ON THE HEIGHT OF DERIVATION TREES*

by

K. Culik II[1]

and

H.A. Maurer[2]

Research Report CS-78-39

Department of Computer Science

University of Waterloo
Waterloo, Ontario, Canada

September 1978

1) Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

2) Institute of Information Processing
Technical University of Graz
Steyrerg. 17, A-8010
Graz, Austria

*) Research in this paper was supported by the Natural Sciences and Engineering Research Council of Canada, Grant No. A 7403, and by the Austrian Federal Ministry of Science and Research.

Abstract

In this paper we study the notion of the height of context-free languages. A CF language L is of height f(n), if for some G generating L and some constant c, each word x of L of length n can be generated by a derivation tree of height $\leq cf(n)$. We show that regular languages are of logarithmic height but that a large class of (nonregular) CF languages is not of any height f(n) essentially smaller than linear. We then show that EOL and ETOL sytems allow to decrease the height of derivation trees in certain instances and that the apparent "gap" between logarithmic and linear for CF languages disappears when considering L systems. The investigations reveal a surprisingly strong relation between the height of derivation trees and other well studied (but still partially unresolved) problems of CF language theory.

1. Introduction

One of the most natural representations of the derivation process for many generative mechanisms (such as CF grammars or EOL systems) is obtained using the notion of a derivation tree.

Derivation trees reflect in a clear manner the derivation process, properties of the generative mechanism used, and the syntactic structure of the words generated. It thus seems to be important to study the influence of properties of derivation trees on the language generated. One such attempt has been made in [2], in which paper a seemingly simple restriction of the "levels" within a tree has made it possible to generate inherently ambiguous languages unambiguously, to characterize a number of well known classes of languages in an elegant fashion, etc.

In the current paper we take another approach. We investigate CF grammars (and other generative devices) with the property that for every word of length n there exists a derivation tree of height not exceeding a certain function of n. This study is motivated by the desire to obtain grammars with as compact and balanced derivation trees as possible.

We establish that regular languages can be generated by very compact trees (at the price of introducing ambiguity) but that a large class of CF languages cannot be generated by trees lower than a linear function of the length of the word generated. Indeed, we conjecture that this class is the class of all nonregular CF languages. We further show

that more powerful generative mechanisms allow a decrease in the height of derivation trees (even without introducing ambiguity) and that for certain generative mechanisms such as e.g. DOL systems a variety of different functions arises as describing the minimal height of derivation trees.

The remainder of this paper is structured as follows: Section 2 contains some general definitions, followed by the introduction of the concept basic to this paper, the notion of height of a grammar and a language. Section 3 contains four main theorems, the second of which we consider the most important one. Its proof is given by a chain of three lemmas. In an extensive discussion after Theorem 2 and its Corollary we try to show how the notion of height of grammars ties in with a number of (still open) basic problems of language theory.

2. Preliminaries

In this section we first briefly review some concepts of language theory, mainly to introduce the notation to be used in what follows. For any concepts not explained and for further details consult some book on language theory such as [4] or [7]. In the second part of this section we present the definitions forming the basis of this paper, such as the notion of height of a grammar and a language.

A CF grammar G is a quadruple $G = (V,\Sigma,P,S)$, where V and Σ are disjoint alphabets of variables and terminals respectively, where $S \in V$ is a start symbol and $P \subseteq V \times (V \cup \Sigma)^*$ is a finite set of productions. Productions (A,z) are usually written as $A \to z$. For any words $x,y \in (V \cup \Sigma)^*$ and any production $A \to z$ we write $xAy \Rightarrow xzy$ and define $\overset{*}{\Rightarrow}$ to be the transitive and reflective closure of $\Rightarrow$. For any word $x \in (V \cup \Sigma)^*$ we define $L_x(G) = \{w \mid x \overset{*}{\Rightarrow} w, w \in \Sigma^*\}$. For $x = S$ this set $L_S(G)$ is generally written as $L(G)$ and called the language generated by G.

An EOL sytem G is a quadruple $G = (V,\Sigma,P,S)$, where V,Σ and S have the same significance as for a CF grammar but where the finite set of productions P is a subset of $(V \cup \Sigma) \times (V \cup \Sigma)^*$ such that for each $\alpha \in V \cup \Sigma$, P contains at least one production $\alpha \to z$. A derivation step $x \Rightarrow y$ is defined by demanding that "rewriting takes place in parallel". I.e. $x = \alpha_1\alpha_2\ldots\alpha_n \Rightarrow z_1z_2\ldots z_n = z$ iff $\alpha_i \to z_i$ is a production of P. Other notions are defined as for CF grammars.

An ETOL system posesses a number of sets of productions $P_1, P_2, \ldots, P_k$ called "tables" (rather than just one such set as in the EOL case). For each derivation step all productions used must be taken from the same table. Otherwise derivations proceed as in EOL systems. An OL system is an EOL system with $V = \phi$ (the empty set) which evidently necessitates the use of a start word (consisting of terminals) instead of a start symbol. A DOL system is an OL sytem with just one production for each symbol.

