X_{A_1}(G)$. Suppose there are words $u_1$ and $v_1$ in $X_{A_1}$ so that $u_1v_1 \neq v_1u_1$. It is easily seen that for each $w_1$ in $\{u_1, v_1\}^* - \epsilon$ there are words $w_2, \cdots, w_k$ in $\Sigma^*$ so that $A_1 \cdots A_k \overset{*}{\Rightarrow} w_1A_1 \cdots w_kA_k \overset{*}{\Rightarrow} w_1x_1 \cdots w_kx_k$. Hence $(u_1, u_2)^* - \epsilon \subseteq X_{A_1}$ and $w_1x_1 \cdots w_kx_k$ is in $W_A$. $\epsilon$ is also in $X_{A_1}$. Thus each word $w_1 \cup \epsilon$ in $\{u_1, u_2\}^*$ is a subword of some word $w_1x_1 \cdots w_kx_k$ in $W_A$. By Lemma 5.3 of [1], $W_A$ is not bounded. This is a contradiction. Therefore $u_1u_2 = u_2u_1$ for every two words $u_1, u_2$ in $X_{A_1}(G)$ i.e. $X_{A_1}(G)$ is a commutative set.

A similar argument shows that $X_{A_2}(G), \cdots, X_{A_k}(G)$ are all commutative sets.

LEMMA 3.2. *If $X_{A_1}(G), \cdots, X_{A_k}(G)$ are all commutative sets for each $k$-tuple $\langle A_1, \cdots, A_k\rangle$ of $G$ of order $k$, then $L(G)$ is bounded.*

PROOF. The proof is by induction on the number of $k$-tuples of nonterminals. Suppose $\langle A_1, \cdots, A_k\rangle$ is the only nonterminal in $G$, apart from $S$. By Lemma 5.2 of [1], $X_{A_1} \subset u_1^*, \cdots, X_{A_k}^* \subseteq u_k^*$ for some words $u_1, \cdots, u_k$ in $\Sigma^*$. Let all the initial and terminal rules of $G$ be $[S \to f_{1j}A_1 \cdots f_{kj}A_k]$, $j=1, \cdots, m$;

$$\begin{bmatrix} A_1 \to w_{1i} \\ \cdot\ \cdot\ \cdot\ \cdot \\ A_k \to w_{ki} \end{bmatrix} \qquad i = 1, \cdots, n.$$

If $y$ be any word in $L(G)$, there is some $S$-derivation of $y$ as $S \Rightarrow f_{1j}A_1 \cdots f_{kj}A_k \overset{*}{\Rightarrow} f_{1j}v_1A_1 \cdots f_{kj}v_kA_k \Rightarrow f_{1j}v_1w_{1i} \cdots f_{kj}v_kw_{ki}$, $1 \leqq j \leqq m$, $1 \leqq i \leqq n$, $v_1, \cdots, v_k$ in $X_{A_1}, \cdots, X_{A_k}$ which are subsets of $u_1^*, \cdots, u_k^*$. Thus

$$L(G) \subseteq \bigcup_{j=1}^{m} \left[ \bigcup_{i=1}^{n} f_{1j}u_1^*w_{1i} \cdots f_{kj}u_k^*w_{ki} \right].$$

Therefore $L(G)$ is bounded.

Suppose that $G$ has $p$ $k$-tuples of variables $\langle A_{1i}, \cdots, A_{ki}\rangle$, $i = 1, \cdots, p$, where $p > 1$ and that the lemma is true for all grammars with fewer than $p$ variables. Let $G_j$ be the grammar obtained from $G$ by deleting all the production rules involving $\langle A_{1j}, \cdots, A_{kj}\rangle$. Let $Y_{A_{1i}}(G_j), \cdots, Y_{A_{ki}}(G_j)$ be the set of words $y_{1i}, \cdots, y_{ki}$ such that $A_{1i} \cdots A_{ki} \overset{*}{\Rightarrow}_{G_j} y_{1i} \cdots y_{ki}$ in $\Sigma^*$. $X_{A_{1i}}(G_j), \cdots, X_{A_{ki}}(G_j)$ being subsets of $X_{A_{1i}}(G), \cdots, X_{A_{ki}}(G)$ are all bounded. By the induction hypothesis $L(G_j)$ is bounded. $Y_{A_{1i}}(G_j), \cdots, Y_{A_{ki}}(G_j)$ consisting of subwords of words in $L(G_j)$ are bounded. Let there be $q$ initial rules $[S \to f_{1j}A_{1j} \cdots f_{kj}A_{kj}]$, $j = 1, \cdots, q$. For each such $j$, consider

