JOURNAL OF COMPUTER AND SYSTEM SCIENCES: 4, 299–307 (1970)

A Characterization of Semilinear Sets*

L. Y. LIU

International Business Machines Corporation, Yorktown Heights, New York 10598

AND

P. WEINER

Computer Science Department, Yale University, New Haven, Connecticut 06520

Received December 2, 1968

In this report, the class of semilinear sets is shown to be the least class of sets which contains all of the stratified semilinear sets and is closed under finite intersection.

I. INTRODUCTION

The notions of linear and semilinear sets were first introduced by Parikh in 1961 [7]; bounded context free languages were introduced by Ginsburg and Spanier in 1964 [3]. The two concepts were related in 1966 by Ginsburg and Spanier [4]. They showed that a subclass of the semilinear sets (called stratified semilinear sets) can be used to completely characterize the bounded context free languages. This characterization makes the class of bounded context free languages very useful in the study of inherent ambiguity [5] and in the construction of counter-examples [2, 6].

In this report, this characterization will be used to prove the following theorem.

THEOREM 1. *The class of semilinear sets is the least class of sets which contains all of the stratified semilinear sets and is closed under finite intersection.*

In the remainder of this section, we introduce notations and definitions, state the known theorems, and prove a preliminary result. The proof of Theorem 1, which is quite complicated, is the subject of the rest of this report.

* This work was supported in part by Bell Telephone Laboratories and NSF Grant GK-1735.

571/4/4-1

DEFINITION 1. A language $L \subseteq \Sigma^*$ is *bounded* if there exists a finite number of finite length strings $w_1, w_2, \ldots, w_n$ such that $L \subseteq w_1^* w_2^* \cdots w_n^*$, where Σ is a finite nonempty set.

The set of natural numbers will be denoted by N. It is natural to associate with any given set

$$S \subseteq N^n \quad (= \underbrace{N \times N \times \cdots \times N}_{n \text{ times}})$$

a bounded language in the following way. For each n-tuple of words $w = \langle w_1, w_2, \ldots, w_n \rangle$, let f_w be the mapping from N^n onto $w_1^* w_2^* \cdots w_n^*$ defined by $f_w(i_1, i_2, \ldots, i_n) = w_1^{i_1} w_2^{i_2} \cdots w_n^{i_n}$. Then if $S \subseteq N^n$, $f_w(S)$ is a bounded language $L \subseteq w_1^* w_2^* \cdots w_n^*$.

DEFINITION 2. A set $S \subseteq N^n$ is a *linear set* if there exist $\alpha_0, \alpha_1, \ldots, \alpha_m$ in N^n such that $S = \{\beta \mid \beta = \alpha_0 + \sum_{i=1}^{m} a_i \alpha_i$ where a_i is in N for $1 \leqslant i \leqslant m\}$. S is denoted by $L(\alpha_0; \alpha_1, \alpha_2, \ldots, \alpha_m)$, and $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is called *the set of periods* of L. α_0 is called the *constant vector* of L. $L(\alpha_1, \alpha_2, \ldots, \alpha_m)$ is used to denote $L(\alpha_0; \alpha_1, \alpha_2, \ldots, \alpha_m)$ when $\alpha_0 = 0$.

DEFINITION 3. A set $S \subseteq N^n$ is a *semilinear set* if it is expressible as a finite union of linear sets.

THEOREM 2 (Ginsburg and Spanier). *If X and Y are semilinear subsets of N^n, then $X \cap Y$ and $X \cup Y$ are semilinear and are effectively calculable from X and Y* [2, 4].

DEFINITION 4. A subset X of N^n is said to be *stratified* if the following two conditions are satisfied: (i) Each element in X has at most two nonzero coordinates, (ii) there are no integers i, j, k, and m where $1 \leqslant i < j < k < m \leqslant n$ such that $x = (x_1, x_2, \ldots, x_n)$ and $x' = (x_1', x_2', \ldots, x_n')$ are in X where $x_i x_j' x_k x_m' \neq 0$.

DEFINITION 5. A set $S \subseteq N^n$ is a *stratified semilinear set* if S is expressible as a finite union of linear sets, each with a stratified set of periods.

The following theorem completely characterizes the semilinear sets corresponding to bounded context free languages [2].

THEOREM 3 (Ginsburg and Spanier). *A language $L \subseteq w_1^* w_2^* \cdots w_n^*$ is a context free language*[1] *if and only if there exists a stratified semilinear set $S \subseteq N^n$ such that $f_{\langle w_1, w_2, \ldots, w_n \rangle}(S) = L$.*

[1] We assume that basic concepts of context free languages and pushdown store machines are familiar to the reader. A good reference is Ginsburg [2].

