A CHARACTERIZATION OF BOUNDED REGULAR SETS

Antonio Restivo

Laboratorio di Cibernetica del CNR
Arco Felice, Napoli.

1. INTRODUCTION

Let X be a finite alphabet and X^* the free monoid generated by X. A subset L of X^* is bounded iff there exists a finite number of words $w_1, w_2, \ldots, w_k$ of X^* such that $L \subset w_1^* w_2^* \ldots w_k^*$. Bounded languages were introduced in [1] by Ginsburg and Spanier as a special family of context-free languages which has "simple" structural properties and is intimately related to certain algebraic concepts. In [2] Ginsburg and Spanier considered also the particular case of bounded regular sets for which derived some characterization results.

In this paper the boundedness of a language is related with the presence in it of strings "without repetitions" of arbitrary length, and a new characterization of bounded regular sets is given (theorem 2). The problem of existence of strings "without repetitions" of arbitrary length was first considered and solved by Thue [3] giving an explicit construction of such strings. Later the same property was rediscovered by other people in different context as symbolic dynamics and the theory of semigroups. Many references on the subject can be found in [4]. To formalize these notions let us now introduce for any positive integer p the following subset of X^* :

$$L_p = \{ f \in X^* \mid f \neq uv^p w \text{ for all } u, v, w \in X^* \text{ with } |v| > 0 \}.$$

In the sequel, if it is necessary, we shall use the symbol $L_p(X)$ to specify the alphabet X on which L_p is defined. We note that if

$p \leqq p'$, $L_p \subset L_{p'}$. The basic result of Thue for our pourposes can be stated as follows

Theorem 1 (Thue). Let $|X|$ be the cardinality of the alphabet X. If $|X| = 2$, L_3 is infinite; if $|X| > 2$, L_2 is infinite.

Given an arbitrary subset L of X^*, we consider now the intersections $L \cap L_p$ for any p. We have the following

Proposition 1 If L is bounded, $L \cap L_p$ is finite for any p.

Proof. Let $L \subset w_1^* w_2^* \ldots w_k^*$. If $f \in L$, there exist k positive integers $n_1, n_2, \ldots, n_k$ such that $f = w_1^{n_1} w_2^{n_2} \ldots w_k^{n_k}$. The length of f is then $|f| = n_1|w_1| + n_2|w_2| + \ldots + n_k|w_k|$. Let r be the maximal length of the words $w_1, w_2, \ldots w_k$. For any $p > o$, if $|f| \geqq pkr$, there exists an index j such that $n_j \geqq p$. Hence, for any $f \in L$ such that $|f| \geqq pkr$, $f \notin L_p$. This completes the proof of the proposition.

Proposition 1 in the case $L = X^*$ gives immediately, by theorem 1, the following result of Ginsburg and Spanier [1] concerning the existence of unbounded sets.

Corollary If $|X| \geqq 2$, X^* is not bounded.

The converse of the proposition 1 is not generally true. The main result of the paper is the following

Theorem 2. Let L be a regular set. L is bounded if and only if $L \cap L_p$ is finite for any p.

The hypothesis that L is regular is essential, as shown by the following counterxample.
Let $|X| \geqq 3$, $x \in X$ and $Y = X \setminus \{x\}$. Consider the language

$$L = \{f = gx^{|g|} \mid g \in Y^*\}$$

L is context - free and is not bounded; however, for any $p > o$, $L \cap L_p$ is finite. Indeed if $f \in L$ and $|f| = 2p$, then $f = gx^p$

for some $g \in Y^p$ Hence $f \notin L_p$

The proof of theorem 2 is based on a certain number of technical lemmas which are reported in the next section without proof. In Section 3 we give in detail the kernel of the proof of the theorem.

2. SOME PRELIMINARY LEMMAS.

In this section we give, without proof, some lemmas which will be used in the proof of theorem 2 in the next section.

The first lemma concerns the problem of the invariance, under a morphism θ, of the sets L_p.

Lemma 1. Let X and Y be two finite alphabets and let θ be a monomorphism from Y^* to X^*. For all words $v \in Y^*$

$$v \in L_p(Y) \Rightarrow \theta(v) \in L_{|\theta|p}(X)$$

where $|\theta|$ is the maximal length of words in $\theta(Y) \subset X^*$.

Lemma 1 is not generally true if Y is not finite and θ is not one-to-one. The next lemma concerns regular sets.

Lemma 2. Let L be a regular set. There exist two positive integers m,k, such that for all $u,v,w \in X^*$ and for all $n \geq m$

$$uv^n w \in L \iff uv^{n+k} w \in L.$$

The following notations will be now introduced. Let f and v be words of X^*. v is a subword of f iff there exist $u, w \in X^*$ such that $f = uvw$. If L is a subset of X^*, we denote by $S(L)$ the set of all the subwords of words in L.

Lemma 3 Let L be a regular set.

$$S(L) \cap L_p \text{ finite for any } p \iff L \cap L_p \text{ finite for any } p.$$

In order to state the next lemma let us now give some other definitions. An element $f \in X^*$ is primitive iff any relation $f = g^k$ implies $f = g$. Let L be a subset of X^*. An element $v \in X^*$ is

an iterating factor of L iff there exist $u,w \in X^*$ such that $uv^*w \cap L$ is infinite. If v is an iterating factor of L and $v = g^k$ for some $g \in X^*$ and $k > o$, then also g is an iterating factor of L. We are here interested only to primitive iterating factors of a set L. The proof of the following lemma makes use of theorem 1.1 of Ginsburg and Spanier in [2].

