# PARIKH-BOUNDED LANGUAGES*

Meera Blattner
University of California, Davis

and

Lawrence Livermore National Laboratory
Livermore, California

and

Michel Latteux
Université de Lille
Villeneuve d'Ascq, France

## ABSTRACT

A string y is in C(x), the commutative image of a string x, if y is a permutation of the symbols in x. A language L is Parikh-bounded if L contains a bounded language B and all x in L have a corresponding y in B such that x is in C(y). The central result in this paper is that if L is context-free it is also Parikh-bounded. Parikh's theorem follows as a corollary. If L is not bounded but is a Parikh-bounded language closed under intersection with regular sets, then for any positive integer k there is an x in L such that $\#(C(x) \cap L) \geq k$. The notion of Parikh-discreteness is introduced.

*This research was supported in part by NSF Grant No. MCS 77-02470.

## I. INTRODUCTION

A language L is a collection of finite length strings over a finite alphabet $\Sigma$. The commutative image C(L) of L is the set of strings y such that y is a permutation of the symbols of some x in L. A language is commutative if L=C(L). Commutative languages arise naturally mathematically and share many properties with those of bounded languages [GS].

A string x is letter-equivalent or Parikh-equivalent to y if x and y have the same number of occurrences of each symbol. If x and y are letter-equivalent then C(x) = C(y).

Section III of this paper contains the main theorem: every context-free language L has associated with it a bounded set $B \subseteq w_1{}^*w_2{}^*\ldots w_n{}^*$, and $B \subseteq L$, for some $n \geq 1$, and each x in L has a y in B such that x and y are letter-equivalent.

The $w_1$, $w_2$, ..., $w_n$ may be regarded as basic words or building blocks of L. Put differently, the theorem states that each string in L has a rearranged counterpart in $\mathcal{B}$. This property of families of languages is called Parikh-boundedness by Latteux and Leguy [LL]. Another proof of Parikh's Theorem, that is, each context-free language is letter-equivalent to a regular set [P], is a corollary of the main theorem.

In the Section IV examples of noncontext-free languages are given that do not have the Parikh-bounded property. More specifically, the nonerasing stack languages and the ETOL do not have the property.

Let $\mathcal{L}$ be a family of languages with the Parikh-bounded property and closed under intersection with regular sets. Then for each nonbounded L in $\mathcal{L}$ and integer k there is a string x in L such that x has more that k letter-equivalent strings in L.

A language L is Parikh-discrete if for all x and y in L C(x) = C(y) implies x = y. It follows from the main theorem that all Parikh-discrete context-free languages are bounded.

Commutative languages were studied by Latteux [L1], [L2], [L3] while Latteux and Leguy proved that the family GRE is Parikh-bounded [LL].

## II. PRELIMINARY DEFINITIONS

Let $\Sigma$ be a finite alphabet. A language L is a set of strings contained in $\Sigma^*$. L is context-free if L is generated by a grammar G where G = (V, $\Sigma$, P, S), V is a finite vocabulary, $\Sigma$ is the set of terminal symbols, $\Sigma \subset V$, S is the start symbol and P is finite a set set of rules $X \to \alpha$, where X is in V-$\Sigma$ and $\alpha$ is in V.*

A context-free grammar G is nonterminal bounded if there is an integer k for G so that every string generated by G has less than k nonterminals. A context-free grammar G is derivation bounded if G has an integer k and each x in L(G) has a derivation such that each string generated in the derivation has less than k nonterminals.

A context-free grammar G is expansive if there is some derivation in G where $X \overset{*}{\Rightarrow} \alpha X \beta X \gamma$, X in V-$\Sigma$, $\alpha,\beta,\gamma$ in V*, and $X \overset{*}{\Rightarrow} w$, w in $\Sigma^+$, and G is nonexpansive otherwise. It is known that if a context-free language L has a nonexpansive grammar then

L(G) is in the family of derivation bounded languages.