The following lemma is useful in reducing problems concerning semilinear sets to problems concerning linear sets.

LEMMA 1. *If a semilinear set S is expressible as a finite union of finite intersections of stratified semilinear sets, then S is expressible as a finite intersection of stratified semilinear sets.*

Proof. Let $S = \bigcup_{i=1}^{m}(\bigcap_{j=1}^{n_i} S_{ij})$ where S_{ij} is a stratified semilinear set for $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n_i$. By the distributive laws of the union and intersection of sets, S can be manipulated into a finite intersection of finite unions of stratified semilinear sets. By definition, the class of stratified semilinear sets is closed under finite union. The theorem follows immediately. Q.E.D.

The proof of Theorem 1 will be spread over the next four sections. First, in Section II, Lemma 1 is used to derive an equivalent theorem. This formulation leads naturally, in Section III, to the use of linear Diophantine equations. An interesting property of these equations is derived. This property is used in Section IV to define certain sets. The proof will be completed in Section V, where these sets are shown to be stratified semilinear sets.

II. AN EQUIVALENT THEOREM

By Theorem 2, the class of semilinear sets is closed under finite intersection. Moreover, every stratified semilinear set is trivially a semilinear set. It follows that the class of sets obtained from the stratified semilinear sets by closure under finite intersection only contains semilinear sets. Thus, Theorem 1 is true if and only if *every* semilinear set may be expressed as a finite intersection of stratified semilinear sets. But every semilinear set is expressible as a finite union of linear sets. These facts allow us to establish Theorem 1 by proving the following theorem, which is equivalent to Theorem 1.

THEOREM 1′. *Every linear set with a zero constant vector is expressible as a finite intersection of stratified semilinear sets.*

Proof (*of equivalence*). Theorem 1 clearly implies Theorem 1′. Suppose, then, that Theorem 1′ is true. If S is a linear set with constant vector α_0, we may express S as a finite intersection of stratified semilinear sets as follows. First express the linear set that has a zero constant vector and the same set of periods as S as a finite intersection of stratified semilinear sets. The desired expression of S is obtained by merely adding α_0 to each of the constant vectors of these stratified semilinear sets. This allows us to express every semilinear set as a finite union of finite intersections of stratified semilinear sets. It follows from Lemma 1 that Theorem 1 is also true. Q.E.D.

III. Linear Diophantine Equations

Suppose that $\alpha_1, \alpha_2, \ldots, \alpha_m$ are fixed vectors[2] from N^n. The linear set $S = L(\alpha_1, \alpha_2, \ldots, \alpha_m)$ is the set $\{\beta \mid \beta = \sum_{i=1}^{m} a_i \alpha_i$, where a_i is in N for $1 \leqslant i \leqslant m\}$. Alternatively, a given vector β in N^n is in S if and only if there exists an element a in N^n such that

$$A \cdot a = \beta$$

where A is the $n \times m$ matrix formed from the α_i, i.e., $A = [\alpha_1, \alpha_2, \ldots, \alpha_m]$. This implies that the problem of deciding whether a given element is in S is equivalent to the problem of deciding the existence of a solution with nonnegative integral components for a set of simultaneous linear Diophantine equations.

The decision problems referred to above are solvable; indeed, an algorithm for this purpose follows directly from an interesting property of linear Diophantine equations. Since we have been unsuccessful in finding this property in the literature, we present it as the next theorem and give a brief proof.

Theorem 4. *Let $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ be a set of fixed vectors from N^n that contains a basis for the vector space R^n $(m \geqslant n)$. Let $A = [\alpha_1, \alpha_2, \ldots, \alpha_m]$ denote the $n \times m$ matrix formed by the given vectors, and let β be an arbitrary vector from N^n. If the system of n simultaneous linear Diophantine equations*

$$A \cdot x = \beta \tag{1}$$

has a solution for x in N^m, say $a = (a_1, a_2, \ldots, a_m)$, then the system has another solution for x in N^m, say $b = (b_1, b_2, \ldots, b_m)$, such that at least $m - n$ of the b_i are less than $\Delta(A)$, the maximum magnitude of all $n \times n$ subdeterminants of the matrix A. Moreover, the set $\{\alpha_i \mid b_i \geqslant \Delta(A)$ for $1 \leqslant i \leqslant m\}$ is a set of linearly independent vectors.