Lemma 4. Let L be a regular set having only a finite number of primitive iterating factors. Then L is bounded.

3. PROOF OF THEOREM 2

Suppose L be a regular set such that $L \cap L_p$ is finite for any p. To prove that L is bounded it is sufficient to prove, by lemma 4, that L has a finite number of primitive iterating factors. The argument is by contradiction. By lemma 3, if $L \cap L_p$ is finite for any p, also $S(L) \cap L_p$ is finite for any p. Let $\gamma(p)$ be the maximal length of words in $S(L) \cap L_p$, and let m,k be positive integers as in the lemma 2. If L has not a finite number of primitive iterating factors, there exists a primitive iterating factor v such that $|v| > \gamma(m)$. The conditions $v \in S(L)$ and $|v| > \gamma(m)$ imply that $v \notin L_m$, and then there exist $a,b,c \in X^*$, with b primitive, such that $v = ab^m c$. Since v is an iterating factor of L, there exist $u,w \in X^*$ and $n > \frac{1}{|v|}\gamma(3(|v|+k|b|))$ such that $uv^n w \in L$. We may write

$$uv^n w = u\underbrace{a\,b^m\,c\,a\,b^m\,c\,a\,b^m\,c\,\ldots\ldots\,a\,b^m\,c}_{n \text{ times}}w \in L$$

By lemma 2, if we substitute, as exponent of each factor b in $uv^n w$, m+k for m, we obtain again a word of L. Consider now the alphabet $Y = \{y_o, y_1\}$ and a map τ from Y to the integers $\{o,1\}$ defined as follows: $\tau(y_o) = o$, $\tau(y_1) = 1$. Let s be a word of Y^* of length n such that $s \in L_3(Y)$. If $s = s_1 s_2 \ldots s_n$, $s_i \in Y$, consider now the following word f of X^*, which, by the above remark, belongs to L:

$$f = uab^{m+\tau(s_1)k}cab^{m+\tau(s_2)k}cab^{m+\tau(s_3)k}c\ldots ab^{m+\tau(s_n)k}cw \in L.$$

The subword h of f defined by the equality $f = uhw$ plays an essential role in the rest of the proof.

We have clearly $|h| > n|v|$.

Let θ be the morphism from Y^* to X^* defined as follows

$$\begin{cases} \theta(y_o) = cab^m \\ \theta(y_1) = cab^{m+k} \end{cases}$$

We prove now that θ is a monomorphism. Let us introduce a new alphabet $Z = \{z_1, z_2\}$, and exprime θ as the composition of the morphism θ_1 from Y^* to Z^* and θ_2 from Z^* to X^* defined as follows:

$$\begin{cases} \theta_1(y_o) = z_1 z_2^m \\ \theta_1(y_1) = z_1 z_2^{m+k} \end{cases} \qquad \begin{cases} \theta_2(z_1) = ca \\ \theta_2(z_2) = b \end{cases}$$

θ_1 is clearly a monomorphism. θ_2, by a well known result in the theory of free monoids [5], is a monomorphism if and only if ca and b are not powers of the same word. Since b is primitive by hipothesis if ca and b are powers of the same word, there exists $q > 0$ such that $ca = b^q$. We have then

$$\begin{cases} c = b^{q_1} b_1 \\ a = b_2 b^{q_2} \end{cases}$$

with $b_1 b_2 = b$ and $q_1 + q_2 = q - 1$. If follows that the word

$$v = ab^m c = b_2 b^{q_2} b^m b^{q_1} b_1 = (b_2 b_1)^{q+m}$$

is not primitive in contradiction with the hypothesis. Then θ_2 is a monomorphism. Since the composition of two monomorphisms is a monomorphism, θ is also a monomorphism . This, with the condition $s \in L_3(Y)$, implies, by lemma 1, $\theta(s) \in L_{3|\theta|}(X)$, with $|\theta| = |v| + k|b|$.

Consider now the word $h \in S(L)$ previously defined. It is easy to see that h is a subword of $\theta(s)$ and then, clearly, h belongs to $L_{3|\theta|}$. We have then $h \in S(L) \cap L_{3|\theta|}$ and $|h| > n|v| > \gamma(3|\theta|)$, a

contradiction. Hence the starting hypothesis is not true and L has only a finite number of primitive iterating factors.
This completes the proof of the theorem.

REFERENCES

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

[2] S. Ginsburg and E.H. Spanier, Bounded regular sets, Proc. Amer. Math. Soc. 17 (1966), 1043-1049.

[3] A. Thue, Uber die gegenseitige Lage gleicher Teile Gewisser Zeicheureihen, Skr. Vid. Kristiania I. Mat. Naturv. Klasse 1 (1912), 1-67.

[4] F. Dejean, Sur un Théorème de Thue, Journal of Combinatorial Theory (A) 13 (1972), 90-99.

[5] R.C. Lyndon and M.P. Schutzenberger, The equation $a^m=b^n c^p$ in a free group, Michigan Math. J. 9 (1962), 289-298.