For all of the above generative devices we talk about derivation, length of derivation, derivation tree, unambiguity etc. in the usual fashion. In particular, the height of a derivation tree is the length of the longest path from the start symbol (axiom) to a leaf of the tree.

In all examples of generative systems, capital letters will denote variables, small letters will denote terminals and S (or indexed, primed or barred version thereof) will be the start symbol.

Languages differing only by ε, the empty word, will be considered equal. Alph (X) denotes the set of all symbols used in the word, or set of words, X.

A finite automaton A is a 5-tuple, $A = (\Phi, \Sigma, \delta, q_0, F)$ where Φ and Σ are finite sets of states and input symbols respectively, where δ: $\Phi \times \Sigma \to \Sigma$ is the transition function, $q_0 \in \Phi$ is the start state and $F \subseteq \Phi$ is a set of final states. δ is extended to $\Phi \times \Sigma^*$ by defining $\delta(q,\varepsilon) = q$ for all $q \in \Phi$ and

$\delta(q,xy) = \delta(\delta(q,x),y)$ for all $q \in \Phi$ and $x,y \in \Sigma^*$. $T(A)$, the set of words accepted by A, is defined by $T(A) = \{x \mid \delta(q_0,x) \in F\}$. Sets accepted by finite automata are called regular.

A gsm g is a (nondeterministic) version of a finite automaton including outputs, $g = (\Phi,\Sigma,\Delta,\delta,q_0,F)$. Φ,Σ,q_0,F have the same significance as for a finite automaton, Δ is a finite set of outputs and δ: $\Phi \times \Sigma \to 2^{\Phi} \times \Delta^*$ defines for each pair $(p,a) \in \Phi \times \Sigma$ a finite set of pairs $(p_1,w_1), (p_2,w_2),\ldots,(p_k,w_k)$ indicating that in state p with input a, for each i, state p_i can be attained, producing output w_i. We write $y = y_1y_2 \ldots y_n \in g(a_1a_2 \ldots a_n)$ with $a_i \in \Sigma$ iff for some sequence $q_0,q_1,\ldots,q_n \in F$ of states $(q_i,y_i) \in \delta(q_{i-1},a_i)$ $(i = 1,2,\ldots,n)$. A gsm g is called bounded erasing on a language L if for some constant c, $y \in g(x)$ implies $|x| \leq c|y|$, where $|z|$ denotes the length of the word z.

Card (M) will denote the cardinality of the set M; for a set M of positive integers $\ell cm(M)$ will denote the least common multiple of all numbers in M.

Let $f(n)$, $g(n)$ be functions of integers ≥ 0 into again such integers. We write $f(n) = O(g(n))$ if for some constant c $f(n) \leq cg(n)$ for all n. We write $f(n) = o(g(n))$ and say $f(n)$ is smaller than $g(n)$ if for each $c > 0$, $f(n) < cg(n)$ for all $n \geq n_0$.

We now introduce the basic notions of this paper. A CF grammar G is called of height $f(n)$, if for each word $x \in L(G)$

there exists a derivation tree of height $O(f(|x|))$. A CF language L is called of <u>height</u> f(n), if for some G with L(G) = L, G is of height f(n).

If a CF grammar G (or a language L) is of <u>height</u> f(n) where

(a) $f(n) = O(\log_\alpha n)$ $(\alpha > 1)$, then we speak of <u>logarithmic</u> height,

(b) f(n) is smaller than n, then we speak of <u>sublinear</u> height.

We will apply the notion of height not only to CF grammars and languages, but quite general to X grammars and languages, where $X \in \{CF, EOL, ETOL, DOL, \text{unambiguous CF}, \ldots\}$. We will then talk of X-height to emphasize the class of generative devices we are considering.

3. Results

One of the main aims of this section is to establish that each regular language is of logarithmic height but that each of a large class of nonregular CF languages cannot be generated by a sub-linear grammar. The first of these results is obtained directly as Theorem 1, the other one, Theorem 2, is based on three Lemmas, two of which are of some interest in themselves.

Theorem 1

Every regular language is of logarithmic height.

Proof:

Let $R \subseteq \Sigma^*$ be an arbitrary regular language. We will exhibit a CF grammar G of logarithmic height which generates R.

Let $A = (\Phi,\Sigma,\delta,q_0,F)$ be a finite automaton accepting R, $R = T(A)$. For each pair of states $p,q \in \Phi$ let $[p,q]$ be a new symbol. Let V be a set of symbols consisting of all such symbols $[p,q]$ and one additional symbol S.