Proof. We assume that the solution a exists, and we show how to obtain the solution b. Consider first the special case $m = n + 1$. Using some simple linear algebra, it is easy to obtain from a another vector $c = (c_1, c_2, \ldots, c_m)$ with *real* components such that (i) $A \cdot c = \beta$, (ii) $c_i \geqslant 0$ for $1 \leqslant i \leqslant m$, (iii) some component of c, say c_j, is zero, and (iv) $\{\alpha_i \mid 1 \leqslant i \leqslant m,\ i \neq j\}$ forms a basis of R^n. Let $B = [\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_m]$, and let δ denote the determinant of B. If x_j is fixed, then the other components of x are uniquely determined by (1) and may be expressed in the form

$$x_i = \frac{d_i + e_i x_j}{\delta} \qquad i = 1, \ldots, j-1, j+1, \ldots, m \tag{2}$$

[2] Strictly speaking, elements of N^n are n-tuples of natural numbers. When convenient, however, we will associate (column) vectors with these elements. The implied vector space is R^n, the space of n-tuples of real numbers over the real field.

where δ and each of the d_i and e_i are integers. Thus a solution b with nonnegative integral components is among the sequence of solutions obtained from (2) by setting $x_j = 0, 1, \ldots, |\delta| - 1$.

For the general case, the desired solution may be obtained from a as follows. Arbitrarily select $n + 1$ vectors that contain a basis of R^n, say $\alpha_1, \alpha_2, \ldots, \alpha_{n+1}$, and form the n equations $\sum_{i=1}^{n+1} x_i\alpha_i = \sum_{i=1}^{n+1} a_i\alpha_i$. Since these equations satisfy the special case condition, there exists a vector $(b_1, b_2, \ldots, b_{n+1})$ such that some component, say b_j, is less than $\Delta(A)$, and $\{\alpha_i \mid i = 1, \ldots, n + 1, i \neq j\}$ is a basis of R^n. Effectively remove the vector α_j by bringing $b_j\alpha_j$ to the right hand side, i.e., fix the value of x_j at b_j and write

$$\sum_{\substack{i=1\\ i\neq j}}^{m} x_i\alpha_i = \beta - b_j\alpha_j = \beta'.$$

The theorem is proved constructively if this procedure is repeated $m - n$ times. Q.E.D.

IV. Decomposition of Linear Sets

In this section we use Theorem 4 to express any linear set with a zero constant vector as a finite union of finite intersections of certain sets. In the next section, we show that each of these sets is a stratified semilinear set. According to Lemma 1, this will complete the proof of Theorem 1′.

Let the linear set be $S = L(\alpha_1, \alpha_2, \ldots, \alpha_m)$, where α_i is in N^p for $1 \leqslant 1 \leqslant m$, and let $A = [\alpha_1, \alpha_2, \ldots, \alpha_m]$ and $P = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$. Thus $S = \{\beta \mid A \cdot x = \beta$ for some x in $N^m\}$. Suppose that the vectors in P span a n-dimensional subspace of R^p, say V. Without loss of generality, assume that the first n rows of A are linearly independent. If β is not in V, then β is certainly not in S. On the other hand, if β is in V, then β is in S if and only if the first n equations of $A \cdot x = \beta$ has a solution with nonnegative integral components. In order to conveniently express these n equations, let α_i', for $1 \leqslant i \leqslant m$, and β' denote vectors in N^n obtained by retaining the first n components of the α_i and β, and let $A' = [\alpha_1', \alpha_2', \ldots, \alpha_m']$ and $P' = \{\alpha_1', \alpha_2', \ldots, \alpha_m'\}$. Now we may express S as

$$S = \{\beta \mid \beta \text{ is in } V, \text{ and } A' \cdot x = \beta' \text{ for some } x \text{ in } N^m\}.$$

In order to apply Theorem 4, we need an enumeration of the distinct subsets of P' that contain exactly n linearly independent elements. Let $P_1, P_2, \ldots, P_q$ denote the enumeration. [Note that $q \leqslant \binom{m}{n}$.] We can now express S as a finite union of sets, namely

$$S = \bigcup_{i=1}^{q} S_i \tag{3}$$

where

$$S_i = \{\beta \mid \beta \text{ is in } V \text{ and } A' \cdot x = \beta' \text{ for some } x = (x_1, x_2, ..., x_m) \text{ in } N^m \text{ such that } x_j < \Delta(A') \text{ if } \alpha_j \text{ is not in } P_i\}.$$

