COMMUTATIVE LINEAR LANGUAGES

by

A. Ehrenfeucht *

and

G. Rozenberg **

CU-CS-209-81

June 1981

* A. Ehrenfeucht
Dept. of Computer Science
University of Colorado, Boulder
Boulder, Colorado 80309

** G. Rozenberg
Institute of Applied Math. and Computer Science
University of Leiden
Leiden, The Netherlands

All correspondence to the second author.
ABSTRACT

It is proved that every commutative linear language is regular. This result follows from a more general one which provides conditions which imposed on an arbitrary language imply its regularity.
INTRODUCTION

The class of regular languages, \( L_R \), forms a very fundamental class of languages within formal language theory (see, e.g., [H] and [S]). The class of context-free languages, \( L_{CF} \), is an important class of languages containing \( L_R \). In order to better understand the structure of languages in \( L_{CF} \) various attempts have been made to provide conditions which imposed on a language in \( L_{CF} \) will "force it" to be regular. Such conditions can be grammatical, that is they are conditions which imposed on a context free grammar imply that its language is regular ("right-linearity" and "non-self-embedding" are examples of such conditions).

Much less is known about conditions which imposed on (the structure of words in) a context-free language will imply that the language is regular, see, e.g., [ABBL]. In an effort to learn more about such conditions one may investigate subclasses of \( L_{CF} \) which are "as small as possible" (and still contain \( L_R \)). A class of languages "very close" to \( L_R \) is the class of linear languages, \( L_{LIN} \). Since linear grammars differ from right-linear grammars only by the fact that the unique nonterminal in a sentential form may generate terminal symbols both to the right and to the left of itself, it looks very plausible that requiring commutativity of a linear language (that is requiring that for every word each permutation of occurrences of letters in it will result in a word also in the language) will force it to be regular.

This conjecture was formulated in [L] which considers various properties of commutative context-free languages. In our paper we demonstrate that this conjecture holds.
0. PRELIMINARIES

We assume the reader to be familiar with the basic theory of context-free languages; in particular with the basic theory of regular and linear languages, see, e.g., [5]. We use mostly standard language theoretic terminology and notation. Perhaps the following points require an additional explanation.

We use \( \mathbb{N} \) to denote the set of nonnegative integers and \( \mathbb{N}^+ \) to denote the set of positive integers. For \( n \in \mathbb{N}^+ \), \( \mathbb{N}^n \) denotes the \( n \)-folded cartesian product of \( \mathbb{N} \). If \( v \in \mathbb{N}^n \) then, for \( 1 \leq i \leq n \), \( v(i) \) denotes the \( i \)-th component of \( v \). If \( v_1, v_2 \in \mathbb{N}^n \) then \( v_1 \preceq v_2 \) if and only if \( v_1(i) \leq v_2(i) \) for each \( 1 \leq i \leq n \).

For a finite set \( Z \), \( \#Z \) denotes its cardinality. For sets \( Z_1, Z_2 \), \( Z_1 - Z_2 \) denotes the set-theoretic difference of \( Z_1 \) and \( Z_2 \).

In the sequel of this paper we consider an arbitrary but fixed alphabet \( \Sigma = \{a_1, \ldots, a_d\} \) where \( d \geq 1 \), and so all languages we consider are over \( \Sigma \).

For a word \( w \), \( \text{alph}(w) \) denotes the set of all letters that occur in \( w \).
For a letter \( a \) and a word \( w \), \( \#_a(w) \) denotes the number of occurrences of \( a \) in \( w \).

Let \( \psi : \Sigma^* \rightarrow \mathbb{N}^d \) be the mapping defined by:
for \( w \in \Sigma^* \), \( \psi(w) = (\#_{a_1}(w), \ldots, \#_{a_d}(w)) \); \( \psi \) is referred to as the Parikh mapping and \( \psi(w) \) as the Parikh vector of \( w \). For \( K \subseteq \Sigma^* \), \( \psi(K) = \bigcup_{w \in K} \psi(w) \).