A language L is bounded if $L \subseteq w_1^* \ldots w_n^*$, where $w_i$ is in $\Sigma^*$, $1 \leq i \leq n$. The commutative image C(L) of L is $\{a_{i_1} a_{i_2} \ldots a_{i_n} \mid a_1 a_2 \ldots a_n$ is in L and $(i_1,\ldots,i_n)$ is a permutation of $(1,2,\ldots,n)\}$, and L is commutative if L = C(L).†

Let $\Sigma = \{a_1, a_2, \ldots, a_n\}$ and define a mapping $\psi$, called the Parikh mapping, from $\Sigma^*$ into $N^n$ by: $\psi(w) = (\#a_1(w),\ldots,\#a_n(w))$ where $\#a_i(w)$ denotes the number of occurrences of $a_i$ in w. Define $\psi(L) = \{\psi(x) \mid x \in L\}$. Languages $L_1$ and $L_2$ in $\Sigma^*$ are called letter-equivalent (or Parikh-equivalent) if $\psi(L_1) = \psi(L_2)$.

A language L is called Parikh-bounded if there is a bounded language B contained in L such that if x is in L there is a y in B so that C(x) = C(y).

A set of strings L is a semilinear set if $\psi(L)$ is the union of linear sets of the form

$$\sum_{i=1,n} C_i(n_{i1},\ldots,n_{ik}) + (d_1,\ldots,d_k) \quad , \text{ for some } k \geq 0 \quad .$$

A language L is called Parikh-discrete if for all x and y in L, C(x) = C(y) implies x = y.

## III. THE PARIKH BOUNDNESS OF CONTEXT FREE LANGUAGES

Theorem 1: (Latteux and Leguy) Greibach's family, denoted by GRE, is the least substitution closed rational cone containing the linear and one counter languages. Every language in this family is Parikh-bounded; namely, it contains a bounded language with the same commutative image.

We extend the results of Theorem 1 to those of Theorem 2. But first we must prove an intermediate result.

Lemma 1: Every context-free language contains a derivation-bounded language with the same commutative image.

Proof: For every context-free language L there is a context-free grammar $G=(V,\Sigma,P,A)$ such that L=L(G) and G and the properties:

i) $P \subseteq N \times (N^2 \cup \Sigma^* N \Sigma^* \cup \{e\})$, $N = V-\Sigma$.

ii) For all $B \to u$ in P, if $u = vCw$, v,w in $V^*$, then $B \neq C$.

Construct a new grammar $G' = (V',\Sigma,P',A)$ as follows:

$N_i = \{B_i \mid B \in N\}$, i in $\{1,2\}$

---

†This definition is not equivalent to that in Harrison [H].

$N' = N \cup N_1 \cup (N_2 - \{A_2\})$
$V' = N' \cup \Sigma$

Let $h_1$ and $h_2$ be homomorphisms from V to V' where $h_1(a) = h_2(a) = a$, if $a \varepsilon \Sigma$, and $h_1(B) = B_1$, $h_2(B) = B_2$ if $B \varepsilon N$. Let:

$P_0 = \{B \to C_1 D,\ B \to CD_1 | B \to CD \text{ in } P \cap N \times N^2\} \cup (P - N \times N^2)$
$P_1 = \{B_1 \to h_1(u) | B \to u \text{ in } P,\ B \neq A\}$
$P_2 = \{B_2 \to h_2(u) | B \to u \text{ in } P,\ B \neq A,\ u \neq vAw \text{ for any } v, w \text{ in } V^*\}$
$P_3 = \{A_1 \to h_2(u) | A \to u \text{ in } P\}$
$P' = P_0 \cup P_1 \cup P_2 \cup P_3$

Now we define another homomorphism $\pi$ which maps the symbols of V' back to V, so $\pi(z) = z$, if $z \varepsilon V^+$, $\pi(B_1) = \pi(B_2) = B$ if $B_i$ in $V' - V$, i in $\{1,2\}$.

In order to prove the lemma we need to establish two claims.

<u>Claim 1</u>: $L(G') \subseteq L(G)$.

<u>Proof</u>: For each derivation in G', $A \Rightarrow s_1 \Rightarrow s_2 \Rightarrow \ldots \Rightarrow w$, we have $\pi(A) \Rightarrow \pi(s_1) \Rightarrow \ldots \Rightarrow \pi(w) = w$ in G.

<u>Claim 2</u>: For each w in L(G) there is a w' in L(G') such that $C(w') = C(w)$.

<u>Proof</u>: The proof is by induction on the length of the derivation. Hence, we shall show that if $A \overset{*}{\underset{G}{\Rightarrow}} x$, x in $V^*$, then there is an x' in $(V')^*$ so that $A \underset{G'}{\Rightarrow} x'$ and $C(\pi(x')) = C(x)$. The result follows when x is a terminal string.

If n = 0 then $A \overset{0}{\underset{G}{\Rightarrow}} A$ and $A \overset{0}{\underset{G'}{\Rightarrow}} A$ so the result holds.