Define $G = (V,\Sigma,P,S)$ with productions P as follows:

$$P = \{S \to [q_0,q] \mid q \in F\} \cup$$
$$\{[p,q] \to [p,r]\,[r,q] \mid p,q,r \in \Phi\} \cup$$
$$\{[p,q] \to a \mid \delta(p,a) = q\}.$$

Augment G by the production $S \to \varepsilon$ iff ε is in R.

We first observe that $L(G) = T(A)$. Consider an arbitrary nonempty word $x = a_1a_2 \ldots a_n$ with $a_i \in \Sigma$ $(1 \leq i \leq n)$. Suppose $x \in T(A)$. Then $\delta(q_0,a_1) = q_1, \delta(q_1,a_2) = q_2, \ldots, \delta(q_{n-1},a_n) = q_n$ with $q_n \in F$. Hence $[q_0,q_1] \to a_1$, $[q_1,q_2] \to a_2, \ldots, [q_{n-1},q_n] \to a_n$ and $S \to [q_0,q_n]$ are productions of P. Observe also that $[q_0,q_i] \to [q_0,q_{i-1}]\,[q_{i-1},q_i]$ for $i = 2,3,\ldots,n$. Hence we have $S \to [q_0,q_n] \overset{*}{\Rightarrow} [q_0,q_1]\,[q_1,q_2] \ldots [q_{n-1},q_n] \overset{*}{\Rightarrow} a_1a_2 \ldots a_n$. Suppose $x \in L(G)$. Then $S \to [q_0,q_n] \overset{*}{\Rightarrow} [q_0,q_1]\,[q_1,q_2] \ldots [q_{n-1},q_n] \overset{*}{\Rightarrow} \overset{*}{\Rightarrow} a_1a_2 \ldots a_n$ with $q_n \in F$ holds. Hence $\delta(q_0,a_1) = q_1$, $\delta(q_1,a_2) = q_2, \ldots, \delta(q_{n-1},a_n) = q_n$, i.e. $x \in T(A)$.

It remains to establish that any word x can be generated by a tree of height $\sim \ell g\,|x|$. It clearly suffices to show, by the first part of the above proof, that any derivation

$$[q_i,q_j] \overset{*}{\Rightarrow} [q_i,q_{i+1}]\,[q_{i+1},q_{i+2}] \ldots [q_{j-1},q_j]$$

requires a derivation tree of height at most $\sim \ell g\,|j - i|$. Since $[q_i,q_j] \to [q_i,q_k]\,[q_k,q_j]$ is a production of P for any k we can choose q_k as (roughly) the midpoint of the sequence $q_i,q_{i+1},\ldots,q_j$. Repeating the process, a tree of the desired small height is obtained. □

Although it is well known that regular languages can be generated by unambiguous CF grammars, every CF grammar of sublinear height generating an infinite language must be ambiguous. This is seen as follows. Let G be an unambiguous CF grammar generating an infinite language. Then there exist terminal words u, v, w, x, y and a

nonterminal A, such that $S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} uvwxy$, S the start symbol of G. By the unambiguity of S, each word uv^iwx^iy has a (unique) derivation of height exceeding i for all $i \geq 0$. Hence G is not sublinear.

Our next goal is to obtain a large class of CF languages none of which is of sublinear height.

We proceed in four steps. We first show that the height of languages does not increase under certain operations. We then establish a normal form result for a certain class of CF languages. Based on this we present a class c_0 of languages none of which is of sublinear height. We finally show that each language of a large class c_1 can be reduced to a language of the class c_0 by the operations mentioned, hereby proving that no language of c_1 is of sublinear height.

Lemma 1

Let L, L' be CF languages of height $f(n)$. Then all of the following languages are also of height $f(n)$:

(1) $L \cup L'$

(2) $L\,L'$

(3) L^*

(4) $L \cap R$, R a regular set

(5) $g(L)$, g a gsm which is bounded erasing.

Proof

(1) and (2) are obvious, and (4) follows from (5). To see (3), let $L = L(G)$, $G = (V,\Sigma,P,S)$ where G is of height $f(n)$. Define $\bar{G} = (V \cup \{\bar{S}\},\Sigma,P \cup \{\bar{S} \to \bar{S}\bar{S} \mid S\},\bar{S})$. Evidently, $L(\bar{G}) = L^*$. It is readily seen that $\bar{G}$ is of height $f(n)$.

To prove (5) we use the "usual triple construction": Let $L = L(G)$, $G = (V,\Sigma,P,S)$, G is of height $f(n)$. We may assume that G is in Chomsky Normal Form, since transformation of a grammar into Chomsky Normal Form increases the height of derivation trees by at most a constant factor.