In this paper we deal with commutative languages. They are defined as follows.
Definition. (i). Let \( w \in \Sigma^* \). The \textit{commutative closure of} \( w \), denoted \( \text{com}(w) \), is defined by \( \text{com}(w) = \{ x \in \Sigma^* : \Psi(x) = \Psi(w) \} \). (ii). A language \( K \) is \textit{commutative} if \( \text{com}(w) \subseteq K \) for each \( w \in K \). (iii). Let \( X \subseteq \Psi(\Sigma^*) \). The \textit{language of} \( X \), denoted \( L(X) \), is defined by \( L(X) = \{ w \in \Sigma^* : \Psi(w) \in X \} \). □

The following result is a direct consequence of the above definition.

Lemma 0.1. (i). Let \( K_1, K_2 \) be commutative languages. \( K_1 \subseteq K_2 \) if and only if \( \Psi(K_1) \subseteq \Psi(K_2) \). (ii). Let \( X \subseteq \Psi(\Sigma^*) \). Then \( L(X) \) is uniquely defined. □

The following result from [La] (somewhat reformulated so that it is suited for our application) will be useful in the sequel.

Proposition 0.1. Let \( X \subseteq \Psi(\Sigma^*) \). There exists a finite set \( F \subseteq X \) such that for every \( v \in X \) there exists a \( u \in F \) such that \( u \leq v \). □
1. PERIODIC LANGUAGES

In this section periodic languages are introduced and investigated. They form a subclass of the class of commutative languages.

Definition. Let \( \rho = v_0, v_1, \ldots, v_d \) be a sequence of vectors from \( \mathbb{N}^d \). We say that \( \rho \) is a base if and only if \( v_i(j) = 0 \) for all \( i, j \geq 1 \) such that \( i \neq j \). We use first(\( \rho \)) to denote \( v_0 \). The \( \rho \)-set, denoted \( \Theta(\rho) \), is defined by \( \Theta(\rho) = \{ v \in \Psi(\Sigma^*) : v = v_0 + \ell_1 v_1 + \ldots + \ell_d v_d \text{ for some } \ell_1, \ldots, \ell_d \in \mathbb{N} \} \).

Note that the \( \rho \)-set is a linear set (see, e.g., [S]). It is easy to see that each base is unique in the following sense.

Lemma 1.1. If \( \rho, \rho' \) are bases such that \( \Theta(\rho) = \Theta(\rho') \) then \( \rho = \rho' \).

Definition. Let \( X \in \Psi(\Sigma^*) \). We say that \( X \) is periodic if and only if there exists a base \( \rho \) such that \( X = \Theta(\rho) \).

In view of Lemma 1.1 for each periodic \( X \in \Psi(\Sigma^*) \) there exists a unique base \( \rho \) such that \( X = \Theta(\rho) \); we say that \( \rho \) is the base of \( X \) and we write \( \rho = \text{base}(X) \).

Definition. A language \( K \) is periodic if and only if \( K \) is commutative and \( \Psi(K) \) is periodic. If \( K \) is periodic then the base of \( \Psi(K) \) is referred to as the base of \( K \), denoted \( \text{base}(K) \).

The following parameters of periodic languages will be considered in the sequel.

Definition. Let \( K \) be a periodic language where \( \text{base}(K) = v_0, v_1, \ldots, v_d \).

(i). The type of \( K \), denoted \( \text{type}(K) \), is the pair of vectors \( (u_1, u_2) \) from \( \mathbb{N}^d \) defined as follows:
\[ u_1 = (v_0(1) \mod v_1(1)), \ldots, v_0(i) \mod v_1(i)), \ldots, v_0(d) \mod v_d(d)) \] and
\[ u_2 = (v_1(1), \ldots, v_1(i), \ldots, v_d(d)). \]