Assume that $A \overset{m}{\underset{G}{\Rightarrow}} x$, for all m<n, implies $A \overset{*}{\underset{G'}{\Rightarrow}} x'$ and $C(\pi(x')) = C(x)$.

If $A \overset{n-1}{\underset{G}{\Rightarrow}} uBv \Rightarrow usv$ then we know there is a derivation: $A \overset{*}{\underset{G'}{\Rightarrow}} u'B'v'$ where B' in $\{B, B_1, B_2\}$ and $C(\pi(u'v')) = C(uv)$, for u, v in $V^*$ and u', v' in $(V')^*$.

<u>Case 1</u>: $B' \neq B_2$.

In this case there is a rule $B' \to s'$ and $\pi(s') = s$, so the result holds.

<u>Case 2</u>: $B' = B_2$ and $s \neq s_1 A s_2$.

The same conclusion may be drawn as in Case 1.

<u>Case 3</u>: $B' = B_2$ and $s = s_1 A s_2$.

If $A \overset{*}{\underset{G'}{\Rightarrow}} u'B_2v'$ then $A \overset{*}{\underset{G'}{\Rightarrow}} yA_1z$ and $A_1 \overset{*}{\underset{G'}{\Rightarrow}} y'B_2z'$ with $yy' = u'$ and $z'z = v'$. If we knew that $A \overset{*}{\underset{G'}{\Rightarrow}} h_1 o \pi(y') B h_1 o \pi(z')$ then we would obtain the desired derivation in G' because:

$$A \overset{*}{\underset{G'}{\Rightarrow}} h_1 o \pi(y') B h_1 o \pi(z') \underset{G'}{\Rightarrow} h_1(\pi(y')s_1) A h_1(s_2 \pi(z')) \overset{*}{\underset{G'}{\Rightarrow}} h_1(\pi(y')s_1)\ yA_1 z h_1(s_2\pi(z')) = x'.$$

Our objective was to show that if $A \overset{n}{\underset{G}{\Rightarrow}} us_1As_2v = x$ then there is a x' such that $C(x) = C(\pi(x'))$ and the x' above satisfies that condition since $\pi(y'y) = u$ and $\pi(zz') = v$.

It remains only to show that if $A_1 \overset{*}{\underset{G'}{\Rightarrow}} y'B_2z'$ then $A \overset{*}{\underset{G'}{\Rightarrow}} h_1 o\pi(y')Bh_1 o\pi(z')$. This will be done by induction on the length of the derivation. If the length is one the result follows. Otherwise we have:

$A_1 \overset{+}{\underset{G'}{\Rightarrow}} y''C_2z'' \underset{G'}{\Rightarrow} y''sz'' = y'B_2z'$. There are again three cases:

Case 1: $y'' = y'B_2y_1''$ and so $z' = y_1''sz''$. $A_1 \overset{+}{\underset{G'}{\Rightarrow}} y'B_2y_1''C_2z''$ and the induction hypothesis implies $A \overset{*}{\underset{G'}{\Rightarrow}} \pi_1(y')B\pi_1(y_1''C_2z'') = \pi_1(y')B\pi_1(y_1'')C_1\pi_1(z'')$ with $\pi_1 = h_1 o\pi$. As $C_2 \to s$ in $P_2$, $C_1 \to \pi_1(s)$ in $P_1$ we have $A \overset{*}{\underset{G'}{\Rightarrow}} \pi_1(y')B\pi_1(y''sz'') = \pi_1(y')B\pi_1(z')$.

Case 2: $z'' = z_1''B_2z'$ by the same reasoning.

Case 3: $s = t_1B_2t_2$ with $y' = y''t_1$, $z' = t_2z''$ and $C_2 \to t_1B_2t_2$. By the inductive hypothesis we have $A \overset{*}{\underset{G}{\Rightarrow}} \pi_1(y'')C\pi_1(z'')$ and clearly $C \to \pi_1(t_1)B\pi_1(t_2)$ in $P_0$ so $A \overset{*}{\underset{G}{\Rightarrow}} \pi_1(y'')\pi_1(t_1)B\pi_1(t_2)\pi_1(z'') = \pi_1(y')B\pi_1(z')$. Now $\pi$ is the identity on $\Sigma^*$ so from claim one and claim two we know that $C(L) = C(L')$.

