RAIRO. INFORMATIQUE THÉORIQUE

A. EHRENFEUCHT

G. ROZENBERG

On subword complexities of homomorphic images of languages

RAIRO. Informatique théorique, tome 16, nº 4 (1982), p. 303-316

<http://www.numdam.org/item?id=ITA_1982__16_4_303_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 16, n° 4, 1982, p. 303 à 316)

ON SUBWORD COMPLEXITIES OF HOMOMORPHIC IMAGES OF LANGUAGES (*)

by A. EHRENFEUCHT (1) and G. ROZENBERG (2)

Communicated by J. BERSTEL

Abstract. — *The subword complexity of a language K is the function π_K on positive integers such that $\pi_K(n)$ equals the number of different subwords of length n appearing in words of K. We investigate the relationship between $\pi_{h(K)}$ and π_K where K is a language and h is a homorphism. This study is also carried out for the special case when K is a DOL language.*

Résumé. — *La complexité des facteurs d'un langage K est la fonction π_K définie sur les entiers positifs telle que $\pi_K(n)$ est le nombre de facteurs distincts de longueur n apparaissant dans les mots de K. Nous étudions les relations existant entre $\pi_{h(K)}$ et π_K, où K est un langage et h est un homomorphisme. Cette étude est menée également dans le cas particulier où K est un DOL langage.*

INTRODUCTION

In the framework of L systems, investigating the subword complexity of a language turned out to be quite useful for the understanding of the role of the deterministic restriction on a rewriting system, *see*, e. g., [1, 2, 3 and 4]. (The subword complexity of a language K is the function π_K on the positive integers such that, for every n, $\pi_K(n)$ equals the number of different subwords of length n appearing in the words of K).

Our paper continues the work in this direction. In particular, we investigate the effect of a homomorphism on the subword complexity of a language, that is, given a language K and a homomorphism h we investigate the relationship between $\pi_{h(K)}$ and π_K.

In the first part of the paper we investigate the situation in the case that K is an arbitrary language. We demonstrate that no "meaningful" lower or upper bounds for the ratio $\pi_{h(K)}(n)/\pi_K(n)$ can be established even in the case that h is a nonerasing homomorphism. We also prove that if a language contains an

(*) Received July 1980, revised July 1981.
(1) Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado, U.S.A.
(2) Institute of Applied Math. and Computer Science, University of Leiden, Leiden, The Netherlands.

R.A.I.R.O. Informatique théorique/Theoritical Informaties, 0399-0540/1982/303/$ 5.00

infinite number of subwords then its subword complexity must be at least linear; in other words, sublinear subword complexities do not exist.

In the second part of the paper we investigate subword complexities of homomorphic images of DOL languages. This class of languages plays an important role in the theory of L systems, while at the same time it is much more difficult to handle than the class of DOL languages, *see*, e. g., [4]. (Let $\mathscr{L}$(DOL), $\mathscr{L}$(H DOL) and $\mathscr{L}$(H_Λ DOL) denote respectively the class of DOL languages, the class of homomorphic images of DOL languages and the class of languages of the form $h(K)$ where h is a nonerasing homomorphism and K is a DOL language). Surprisingly enough it turns out that the subword complexity of a language in $\mathscr{L}$(H DOL) is bounded by a function of order n^2; in this way there is no difference between DOL and HDOL languages (*see* [2]). We also show that in the general case of $\mathscr{L}$(H DOL) one cannot have the theory of subword complexity "sensitive to" natural structural restrictions on the underlying DOL systems; it is known (*see* [2]) that such a theory exists for the class $\mathscr{L}$(DOL). However if one considers $\mathscr{L}$(H_Λ DOL), such a theory is again possible.

The reader is assumed to be familiar with the theory of DOL systems, *see*, e. g., [4].