Let $M = (\Phi,\Sigma,\Delta,\delta,q_0,F)$ be a gsm, bounded erasing on L. Let $\bar{V} = \{[p,A,q] \mid p,q \in \Phi, A \in V\} \cup \{S\}$. Define $\bar{G} = (\bar{V},\Sigma,\bar{P},S)$, where $\bar{P}$ consists of the following productions:

$$\bar{P} = \{S \to [q_0,S,q] \mid q \in F\} \cup$$
$$\{[p,A,q] \to [p,B,r]\,[r,C,q] \mid A \to BC \in P,\ p,r,q \in \Phi\} \cup$$
$$\{[p,A,q] \to w \mid A \to a \in P,\ \delta(p,q) \ni (q,w)\}.$$

It is clear that $L(\bar{G}) = g(L)$. Further, for every word $y \in g(x)$, $x \in L$, y has a derivation tree of height t (with respect to $\bar{G}$) provided that x has a derivation tree of such height (with respect to G). By assumption of linear-bounded erasing, for some c depending only on g, $|y| > c|x|$. Hence $g(L)$ is of height $f(n)$ if this is true of L. □

Our next Lemma provides a somewhat technical normal form result required later.

Lemma 2

Let $L \subseteq a^*b^*$ be a CF language, $L = L(H)$, H of height $f(n)$. Then there exists a grammar $G = (V,\{a,b\},P,\bar{S})$ of height $f(n)$ generating L and satisfying conditions (1) - (4):

(1) $V = V_a \cup V_b \cup V_M \cup \{S\}$, (disjoint union).

(2) $P = P_a \cup P_b \cup P_T \cup P_{Reg} \cup P_{Start} \cup P_{Loop}$, (disjoint union).

(3) $P_a = \{A_p \to A_pA_p \mid a^p \mid \varepsilon \mid A_p \in V_a\}$;

$P_b = \{B_q \to B_qB_q \mid b^q \mid \varepsilon \mid B_q \in V_b\}$;

P_T consists only of productions of the form $\alpha \to a^mb^m$ where $\alpha \in V_M \cup \{\bar{S}\}$, $m,n \geq 0$;

P_{Reg} consists only of productions of the form $\bar{S} \to \alpha a^mb^n\beta$ where $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}$, $m,n \geq 0$;

P_{Start} consists only of productions of the form $\bar{S} \to \alpha a^mDb^n\beta$ where $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}$, $\alpha\beta \neq \varepsilon$, $m,n \geq 0$;

P_{Loop} consists only of productions of the form $D \to a^mEb^n$ where $D,E \in V_M$, $m,n \geq 0$.

(4) $D \overset{*}{\Rightarrow} a^mDb^n$ for $D \in V_M$ implies $m > 0$ and $n > 0$.

Proof

We will prove the Lemma by systematically transforming H without changing the language generated and without increasing the height of derivation trees involved.

We may assume that H does not contain productions of the form $A \to \varepsilon$ and $A \to B$, $A,B \in V^{(H)}$. (Removing such productions can clearly be done without increasing the height of H.) Also, we may assume that $L_A(H)$ is infinite for each $A \in V^{(H)}$. (Or else each occurrence of A could be replaced by all words of $L_A(H)$.)

$$\text{Let } V_a^{(H)} = \{A \in V^{(H)} \mid L_A(H) \subseteq a^*\},$$

$$V_b^{(H)} = \{B \in V^{(H)} \mid L_B(H) \subseteq b^*\},$$

$$V_M^{(H)} = \{C \in V^{(H)} \mid \text{alph}(L_C(H)) = \{a,b\}\}.$$

Let $p: c \to c_1 c_2 \ldots c_n$, $c_i \in V \cup \{a,b\}$ be any production of $P^{(H)}$.

Case 1: For some i, $c_i \in V_M^{(H)}$. Then $L_{c_1 \ldots c_{i-1}}(H) \subseteq a^*$, $L_{c_{i+1} \ldots c_n}(H) \subseteq b^*$, and both sets are regular. Hence we may replace p by a finite number of productions of the type $c \to \alpha a^m c_i b^n \beta$ with $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}$, $m,n \geq 0$, where here and in what follows V_a and V_b are sets of variables of the type A^p, B^q respectively, for which productions as specified in point (3) of the formulation of the Lemma are provided.

Case 2: For some i, $C_i \in V_a^{(H)}$ and $C_{i+1} \in V_b^{(H)}$. Then $L_{C_1 \ldots C_i}(H) \subseteq a^*$, $L_{C_{i+1} \ldots C_n}(H) \subseteq b^*$ and both sets are regular. Hence we may replace p by a finite number of productions of the form $C \to \alpha a^m b^n \beta$, where $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}$. By replacing every production p by a set of productions as described, and by adding productions $A_p \to A_p A_p \mid a^p \mid \varepsilon$ for each new variable A_p, and $B_q \to B_q B_q \mid b^q \mid \varepsilon$ for each new variable B_p, we finally obtain an equivalent grammar $E = (V^{(E)}, \{a,b\}, P^{(E)}, S)$ of height $f(n)$ as follows:

$V^{(E)} = V_a \cup V_b \cup V_M^{(E)}$, $S \in V_M^{(E)}$, and $P^{(E)} = P_a \cup P_b \cup P_T^{(E)} \cup P_{Loop}^{(E)}$

where P_a, P_b are sets as described earlier, where $P_T^{(E)}$ consists of productions of the form $X \to a^m b^n$ for $X \in V_M^{(E)}$ and $m,n \geq 0$; and where P_{Loop} consists of productions of the form

$X \to \alpha a^m Y b^n \beta$ with $X,Y \in V_M^{(E)}$, $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}$, $m,n \geq 0$.