(ii) The size of \( K \), denoted size(\( K \)), is defined by:
\[
\text{size}(K) = \max_{1 \leq i \leq d} \{\max\{u_1(i), u_2(i)\}\} \quad \text{where} \quad \text{type}(K) = (u_1, u_2). \quad \square
\]

Example. Let \( \Sigma = \{a_1, a_2, a_3, a_4\} \) and let \( K \) be the periodic language such that \( \text{base}(K) = (1, 6, 8, 0), (2, 0, 0, 0), (0, 3, 0, 0), (0, 0, 0, 0), (0, 0, 0, 7) \). Then \( \text{type}(K) = (u_1, u_2) \) where \( u_1 = (1, 0, 8, 0) \) and \( u_2 = (2, 3, 0, 7) \);
\[
\text{size}(K) = \max\{2, 3, 8, 7\} = 8. \quad \square
\]

The following result is very basic for periodic languages.

**Theorem 1.1.** Every periodic language is regular.

**Proof.**

Let \( K \) be a periodic language and let \( \text{base}(K) = v_0, v_1, \ldots, v_d \). Clearly a word \( w \in \Sigma^* \) is in \( K \) if and only if, for every \( i \in \{1, \ldots, d\} \),
\[
\#_{a_i}(w) \geq v_0(i) \quad \text{and} \quad \#_{a_i}(w) = v_0(i) \mod v_1(i)) \quad \text{.........................(1)}
\]
Consequently \( K = K_1 \cap \ldots \cap K_d \) where \( K_i = \{w \in \Sigma^* : (1) \text{ holds}\} \) for \( 1 \leq i \leq d \).

It is easily seen that each \( K_i, 1 \leq i \leq d \), is regular and so \( K \) is regular. \( \square \)

Next we will provide conditions which imposed on an arbitrary language will force it to be a finite union of periodic languages.

**Lemma 1.2.** Let \( K_1, K_2 \) be periodic languages such that \( \text{type}(K_1) = \text{type}(K_2) \).
If \( \text{first}(\text{base}(K_1)) \leq \text{first}(\text{base}(K_2)) \) then \( K_2 \subseteq K_1 \).

**Proof.**

Obvious. \( \square \)

**Lemma 1.3.** Let \( F \) be a family of periodic languages such that all languages in \( F \) are of the same type. There exists a finite family of languages \( L \subseteq F \) such that \( \bigcup_{K \in F} K = \bigcup_{K \in L} K \).
Proof.

Let \( X_F \subseteq \psi(\Sigma^*) \) be defined by \( X_F = \{ v : v = \text{first}(\text{base}(K)) \text{ for some } K \in F \} \). By Proposition 0.1, \( X_F \) contains a finite set of vectors \( \{ z_1, \ldots, z_\ell \} \), \( \ell \geq 1 \), such that for each \( v \in X_F \), \( z_j \leq v \) for some \( j \in \{1, \ldots, \ell\} \).

(2) Now let, for each \( j \in \{1, \ldots, \ell\} \), \( K_j \) be a language from \( F \) such that \( u_j = \text{first}(\text{base}(K_j)) \) and let \( L = \{ K_1, \ldots, K_\ell \} \). Then the result follows from (2) and from Lemma 1.2. \( \square \)

Lemma 1.4. Let \( F \) be a family of periodic languages such that there exists a \( q \in \mathbb{N}^+ \) such that \( \text{size}(K) \leq q \) for each \( K \in F \). Then there exists a finite family of languages \( L \subseteq F \) such that \( \bigcup_{K \in F} K = \bigcup_{K \in L} K \).

Proof.

Let \( F \) satisfy assumptions of the lemma. Since \( \text{size}(K) \leq q \) for each \( K \in F \), the number of different types of languages in \( F \) is finite. Consequently there exists a positive integer \( r \) such that \( F = F_1 \cup \ldots \cup F_r \) where, for each \( i \leq j \leq r \), all languages in \( F_j \) are of the same type. Hence the result follows from Lemma 1.3. \( \square \)

Theorem 1.2. Let \( K \) be a language. If there exists a \( q \in \mathbb{N}^+ \) such that for each \( w \in K \) there exists a periodic language \( L_w \subseteq K \) where \( w \in L_w \) and \( \text{size}(L_w) \leq q \) then \( K \) is a finite union of periodic languages.