I. PRELIMINARIES

For a finite set K, $\# K$ denotes its cardinality; if $\# K=1$ then we often identify K with its element. For an integer x, $abs\, x$ denotes the absolute value of x. For a word y, $|y|$ denotes its length and $alph\, y$ denotes the set of all letters occuring in y; Λ denotes the empty word. If $y \neq \Lambda$ then *first* y denotes the leftmost letter in y, *last* y denotes the rightmost letter in y and y^∞ denotes the infinite in both directions word consisting of catenations of y only. For a positive integer n, $pref_n\, y$ denotes the prefix of y consisting of the n leftmost letters in y (if $|y|<n$ then $pref_n\, y=y$) and $pref\, y$ denotes the set of all prefixes of y. Analogously we use the notation $suf_n\, x$ and $suf\, x$ for suffixes. Also $sub_n\, y$ denotes the set of all subwords of y of length n and $sub\, x$ denotes the set of all subwords of y. For a language K, $pref_n\, K=\{pref_n\, y : y \in K\}$, $pref\, K=\bigcup_{y \in K} pref\, y$, $sub_n\, K=\{sub_n\, y : y \in K\}$ and $sub\, K=\bigcup_{y \in K} sub\, y$.

Given an alphabet Σ (fixed in the considerations) and $\Delta \subseteq \Sigma$, $pres_\Delta$ is the homomorphism on Σ^* defined by: $pres_\Delta\, x=\Lambda$ if $x \in \Sigma \backslash \Delta$ and $pres_\Delta\, x=x$ if $x \in \Delta$. For a homomorphism h on Σ^*, $maxr\, h=\max\{|h(x)| : x \in \Sigma\}$.

A DOL system G is specified in the form $G=(\Sigma, h, \omega)$ where Σ is its alphabet, h its homomorphism and ω its axiom; $L(G)$ denotes the language of G while $E(G)$ denotes its sequence. Also $maxr\ G = maxr\ h$. We say that G is *everywhere growing* if, for every $x \in \Sigma$, $|h(x)| \geqq 2$, and G is *uniformly growing* if there exists a positive integer $t \geqq 2$ such that, for every $x \in \Sigma$, $|h(x)| = t$; as usual we say that $L(G)$ is *everywhere growing* and *uniformly growing*, respectively. $\mathscr{L}$(DOL) denotes the class of all DOL languages, $\mathscr{L}$(HDOL) denotes the class of all homomorphic images of DOL languages and $\mathscr{L}(H_\Lambda \mathrm{DOL})$ denotes the set of all languages of the form $h(K)$ where h is a Λ-free homomorphism and K is a DOL language.

For a language K, its *subword complexity* is the function π_K on positive integers defined by: $\pi_K(n) = \# sub_n K$.

Since problems considered in this paper are trivial otherwise, we consider infinite languages only (unless indicated otherwise); in particular we consider only DOL systems G such that $L(G)$ is infinite.

To avoid cumbersome technicalities, most of the proofs in this paper are presented in a rather informal way. We are convinced that the reader (familiar with the theory of DOL systems) can complete all formal details in the proofs, if necessary.

II. ARBITRARY LANGUAGES

In this section we investigate the effect of a homomorphism on the subword complexity of a language, that is we investigate the relationship between $\pi_{h(K)}$ and π_K where K is a language and h is a homomorphism.

We start by establishing the lower bound on the subword complexity of a language. Our first result says that there do not exist sublinear (but not constant) subword complexities.

THEOREM 1: *Let K be a language. Either:*

(1) *$\pi_K(n) \geqq n+1$ for every positive integer n, or*

(2) *there exists a positive integer C such that $\pi_K(n) \leqq C$ for every positive integer n; moreover, in this case $K \subseteq K_0 \cup \bigcup_{i=1}^{r} K_i$ where $r \geqq 1$, K_0 is a finite language and, for every $i \in \{1, \ldots, r\}$, there exist words $x_{i,1}$, $x_{i,2}$ and y_i such that $K_i = x_{i,1} y_i^+ x_{i,2}$.*

Proof: Let $w \in sub\,K$, $K \subseteq \Delta^*$ (certainly we can assume that $\# \Delta \geqq 2$; otherwise the theorem trivially holds). We say that w is *deep* if for every positive integer n there exist words x, y such that $|x| > n$, $|y| > n$ and $xwy \in K$.