$$(**) \qquad f_{1j}X_{A_{1j}}(G)g_{1i}Y_{A_{1i}}(G_j) \cdots f_{kj}X_{A_{kj}}(G)g_{ki}Y_{A_{ki}}(G_j),$$

$i$ in $\{1, \cdots, p\} - \{j\}$ where

$$\begin{bmatrix} A_{1j} \to g_{1i}A_{1i} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ A_{kj} \to g_{ki}A_{ki} \end{bmatrix}$$

are all the rules of $G$ with $\langle A_{1j}, \cdots, A_{kj}\rangle$ occurring on the left side. (When the above rule is terminal, the $Y$'s are empty.) Since there are only a finite number of such rules the sets (**) are bounded. The proof is completed by noting that

$$L(G) \subseteq \bigcup_{j=1}^{q} f_{1j}X_{A_{1j}}g_{1i}Y_{A_{1i}} \cdots f_{kj}X_{A_{kj}}g_{ki}Y_{A_{ki}}.$$

Combining Lemmas 2.1 and 2.2 we get

THEOREM 3.1. *A necessary and sufficient condition that an EML*

*$L(G) \neq \emptyset$ be semilinear is that $X_{A_1}(G), \cdots, X_{A_k}(G)$ be all commutative for each variable $\langle A_1, \cdots, A_k \rangle$ in $G$ of order $k$.*

LEMMA 3.3. *For each variable $\langle A_1, \cdots, A_k \rangle$ in $G$ of order $k$, $X_{A_1}(G), \cdots, X_{A_k}(G)$ are regular sets and effectively determined.*

The proof is obvious from the definition of an EMG that all rules except the initial rules consist of $k$ left-linear rules.

Now from Lemma 2.3, and Lemmas 5.7 and 5.8 of [1] and the proof of Lemma 2.2, the following decision theorem is immediate.

THEOREM 3.2. (a) *It is decidable whether or not a given EML $L(G)$ is bounded.*

(b) *If $L(G)$ is bounded then words $w_1, \cdots, w_t$ in $\Sigma^*$ can be effectively found so that $L(G) \subseteq w_1^* \cdots w_1^*$.*

THEOREM 3.3. *If $L_1$, $L_2$ are EML and one of them is semilinear, then it is solvable whether* (a) $L_1 \subseteq L_2$ *and whether* (b) $L_2 \subseteq L_1$.

Proof is immediate from the proof of Theorem 6.3 of [1] using the corresponding results for EML obtained in Theorems 2.2 and 1.1.

COROLLARY. *If $L_1$, $L_2$ are EML and one of them is semilinear then it is solvable whether $L_1 = L_2$.*

Several of the mathematical properties of semilinear sets proved in [4] can also be established by considering bounded EML.

## REFERENCES

**1.** S. Ginsburg and E. H. Spanier, *Bounded* ALGOL-*like languages*, Trans. Amer. Math. Soc. **113** (1964), 333–368.

**2.** ———, *Semigroups, Presburger formulas, and languages*, Pacific J. Math. **16** (1966), 285–296.

**3.** ———, *Bounded regular sets*, Proc. Amer. Math. Soc. **17** (1966), 1043–1049.

**4.** S. Ginsburg and J. Ullian, *Ambiguity in contextfree languages*, J. Assoc. Comput. Mach. **13** (1966), 62–89.

**5.** R. Siromoney, *On equal matrix languages*, Information and Control, (to appear).

**6.** ———, *k-turn one-way stack automata*, (to appear).

MADRAS CHRISTIAN COLLEGE, INDIA