We now want to replace the productions of P_{Loop} by productions of the form $X \to a^m Y b^n$. Variables of V_a and V_b will be produced as required once and for all in a starting step. This is possible by noting that once a variable A_p or B_q has been introduced, it is not necessary that a variable $X \in V_M^{(E)}$ introduces a further A_p or B_q: such further A_p's or B_q's can be introduced by $A_p \to A_p A_p$ or $B_q \to B_q B_q$, instead.

For every $X \in V_M^{(E)}$, $U \subseteq V_a$, $W \subseteq V_b$ let $[U, X, W]$ be a new symbol. Let $R(U)$ be the regular set $\prod_{A \in U} L_A(E)$; since $R(U) \subseteq a^*$ there exists a finite set $P(U)$ of words a^m and $a^m A_p$ such that each word of $R(U)$ is either in $P(U)$ or can be obtained from a word in $P(U)$ by productions of the form $A_p \to A_p A_p \mid a^p \mid \varepsilon$. Add all symbols A_p not yet in V_a to V_a, and add all productions for such an A_p to P_a. Similarly, let $R(W)$ be the regular set $\prod_{B \in W} L_B(E)$, define $P(W)$ as a finite set of words b^n and $b^n B_q$ analogously, and enlarge V_b and P_b as may be necessary.

We now construct a grammar $G = (V, \{a,b\}, P, \bar{S})$ which we will then slightly modify to obtain a grammar of desired form.

Let $V = V_a \cup V_b \cup V_M \cup \{S\}$ where

$V_M = \{[U,X,W] \mid X \in V_M^{(E)},\ U \subseteq V_a,\ V \subseteq V_b\}$, and

$P = P_a \cup P_b \cup P_T \cup P_{Start} \cup P_{Loop}$ as follows:

P_a, P_b as described earlier;

$P_{Start} = \{\bar{S} \to \alpha[U,S,W]\beta \mid [U,S,W] \in V_M,\ \alpha \in P(U),\ \beta \in P(W)\}$;

$P_T = \{\bar{S} \to X \mid S \to X \in P_T^{(E)}\} \cup$

$\{[\phi,X,\phi] \to a^m b^n \mid X \in V_M^{(E)},\ X \to a^m b^n \in P_T^{(E)}\}$;

$P_{Loop} = \{[U,X,W] \to a^m[U \setminus \{\alpha\},\ Y, W - \{\beta\}]b^n \mid X \to \alpha a^m Y b^n \beta$ is production of $P_{Loop}^{(E)}$, $X,Y \in V_M^{(E)}$, $\alpha \in V_a \cup \{\varepsilon\}$, $\beta \in V_b \cup \{\varepsilon\}\}$.

Note that the grammar G obtained at this point is equivalent to the original grammar H, has height $f(n)$ and is already of the desired form except for two points:

(i) P_{Start} may contain productions $\bar{S} \to \alpha a^m Y b^n \beta$ with $\alpha \neq \varepsilon$ and $\beta \neq \varepsilon$,

(ii) $D \overset{*}{\Rightarrow} a^m D b^n$ for $D \in V_M$ with either $m = 0$ or $n = 0$ is possible.

We will first show that $L_{\alpha a^m Y b^n \beta}(G)$ for $\alpha \neq \varepsilon$ and $\beta \neq \varepsilon$ is a regular set. Hence each production $\bar{S} \to \alpha a^m Y b^n \beta$ with $\alpha \neq \varepsilon$ and $\beta \neq \varepsilon$ can be replaced by a finite set of productions as described in condition (3) of the Lemma under P_{Reg}.

For any $X \in V_M$ let

$$g(X) = \{(a^i, b^j) \mid X \Rightarrow a^{i_1} X_1 b^{j_1} \Rightarrow a^{i_2} X_2 b^{j_2} \Rightarrow \ldots$$

$$\Rightarrow a^{i_k} X_k b^{j_k} \Rightarrow a^i X b^j,\ k \geq 0,\ X_1, X_2, \ldots, X_k \text{ all different}\}.$$

Define $g(V_M) = \bigcup_{X \in V_M} g(X)$ and $\tau = \text{card}(g(V_M))$. (Observe that $g(X)$ describes all "elementary loops" on X). Define $\rho = \text{card}(V_M)$ and let

$$\mu = \ell cm \{pq \mid \bar{S} \to A_p a^m Y b^n B_q \in P_{Start}\}.$$

Consider a derivation tree starting with $\bar{S} \to A_p a^m Y b^n B_q$ for a word x such that the tree contains more than $\mu \cdot \tau \cdot \rho$ productions for symbols of V_M. Thus, the tree contains $\mu \cdot \tau$ independent (i.e.