In each of these sets there are exactly $m - n$ components of x that must be less than $\Delta(A')$. We call these components the *constrained variables*, and the remaining n components the *free variables*. For each P_i, let $\{y_j^{(i)} \mid 1 \leqslant j \leqslant n\}$ and $\{z_j^{(i)} \mid 1 \leqslant j \leqslant m - n\}$ denote the free and constrained variables, respectively. We next express each S_i as a finite union of sets that have fixed values of the constrained variables. Define $S^{(1)}_{(a_1, a_2, ..., a_{m-n})}$ to be the set

$$\{\beta \mid \beta \text{ is in } S_i \text{ and } z_1^{(i)} = a_1, z_2^{(i)} = a_2, ..., z_{m-n}^{(i)} = a_{m-n}\}.$$

An ordered set of $m - n$ nonnegative integers is called an *allowable configuration* if each component is less than $\Delta(A')$. Let $\gamma_1, \gamma_2, ..., \gamma_s$ denote an enumeration of the distinct allowable configurations ($s = \Delta(A')^{m-n}$). Then

$$S_i = \bigcup_{j=1}^{s} S_{\gamma_j}^{(i)}. \tag{4}$$

By combining (3) and (4), we have

$$S = \bigcup_{\substack{i=1,...,q \\ j=1,...,s}} S_{\gamma_j}^{(i)}.$$

We complete this section by expressing each set $S_{\gamma_j}^{(i)}$ as an intersection of p sets. There is one of these sets for each free variable. The other $p - n$ sets insure that the vector β is in the subspace spanned by the given periods. It is easy to see that

$$S_{\gamma_j}^{(i)} = \bigcap_{k=0}^{n} S_k^{(i,j)}$$

where

$$S_0^{(i,j)} = \{\beta \mid \beta \text{ is in } N^p \text{ and } \beta \text{ is in } V\}$$

and, for $k = 1, 2, ..., n$,

$$S_k^{(i,j)} = \{\beta \mid \beta \text{ is in } N^p \text{ and } A' \cdot x = \beta' \text{ for some } x \text{ in } R^m \text{ such that } y_k^{(i)} \text{ is in } N \text{ and the constrained variables corresponding to } P_i \text{ are equal to the allowable configuration } \gamma_j\}.$$

It is convenient to express this set in a slightly different form. Note that since the constrained variables are fixed at the allowable configuration γ_j, we may use Cramer's rule to determine the values of the free variables. Thus, for every $l = (i, j, k)$, there exist integers b_l and δ_l, and a row vector $c_l = (c_1{}^l, c_2{}^l, \ldots, c_m{}^l)$ with integral components such that

$$S_k^{(i,j)} = \left\{\beta \,\middle|\, \beta \text{ is in } N^p \text{ and } y_k^{(i)} = \frac{b_l + c_l \cdot \beta'}{\delta_l} \text{ is in } N\right\}.$$

Finally, we express the set $S_0^{(i,j)}$ (which is actually independent of both i and j) as a finite intersection of $p - n$ sets. First, express each of the $p - n$ rows of A that is not in A' as a linear combination of the rows of A'. Each of these expressions may be used to form a row vector $c_i = (c_1{}^i, c_2{}^i, \ldots, c_p{}^i)$, with integral components, such that $c_i \cdot A = 0$ for $1 \leqslant i \leqslant p - n$. Now β is in V if and only if $c_i \cdot \beta = 0$ for $1 \leqslant i \leqslant p - n$. Thus we may express $S_0^{(i,j)}$ as

$$\bigcap_{i=1}^{p-n} \{\beta \mid \beta \text{ is in } N^p \text{ and } c_i \cdot \beta = 0\}.$$

V. Verification that Sets are Stratified Semilinear

In Section IV it was shown that every linear set S with a zero constant vector may be expressed as a finite union of finite intersections of certain sets. These sets were of two distinct types. We complete the proof of Theorem 1 by showing that both types of sets are stratified semilinear sets. In each case, the characterization theorem of Ginsburg and Spanier will be used to convert the problem into one concerning pushdown store machines and context free languages.

Theorem 5. *Let $c = (c_1, c_2, \ldots, c_n)$ be a row vector with integral components. Then the set*

$$S = \{\beta \mid \beta \textit{ is in } N^n \textit{ and } c \cdot \beta = 0\}$$

is a stratified semilinear set.

Proof. Let $L = f_{\langle a_1, a_2, \ldots, a_n \rangle}(S)$, where $a_1, a_2, \ldots, a_n$ are distinct symbols. By Theorem 2, it suffices to show that L is context free. Since one way non-deterministic pushdown automata (1NPDA) accept precisely the context free languages,[3] we will prove this theorem by constructing a 1NPDA M such that Null $(M) = L$.[4]

[3] A definition of INPDA may be found on page 59 of [2].

[4] A definition of Null (M) for a 1NPDA M may be found on page 63 of [2].