Proof.

Assume that \( K \) satisfies the assumptions of the theorem. Then \( K = \bigcup_{w \in K} L_w \) where the family \( F = \{ L_w : w \in K \} \) satisfies the assumptions of Lemma 1.4. Thus the theorem follows from Lemma 1.4. \( \square \)
Corollary 1.1. Let $K$ be a language. If there exists a $q \in \mathbb{N}^+$ such that for each $w \in K$ there exists a periodic language $L_w \subseteq K$ where $w \in L_w$ and $\text{size}(L_w) \leq q$ then $K$ is regular.

Proof.

The corollary follows directly from Theorems 1.1 and 1.2. ☐
2. COMMUTATIVE LINEAR LANGUAGES

In this section we will consider commutative linear languages. In particular we will provide their representation through periodic languages.

Theorem 2.1. A language $K$ is a commutative linear language if and only if $K$ is a finite union of periodic languages.

Proof.

Assume that $K$ is a finite union of periodic languages. Then, by Theorem 1.1, $K$ is a commutative regular language and so a commutative linear language.

To prove that a commutative linear language is a finite union of periodic languages we proceed as follows.

Let $K$ be a commutative linear language and let $G = (\Omega, \Sigma, P, S)$ be a linear grammar generating $K$, so that $L(G) = K$. Clearly we can assume that each production of $G$ is in one of the following three forms:

$A \rightarrow Ba$, $A \rightarrow aB$ and $A \rightarrow a$ where $A$, $B$ are nonterminals ($A, B \in \Omega - \Sigma$) and $a$ is a terminal ($a \in \Sigma$).

By Theorem 1.2 it suffices to prove the following result.

Lemma 2.1. There exists a $q \in \mathbb{N}^+$ such that for every $w \in K$ there exists a periodic language $L_w \subseteq K$ where $w \in L_w$ and $size(L_w) \leq q$.

Proof of Lemma 2.1.

Let $m = \#\Omega$. We define the sequence $\{q_i\}_{i \geq 1}$ of positive integers as follows:

$q_1 = m + 1$ and $q_{i+1} = (q_1 + \ldots + q_i + 1)(m+1)$ for $i \geq 1$.

Then we set $q = 2q_{m+1}$

Let $w \in K$. Let $\rho = v_0, v_1, \ldots, v_d$ be the base defined as follows.

$v_0 = \psi(w)$.

If $1 \leq i \leq d$ is such that $v_0(i) \leq q$ then $v_i(i) = 0$. 


If for every $i \in \{1, \ldots, d\}$, $v_0(i) \leq q$ then all components of $\rho$ are defined and we are done. Otherwise we proceed as follows.

Let $\{b_1, \ldots, b_s\}$ be all the letters from $\text{alph}(w)$ such that $\#_{b_j}^q(w) > q$ for $1 \leq j \leq s$.

Now let $w' = b_1^{q_1} \ldots b_s^{q_s} u b_s^{q_s} \ldots b_1^{q_1}$ where $u$ is a fixed word such that $b_1^{q_1} \ldots b_s^{q_s} u b_s^{q_s} \ldots b_1^{q_1} \in \text{com}(w)$. Since $q = 2q_m$, $w'$ is well defined.

For $1 \leq i \leq s$ we refer to the leftmost occurrence of $b_i^{q_i}$ in $w'$ as the left $i$-block and to the rightmost occurrence of $b_i^{q_i}$ in $w'$ as the right $i$-block; the left $i$-block together with the right $i$-block form the $i$-block of $w'$.