non-overlapping) elementary loops on symbols of V_M. Hence one elementary loop $X \overset{*}{\Rightarrow} a^i X b^j$ (i.e. $(i,j) \in g(V_M)$) has been used (at least) μ times. Consider μ such loops. They have introduced $i \cdot \mu = i \cdot p \cdot q \cdot c$ a's and $j \cdot \mu = j \cdot p \cdot q \cdot c$ b's. We eliminate the μ loops under discussion and generate $a^{i \cdot p \cdot q \cdot c}$ and $b^{j \cdot p \cdot q \cdot c}$ by using the productions $A_p \to A_p A_p \mid a^p$ and $B_q \to B_q B_q \mid b^q$, instead.

Repeating this argument a derivation tree for x starting with $\bar{S} \to A_p a^m Y b^n B_q$ is obtained with at most $\mu \cdot \tau \cdot \rho$ productions for symbols of V_M. Hence nonregular productions are required just a finite number of times. Each production $\bar{S} \to A_p a^m Y b^n B_q$ thus generates a regular set only. We can therefore remove productions of this form from P_{Start} and replace them by a finite set of productions P_{Reg} as described in the statement of the Lemma.

To complete the proof, it remains to assure condition (4). It is easy to see that derivations of the form $D \overset{*}{\Rightarrow} a^m D b^n$ with $mn = 0$ can be prohibited without changing the language by allowing the generation of $(a^m)^*$ and $(b^n)^*$ by means of suitable start productions. We leave the details to the reader. □

Using the above Lemma we are now able to establish that any CF nonregular subset of a^*b^* cannot be generated by low derivation trees.

Lemma 3

Let $L \subseteq a^*b^*$ be a CF language which is generated by some grammar H of sublinear height. Then L is regular.

Proof

We may assume $L = L(G)$, where G is as described in Lemma 2. Let Π be the product of all p's such that $\bar{S} \to A_p a^m b^n A_q$ or $\bar{S} \to A_p a^m D b^n$ is a production of P.

Consider some production $\bar{S} \to A_p a^i D b^j$. We will show that $Z = L_{A_p a^i D b^j}(G) = (a^p)^* Y$ where $Y = L_{a^i D b^j}(G)$ is regular. To this end, consider an arbitrary word $x = uv \in Z$ with $v \in M$, where v contains sufficiently many b's, $v = a^m b^n$. Since it must be possible to generate $v_t = a^{tp+m} b^n$ for $t = 0,1,2,\ldots,\frac{\Pi}{p}$ by means of starting productions $\bar{S} \to A_r a^k b^\ell B_s$ (or else no generation tree of v_t of sublinear height would exist) we have:

$$\bar{S} \to A_{r_t} a^{k_t} b^{k_t} B_{s_t} \overset{*}{\Rightarrow} a^{tp+m} b^n \quad \text{for} \quad t = 0,1,2,\ldots,\frac{\Pi}{p}.$$

Since each r_t divides Π by definition of Π we also have

$$\bar{S} \to A_{r_t} a^{k_t} b^{k_t} B_{s_t} \overset{*}{\Rightarrow} a^{\Pi \cdot k + t \cdot p + m} b^n \text{ for } k \geq 0,\ t = 0,1,2,\ldots,\frac{\Pi}{p}.$$

Thus, all words $a^{tp+m} b^n$ $(t \geq 0)$ can be generated via a starting production $\bar{S} \to A_p a^k b^\ell B_q$. Hence, productions of the form $S \to A_p a^i D b^j$ are not required except possibly for the generation of words with a

bounded number of b's. The set of such words is regular. Therefore, all productions $\bar{S} \to A_p a^i D b^j$ can be eliminated by increasing the set P_{Reg}. Analogously, all productions $\bar{S} \to a^i D b^j B_q$ can be removed by increasing P_{Reg}. Productions $\bar{S} \to a^m D b^n$ can be removed without even changing P_{Reg}, since any word generated via such production must also have a derivation with another starting production, if only sublinear derivation trees are considered.

We have thus transformed G into an equivalent grammar whose only production are of the type P_a, P_b, P_T and P_{Reg} as described in condition (3) of Lemma 2. Hence $L = L(G)$ is regular. □

Combining Lemma 1 and Lemma 3 we obtain our main result.

Theorem 2

If L is a CF language such that for some gsm g which is bounded erasing on L, g(L) is a nonregular subset of a*b*, then L is not of sublinear height.