Let $M = (K, \Sigma, \Gamma, \delta, q_0, Z_0, F)$. M forms the linear combination $c \cdot \beta$ on the stack and checks whether the result is zero.

(i) $K = \{[q_i, m] \mid 0 \leqslant m \leqslant |c_i| \text{ and } 1 \leqslant i \leqslant n\} \cup \{q_0\}$.

q_i remembers that the last symbol M reads is a_i. m remembers that m markers are stored in the finite control.

(ii) $\Sigma = \{a_1, a_2, ..., a_n\}$.

(iii) $\Gamma = \{Z_0, Z_1, Z_2\}$.

Z_1 is a marker for "+1". Z_2 is a marker for "−1".

(iv) $\delta(q_0, a_i, Z_0) = ([q_i, 0], Z_0 Z_l^{|c_i|})$ where $l = 1$ if $c_i > 0$ and $l = 2$ if $c_i \leqslant 0$ for $1 \leqslant i \leqslant n$.

For $1 \leqslant i \leqslant j \leqslant n$, $\delta([q_i, 0], a_j, Z_0) = ([q_j, 0], Z_0 Z_l^{|c_i|})$, where $l = 1$ if $c_j > 0$ and $l = 2$ if $c_i \leqslant 0$.

For $1 \leqslant i \leqslant j \leqslant n$, $\delta([q_i, 0], a_j, Z_l)$
$= ([q_j, 0], Z_l Z_l^{|c_i|})$ if $l = 1$ and $c_j \geqslant 0$ *or* if $l = 2$ and $c_j \leqslant 0$.
$= ([q_j, |c_j| - 1], \epsilon)$ if $l = 1$ and $c_j < 0$ *or* if $l = 2$ and $c_j > 0$.

For $1 \leqslant i \leqslant n$, $\delta([q_i, m], \epsilon, Z_l)$
$= ([q_i, m - 1], \epsilon)$ if $l = 1$ and $c_i < 0$ *or* if $l = 2$ and $c_i > 0$
$= ([q_i, 0], Z_l Z_l^{m})$ if $l = 1$ and $c_i > 0$ *or* if $l = 2$ and $c_i < 0$.

For $1 \leqslant i \leqslant n$, $\delta([q_i, 0], \epsilon, Z_0) = ([q_i, 0], \epsilon)$.

It is easy to show that Null $(M) = L$. Q.E.D.

THEOREM 6. *Let b and δ be integers, and let $c = (c_1, c_2, ..., c_n)$ be a row vector with integral components. Then the set*

$$S = \left\{\beta \;\middle|\; \beta \text{ is in } N^n \text{ and } \frac{b + c \cdot \beta}{\delta} \text{ is in } N\right\}$$

is a stratified semilinear set.

Proof. Let $L = f_{\langle a_1, a_2, ..., a_n\rangle}(S)$, where $a_1, a_2, ..., a_n$ are distinct symbols. As in the proof of Theorem 5, it suffices to construct a 1NPDA M such that Null $(M) = L$. The formal construction is left to the reader. Informally, M operates as follows: The sum $b + c \cdot \beta$ is represented on the stack by the end of the scan of the input tape. It is then checked if this sum is of the same sign as δ, and if it is an integral multiple of δ. Only if both these conditions are satisfied is the stack emptied. Q.E.D.

REFERENCES

1. J. EVEY, The theory and application of pushdown store machines, Harvard University Doctoral Thesis, Cambridge, Mass., 1963.
2. S. GINSBURG, "The Mathematical Theory of Context Free Languages," McGraw–Hill, New York, 1966.
3. S. GINSBURG AND E. H. SPANIER, Bounded ALGOL-like languages, *Trans. Am. Math. Soc.* **113** (1964), 333–368.
4. S. GINSBURG AND E. H. SPANIER, Semigroups, Presburger formulas, and languages, *Pacif. J. Math.* **16** (1966), 285–296.
5. S. GINSBURG AND J. S. ULLIAN, Ambiguity in context free languages, *J. Assoc. Comput. Mach.* **13** (1966), 62–89.
6. L. Y. LIU AND P. WEINER, An infinite hierarchy of finite intersection of context free languages, Computer Sciences Laboratory Technical Report Number 65, Electrical Engineering Department, Princeton University, Princeton, N.J., June, 1968.
7. R. J. PARIKH, Language generating devices, M.I.T. Research Laboratory of Electronics Quarterly Progress Report, Number 60, pp. 199–212, 1961.
8. R. J. PARIKH, On context-free languages, *J. Assoc. Comput. Mach.* **13** (1966), 570–581.