Consider a derivation tree $D$ of $w$ in $G$; the path of $D$ originating in its root and ending on a leaf of $D$ such that the direct ancestor of the last node (the leaf) has one descendant only is called the spine of $D$ and denoted $\tau$. A sequence of consecutive nodes of $\tau$ is called a segment (of $\tau$). The label of a node $e$ of $\tau$ is denoted by $\ell(e)$. If $\rho = e_1 \ldots e_k e_{k+1}$ is a segment of $\tau$ such that $k \geq 1$, $e_1, \ldots, e_{k+1}$ are nodes of $\tau$, $\ell(e_1) = \ell(e_{k+1})$ and $\ell(e_j) \neq \ell(e_1)$ for $2 \leq j \leq k$ then $\rho$ is called a repeat (of $\tau$); $e_1 \ldots e_k$ is the front of $\rho$ (denoted $\text{front}(\rho)$). The contribution of a segment $\mu$ of $\tau$ are the occurrences in $w'$ which are "derived" from nodes of $\mu$ (in other words, those occurrences in $w'$ which have ancestors among the nodes of $\mu$).

The following technical result is very crucial to our proof of Lemma 2.1.

**Claim 2.1.** For every $1 \leq i \leq s$ there exists a repeat $\mu$ on $\tau$ such that the contribution of $\text{front}(\mu)$ is contained in the $i$-block of $w'$.

**Proof of Claim 2.1.**

The proof goes by induction on $i$, $1 \leq i \leq s$.

Let $i = 1$.

Consider the segment of $\tau$ consisting of its first $(m+1)$ nodes.
Since \( q_1 = m + 1 \) it is clear that this segment contributes only to the first block of \( w' \). On the other hand, the length of this segment is \((m+1)\) and so it must contain a repeat. Hence the claim holds for \( i = 1 \).

Assume that the claim holds up to the \((i-1)\)-block where \( 2 \leq i \leq s \). We will demonstrate now that it holds for the \( i \)-block of \( w' \).

Let \( U \) be the rightmost occurrence of \( b_{i-1} \) in the left \((i-1)\)-block of \( w' \) and let \( T \) be the leftmost occurrence of \( b_{i-1} \) in the right \((i-1)\)-block of \( w' \). Let \( O_U \) be the ancestor of \( U \) on \( \tau \) and let \( O_T \) be the ancestor of \( T \) on \( \tau \).

Thus we have the following situation (we have assumed that \( O_U \) is closer to the root than \( O_T \); clearly we can assume it without loss of generality).
Clearly all nodes above \( Q_U \) contribute either to the left of \( U \) or to the right of \( T \). Now let \( Q_1, \ldots, Q_\ell \) be all the nodes strictly between \( Q_U \) and \( Q_T \) such that they contribute to the right of \( T \).

Since \( \begin{vmatrix} q_1 & q_2 & \cdots & q_{i-2} & q_{i-1} \\ b_1 & b_2 & \cdots & b_{i-2} & b_{i-1} \end{vmatrix} = q_1 + \cdots + q_{i-1} \), clearly we have \( \ell + 1 \leq q_1 + \cdots + q_{i-1} \) \((3)\).

Now let \( z_1, \ldots, z_\ell, z_{\ell+1} \) be segments of \( \tau \) defined as follows:

- \( z_1 \) consists of all the nodes strictly between \( Q_U \) and \( Q_1 \),
- \( z_2 \) consists of all the nodes strictly between \( Q_1 \) and \( Q_2 \),
- \( \vdots \)
- \( z_\ell \) consists of all the nodes strictly between \( Q_{\ell-1} \) and \( Q_\ell \),
- \( z_{\ell+1} \) consists of all the nodes strictly between \( Q_\ell \) and \( Q_T \).

We consider now separately two cases.

Case 1. At least one of the segments \( z_1, \ldots, z_\ell \) consists of more than \( m \) nodes.

Let \( i_0 \) be the smallest index \( j \) such that \( z_j \) consists of more than \( m \) nodes.

In \( z_{i_0} \), we consider the segment \( \gamma \) consisting of the first \((m+1)\) nodes.