Proof

Suppose L can be generated by some sublinear grammar. Then g(L) can be generated by some sublinear grammar by Lemma 1. Since g(L) is a nonregular CF subset of a*b* this contradicts Lemma 3. □

Theorem 2 has a number of interesting aspects. It could be that the following statement (A) holds:

(A): For every nonregular CF language L there exists a bounded erasing gsm g such that g(L) is a nonregular subset of a*b*.

By Theorem 2, (A) would immediately imply (B):

(B): No nonregular CF language L is of sublinear height.

We conjecture that (B) is true. Whether (B) can be established by proving (A) is somewhat doubtful, since proving (A) would imply (C):

(C): It is undecidable whether a CF grammar G with $L(G) \subseteq a^*b^*$ generates a regular set.

Statement (C) can be deduced from (A) roughly as follows:

It is known to be undecidable of degree 2 whether a CF grammar generates a regular set, cf. [7,p.286]. The validity of (A) and an algorithm for deciding the regularity of a CF subset of a*b* would, however, provide a semi-algorithm for deciding non-regularity of a CF language: we examine, one by one, all of the enumerable many gsm's $g_1, g_2, g_3, \ldots$. For each such g_i, (CF languagesare effectively closed under gsm mappings and the inclusion of a CF language in a regular set is decidable) we check whether $g_i(L) \subseteq a^*b^*$ holds and if so, if $g_i(L)$ is nonregular. If L is indeed nonregular this procedure will terminate after a finite amount of time.

The general problem of deciding for an arbitrary CF bounded language L (cf. [3]) whether it is regular is, to our knowledge, still open. Even the "simple" subcase (C) does not seem to have been solved yet, despite the fact that already [3] contains a comparatively elegant representation theorem for CF subsets of a*b*.

Corollary 1

Let L be a CF language such that for some words u,v,w,x,y with $v \neq \varepsilon$, $x \neq \varepsilon$ the language $L \cap uv^* w x^*y$ is nonregular. Then L is not of sublinear height.

Proof:

Suppose L is of sublinear height. By (4) of Lemma 1 $L \cap uv^* w x^*y$, is also of sublinear height.

We now construct a gsm g, bounded erasing on $uv^* w x^*y$, such that $g(L \cap uv^* w x^*y)$ is a nonregular subset of a^*b^*. This contradicts Theorem 2.

The gsm g is simple enough to be described intuitively: first, the prefix u is read without producing any output, then, g switches to some new set of states, outputting one a everytime it finishes reading a word v; next, g switches nondeterministically into a new set of states, reads w without producing output, switches to still another set of states in which reading x produces one symbol b for each x; finally, a last set of states is nondeterministically entered in which y is read without output.

We note that for $z \in L \cap uv^* w x^*y$ we have $g(z) = \{a^i b^j \mid uv^i w x^j y = z\}$. Thus, the gsm mapping f producing the word v for each a (except for producing uv for the first a) and producing x for each b (except for wx for the first and xy for the last b) is the (mathematical) inverse of g, i.e.,

It is easy to see that (B) can also be derived by obtaining a space-bound theorem on pda's: Let us call a pda P <u>sublinear</u> if for some $f(n) = o(n)$ and for every word X accepted by P there is a computation accepting X during which the stack never contains more than $f(|X|)$ symbols. Result (D) (which we conjecture to hold) would imply (B) as is seen readily.

(D): No nonregular CF language is accepted by a sublinear pda.

Incidently, (B) does not necessarily imply (D).

One intuitive reason why sublinear grammars do not seem to be able to generate CF nonregular language is the fact that sublinear derivation trees seem to allow "pumping" in too many places. That this approach, somewhat surprisingly, cannot be successful is demonstrated by the following example.

Let $L = \{x \in \{a,b\}^* \mid \#_a(x) = \#_b(x)$, neither a^3 nor b^3 occur as subword of $x\}$.

Despite the fact that $x = x_1 u x_2 \in L$ and $|u| \geq 3$ implies $u = u_1u_2u_3u_4u_5$ with $|u_2u_4| \neq 0$ and $x_1u_1u_2^iu_3u_4^iu_5x_2 \in L$ for $i \geq 0$ (i.e. despite the fact that within every reasonably long subword pumping is possible), L is clearly nonregular CF.

After this lengthy discussion we turn to a special case of Theorem 2 which we think is interesting in its own right.

$f(g(L \cap uv^* w x^*y)) = L \cap uv^* w x^*y$. Since f preserves regularity, $g(L \cap uv^* w x^*y)$ cannot be regular, completing the proof.

Some remarks concerning the above Corollary 1 are in order. In view of various "pumping lemmas" one might think that statement (E) holds.

(E): If L is a nonregular CF then there exist words u,v,w,x,y with $v \neq \varepsilon$, $x \neq \varepsilon$ such that $L \cap uv^* w x^*y$ is nonregular.