Let $P_t = \{X \to w | w \in \Sigma^*\} \cap P$ and $G_0$ be the linear grammar $G_0 = (V \cup N_1, \Sigma \cup N_1, P_t, A)$ then $L_0 = L(G_0)$ is a linear language and $L'$ is obtained from $L_0$ by substituting for each $B_1$ in $\Sigma \cup N_1$ the set derived from $B_1$ in $G'$. Let $G_1 = (V_1, \Sigma, P_1 \cup P_2 \cup P_3, B_1)$ where $V_1 = \Sigma \cup N_1 \cup (N_2-\{A_2\})$, for each $B_1$ in $N_1$, and t be the substitution $t(a) = a$, for all a in $\Sigma$, and $t(B_1) = L(G_1)$ for all $B_1$ in $N_1$, then $t(L_0) = L'$.

The proof that $L = L(G)$ contains a derivation bounded language with the same commutative image will be made by induction on the number of nonterminals. We assume that G has properties i) and ii). If $N = 1$ then by property ii) the productions are all of the type $A \to w$, $w \in \Sigma^*$, hence $L(G)$ is finite.

Assume that G has n+1 nonterminals. If $L'$ contains a derivation bounded language with the same commutative image then we are finished since $C(L) = C(L')$ and $L' \subseteq L$. Since $L'$ is $t(L_0)$ it suffices to show the property for each $G_1$ since the family of derivation bounded languages is closed by substitution.

Consider the grammar $G_1' = (V_1', \Sigma \cup \{A_1\}, P_1, B_1)$, $V_1' = N_1 \cup \Sigma$. This grammar has properties i) and ii) and $\#(N_1-\{A_1\}) = \#(N)-1$. Now one can use the induction hypothesis and so for each $B_1$ in $N_1-\{A_1\}$ there is a language $L_{B_1} \subseteq L(G_1')$, $L_{B_1}$ derivation bounded, and $C(L_{B_1}) = C(L(G_1'))$. Now observe that $L(G_1)$ is obtained from $L(G_1')$ by replacing each $A_1$ by $L(G_2)$ where $G_2 = (V_2, \Sigma, P_2 \cup P_3, A_1)$ and $V_2 = \Sigma \cup \{A_1\} \cup (N_2-\{A_2\})$ so to finish the proof we must show that $L(G_2)$ contains a derivation bounded language with the same commutative image. But it suffices to show this for $L(G_2')$, where $G_2' = (\Sigma \cup N_2-\{A_2\}, \Sigma, P_2, B_2)$ for each $B_2$ in $N_2-\{A_2\}$. Properties i) and ii) hold for $G_2'$ and $\#(N_2-\{A_2\}) = \#(N)-1$, so the inductive hypothesis is applied and the proof is finished.

Every derivation bounded language is in GRE [G] so Lemma 1 and Theorem 1 imply:

Theorem 2: The context-free languages are Parikh bounded.

Since it is easy to show directly that the derivation bounded languages are semilinear, Lemma 1 implies:

Corollary: (Parikh's Theorem) Every context-free language is semilinear.

## IV. NONCONTEXT-FREE LANGUAGES AND THE PARIKH BOUNDED PROPERTY

Proposition: The nonerasing counter stack languages are not Parikh-bounded.

Observe that $L_1 = \{1010^2 1 \ldots 10^h 1 | h \geq 1\}$ is a nonerasing counter language. One need not go far from the context-free to find examples of languages that are not Parikh-bounded. Since $L_1$ is not semilinear the natural conjecture arises as to whether all languages that are semilinear are Parikh-bounded.

Proposition: $L_2 = L_1 \cup \{1^i 0^j | C(1^i 0^j) \cap L_1 = \emptyset\}$ is semilinear but not Parikh-bounded.

Since $\psi(L_1) \subseteq \psi\{1^i 0^j | i<j\}$ and $\psi\{1^i 0^j | i<j\}$ is semilinear the result follows. However $L_2 \cap 1^*0^*$ is not semilinear even though $\psi(L_2) = \psi(1^*0^*)$. The authors were unable to find an example of a family of languages that were semilinar under closure with regular sets that were not Parikh-bounded.

Proposition: The OL, EOL, ETOL, DOL, EDOL are not Parikh bounded.

An example proves this result. $P = \{2 \rightarrow 201,\ 1 \rightarrow 01,\ 0 \rightarrow 0\}$. Let $G = (\{2,0,1\},P,2)$ then $L(G)$ is in all of the above.