Clearly, this segment contains a repeat; say \( w \). Note that all the nodes from \( z_1, z_2, \ldots, z_{i-1}, z_{i_0} \), \( \gamma \) contribute to the right of \( U \) (but to the left of \( T \)).

The number of occurrences contributed to \( w \) by all the nodes from \( z_1, \ldots, z_{i_0}, \gamma \) is not greater than \((\ell+1)(m+1)\) and so by \((3)\) it is not greater than \((q_1 + \cdots + q_{i-1} + 1)(m+1)\). Since the length of the left and the right \( i \)-block equals \( q_i \), this means that all occurrences contributed by nodes from \( z_1, \ldots, z_{i-1}, \gamma \) are within the \( i \)-block.

Thus in this case the claim holds for the \( i \)'th block.
Case 2. Each of the segments $z_1, \ldots, z_{\ell+1}$ consists of no more than $m$ nodes.

Clearly in this case the number of occurrences contributed to $w'$ by all the nodes from $z_1, \ldots, z_{\ell+1}$ does not exceed $(\ell+1)m$ and (because the length of the left and right $i$-block is $q_i$) all of these occurrences are within the $i$-block. Moreover, from (3) and from the definition of $q_i$ it follows that if we consider the segment $\rho$ of $\tau$ consisting of $(m+1)$ nodes immediately following $0_{\tau}$ then all the nodes from $\rho$ will contribute to the $i$-block of $w'$. But $\rho$ must contain a repeat and so also in this case the claim holds for the $i$'th block.

Hence we have completed the induction and the claim holds. \(\square\)

Now that the claim is proved we complete the definition of $\rho$ as follows.

Let for each $i \in \{1, \ldots, s\}$, $k(b_i)$ be the length of the front of a repeat $u$ on $\tau$ which satisfies the statement of Claim 2.1 and has the shortest length. If $b_i = a_j$ for $1 \leq j \leq d$, then we set $v_j(j) = k(b_i)$. Thus $\rho$ is now completely defined; $\rho = v_0, v_1, \ldots, v_d$.

We set $L_{w'} = L(\Theta(\rho))$. In order to show that $L_{w'} \subseteq K$ it suffices to show (see Lemma 0.1) that $\Theta(\rho) \subseteq \Psi(K)$.

Let $v \in \Theta(\rho)$, hence $v = v_0 + \ell_1 v_1 + \ldots + \ell_d v_d$ where $\ell_1, \ldots, \ell_d \in \mathbb{N}$.

If $v_i(i) = 0$ for $1 \leq i \leq d$ then in the derivation tree $D$ of $w'$ (from the proof of the above claim) we will "iterate" $\ell_i$ times a repeat of the length $k(a_i)$ contributing to the $i$-block (and we do it for each $i$ satisfying $v(i) \neq 0$). In this way we get the word $w'(\ell_1, \ldots, \ell_d)$ such that
\[
\Psi(w'(\ell_1, \ldots, \ell_d)) = v. \text{ Thus } v \in \Psi(K).
\]

Consequently $\Theta(\rho) \subseteq \Psi(K)$ and so $L_{w'} \subseteq K$. Clearly $\text{size}(L_{w'}) \leq q$. Finally we notice that $w \in L_{w'}$ (because $w' \in \text{com}(w)$) and so if we set $L_w = L_{w'}$, the lemma holds. \(\square\)
But Lemma 2.1 together with Theorem 1.2 proves the "only if" part of the theorem.

Consequently the theorem holds. □

The following corollary of Theorem 2.1 solves an open problem from [L].

Corollary 2.1. If $K$ is a commutative linear language then $K$ is regular.

Proof.
Directly from Theorems 2.1 and 1.1. □

Also, directly from Theorem 2.1 we get the following result.

Corollary 2.2. A language is commutative and regular if and only if it is a finite union of periodic languages. □
REFERENCES


ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NSF grant MCS 79-03838. They are indebted to J. Kleijn and R. Verraedt for comments concerning the first draft of this paper.