By virtue of Corollary 1, (E) would certainly imply (B). Unfortunately, proving or disproving (E) does not seem to be easy. Deep results of a related nature have been obtained e.g. in [1] and [6] but do not seem to be applicable to resolving (E). We would like to point out that proving (E) would require to use both that L is nonregular and that L is CF. Observe that for the simple non CF language $L = \{a^n b^n c^n \mid n \geq 1\}$ and every choice of u,v,w,x,y the set $L \cap uv^* w x^*y$ is even finite.

We now turn our attention to the height of derivation trees in L systems. In view of the rather difficult problems arising even with CF grammars we cannot expect a systematic treatment of all aspects of the height of derivation trees for L systems. We will briefly demonstrate, however, a surprising richness of results indicating that a further study is warranted.

By Corollary 1, $\{a^n cb^n \mid n \geq 1\}$ is not of sublinear (CF)-height. It is, however, of logarithmic EOL height. Thus, the

transition to more powerful generative mechanisms allows to reduce the height of derivation trees. We conjecture that a similar phenomenon occurs in the transition from EOL to ETOL systems.

Theorem 3

$L_1 = \{a^n c b^n \mid n \geq 1\}$ is of logarithmic EOL height.

$L_2 = \{a^n b^n c^n \mid n \geq 1\}$ is of logarithmic ETOL height.

Proof

(a) Consider the EOL system G with productions: $S \to S \mid ASB \mid c$, $A \to AA \mid a$, $B \to BB \mid b$, $a \to N, b \to N, c \to N$, $N \to N$. Evidently, $L(G) \subseteq a^n c b^n$. To see that every word $a^n c b^n$ can be obtained by a tree of height $O(\log n)$ consider the binary presentation. $n = b_1 b_2, \ldots, b_k$. By using as i-th production for S the production $S \to ASB$ if $b_i = 1$ and $S \to S$ if $b_i = 0$, the desired word is obtained. (Thus, G is even unambiguous. Quite contrary to the CF case, logarithmically high derivation trees can be obtained by unambiguous EOL systems).

(b) Consider the ETOL system G with productions:

Table 0: $\bar{A} \to \bar{A}$, $\bar{B} \to \bar{B}$, $\bar{C} \to \bar{C}$, $S \to S$, $A \to AA$, $B \to BB$, $C \to CC$, $a \to a$, $b \to b$, $c \to c$

Table 1: $\bar{A} \to A\bar{A}$, $\bar{B} \to B\bar{B}$, $\bar{C} \to C\bar{C}$, $S \to S$, $A \to AA$, $B \to BB$, $C \to CC$, $a \to a$, $b \to b$, $c \to c$

Table 3: $S \to \bar{A}\bar{B}\bar{C}$, $\bar{A} \to a$, $\bar{B} \to a$, $\bar{C} \to a$, $A \to a$, $B \to a$,

$C \to c$, $a \to a$, $b \to b$, $c \to c$

Let $n = b_1 b_2 \ldots b_k$ be the binary representation of some positive integer n. Applying first Table 3, then Table b_1, Table b_2,..., Table b_k, and finally Table 3, $a^n b^n c^n$ is obtained. Hence $L(G) = L_2$, and G is of logarithmic height.

(c) We conjecture that L_2 is not of sublinear EOL height. The proof of this fact seems to require a detailed study of the generation of L_2 by an EOL system. It appears to be possible but rather complicated.

Because of the close relationship of the growth functions of of DOL systems and the heights of their derivations trees results on growth functions lead to anologous results on the height of derivation trees. In particular we have the following:

Theorem 4

For each function $f(n)$ of the set $\bigcup_{k=2}^{\infty} \{n^{\frac{1}{k}}, \log_k n\}$ there exists a language L such that L is of DOL height $f(n)$, but of no smaller DOL height.

Proof

For each DOL system G there exist constants c_1, c_2 such that the n-longest word of $L(G)$ has a derivation tree of height k with $c_1 n \leq k \leq c_2 n$. Hence Theorem 4 follows directly from the well-known results on the growth functions of DOL system, cf. [5] and [8].

References

[1] Boasson, L.: Un critére de rationalité des languages algébriques, ICALP 1972, North Holland, Amsterdam, (1973), 359-365.

[2] Culik, K. II, Maurer, H.A.: Tree controlled grammar, Computing 19 (1977), 129-139.

[3] Ginsburg, S.: The mathematical theory of context free languages, McGraw Hill, New York, (1966).

[4] Harrison, M.A.: Introduction to formal language theory, Addison-Wesley, Reading, Mass., (1978).

[5] Hermann, G., Rozenberg, G.: Developmental systems and languages, North Holland, Amsterdam (1974).

[6] Nivat, M.: Une propriété des languages compilables, C.R. Academic Sciences, Series 1, 267 (1968), 244-246.

[7] Salomaa, A.: Formal languages, Academic Press, New York, (1973).

[8] Salomaa, A., Soittola, M.: Automata theoretic aspects of formal power series, Springer, New York (1978).