A simple observation that follows from Theorem 2 is:

Proposition: Let $L_1$ and $L_2$ be languages such that:

1. $L_1 \subseteq L_2$
2. $C(L_1) = C(L_2)$
3. $L_1$ is context-free

Then $L_2$ is Parikh-bounded.

## V. PARIKH-DISCRETENESS

In this section we examine the number of times strings with the same commutative images may occur in a language.

A language L is Parikh-discrete if for all x and y in L, $C(x) = C(y)$ implies $x = y$.

Within the context-free the Parikh-discrete languages must be bounded languages by Theorem 2.

Proposition: If L is Parikh-discrete and context-free then L is bounded.

If L is not bounded (and not necessarily context-free) then what can we say about the number of occurrences of strings in L with the same commutative image? Surprisingly enough, if $\mathcal{L}$ is a Parikh-bounded family closed under intersection with regular sets then for each L in $\mathcal{L}$ which is not bounded we may not put a limit on the number of occurrences of strings with the same commutative image. That is, for any integer k there are strings $x_1, x_2, \ldots, x_n$ in L, $x_i \neq x_j$, if $i \neq j$, and $C(x_1) = C(x_2) = \ldots = C(x_n)$, $n \geq k$.

Theorem 3: Let $\mathcal{L}$ be a family of Parikh-bounded languages closed under intersection with regular sets. If L is not bounded and L is in $\mathcal{L}$, then for all $k > 1$, there is a string w in L such that $\#(C(w) \cap L) > k$.

Proof: By induction on k. Trivial for $k = 1$. Assume that the result holds for all $k' < k$. By the definition of Parikh-boundedness we know that $\psi(L) = \psi(L \cap w_1^* \ldots w_n^*)$ for some bounded set $w_1^* \ldots w_n^*$. Now let use consider $L' = L - w_1^* \ldots w_n^* = L \cap \overline{w_1^* \ldots w_n^*}$. We know L' is in $\mathcal{L}$ and L' is not bounded (or $L \subsetneq L' \cup w_1^* \ldots w_n^*$ would be bounded). So for all w in L we know $\#(C(w) \cap L') < \#(C(w) \cap L)$. By the induction hypothesis the theorem is proved.

## VI. SUMMARY AND FUTURE RESEARCH DIRECTIONS

The authors are continuing to examine the properties of Parikh-boundedness and Parikh-discreteness. Parikh-boundedness for some other subfamilies of the context-sensitive languages is known.

The Parikh-discrete languages may also be subdivided into those that are a-discrete. A Parikh discrete language L is a-discrete if for all x and y in L if the number of a's in x is equal to the number of a's in y implies $x = y$. As an example, $\{a^n b^n | n \geq 1\}$ is a-discrete while $\{a^i b^j | i < j\}$ is not. Parsing a-discrete languages may be very efficient since only the occurrences of a's are required to discriminate between strings in the language.

The authors are examining the discomposition properties of Parikh-discrete languages are well as those operations that preserve Parikh-discreteness and a-discreteness.

## ACKNOWLEDGMENT

This work was done while Meera Blattner was visiting the University of Paris 7. The authors would like to than Maurice Nivat and the University of Paris 7 for making this collaboration possible. The authors would also like to thank Faith Fich for her comments and suggestions.


## REFERENCES


G Greibach, S., "Chains of full AFL's," Math. Systems Theory, Vol. 4, No. 3, 1970, pp. 231-242.

GS Ginsburg, S. and Spanier, E. H., "Bounded ALGOL-like languages," Trans. of the AMS, Vol. 113, 1964, pp. 333-368.

H Harrison, M., Introduction to Formal Language Theory. Addison-Wesley, Reading, Mass. 1978.

L1 Latteux, M., "Languages Commutatifs," These Sc. Math., Lille, 1978.

L2 Latteux, M., "Mots infinis et languages commutatifs," RAIRO Informatique theorique / Theor. Comp. Science, Vol. 12, No. 3, 1978, pp. 185-192.

L3 Latteux, M., "Cones rationnels commutatifs," JCSS, Vol. 18, No. 3, June 1979, pp. 307-333.

LL Latteux, M. and Leguy, J., "Une propriete de la famille GRE." Fundamentals of Computation Theory, FCT 1979. Akademie-Verlag, Berlin 1979.

P Parikh, R. J., "On context-free languages," JACM, Vol. 13, 1966, pp. 570-581.