We prove the theorem essentially by analyzing deep subwords (of K). We use $dsub\,K$ to denote the set of deep subwords of K and, for a nonnegative integer n, $dsub_n K$ denotes the set of deep subwords of K of length n.

(i) Every (infinite) language contains infinitely many deep subwords.

This is obvious.

(ii) If w is a deep subword then there exist letters a, b in Δ such that wa and bw are deep subwords.

This is obvious.

(iii) For every nonnegative integer n, $\# dsub_{n+1} K \geqq \# dsub_n K$.

This follows directly from (ii).

(iv) If for every nonnegative integer n, $\# dsub_{n+1} K > \# dsub_n K$ then, $\pi_K(n) \geqq n+1$ for every nonnegative integer n.

This follows from the fact that $\pi_K(n) \geqq \# dsub_n K$ for every nonnegative integer n and moreover $\# dsub_1 K \geqq 2$. Thus (the first case of the statement of) the theorem holds.

(v) If the assumption of (iv) does not hold then there exists a positive integer n_0 such that $\# dsub_{n_0} K = \# dsub_{n_0+1} K$.

This is obvious.

In the rest of this proof n_0 will be a fixed constant satisfying (v).

(vi) For every $m > n_0$, $\# dsub_m K = \# dsub_{n_0} K$.

This is proved as follows.

Let, for $w \in dsub\,K$, $R(w) = \{ a \in \Delta : wa \in dsub\,K \}$ and $L(w) = \{ b \in \Delta : bw \in dsub\,K \}$. By (ii) we know that both $R(w)$ and $L(w)$ are nonempty. However, since $\# dsub_{n_0} K = \# dsub_{n_0+1} K$, if $|w| = n_0$ then $\# R(w) = \# L(w) = 1$. We will also use $R(w)$ and $L(w)$ to denote the unique elements of $R(w)$ and $L(w)$ respectively.

Now let $z_1, z_2 \in dsub_m K$ where $m > n_0$.

If $z_1 \neq z_2$ then $\mathrm{pref}_{n_0} z_1 \neq pref_{n_0} z_2$. This is seen as follows. If $z_1 \neq z_2$ but $pref_{n_0} z_1 = pref_{n_0} z_2$ then let q be the smallest positive integer such that the letter occurring on the q-th position in z_1 (say a_1) is different than the letter occurring on the q-th position in z_2 (say a_2). Thus $z_1 = u_1 a_1 x_1$ and $z_2 = u_1 a_2 x_2$ where $|u_1| = q-1$. Let $u = suf_{n_0} u_1$. Then $ua_1 \in dsub_{n_0+1} K$ and $ua_2 \in dsub_{n_0+1} K$, which implies that $\# dsub_{n_0+1} K > \# dsub_{n_0} K$; a contradiction. Consequently it must be that if $z_1 \neq z_2$ then $pref_{n_0} z_1 \neq pref_{n_0} z_2$.

This, however, implies that $\# dsub_m K = \# dsub_{n_0} K$ and so (vi) holds.

(vii) Now for each $w \in dsub_{n_0} K$ we construct the double infinite sequence $\rho(w)$ by appending to the right of w consecutively $R(w)$, $R(suf_{n_0}(w R(w)))$, . . . and to the left of w consecutively $L(w)$, $L(pref_{n_0}(L(w)w))$, . . . Then $\rho(w)$ is periodic, meaning that there exists a nonnegative integer m such that for each integer i, $f_{\rho(w)}(i)=f_{\rho(w)}(i-m)=f_{\rho(w)}(i+m)$ where $f_{\rho(w)}$ is the function (from integers into Δ) defining $\rho(w)$.

This follows because $dsub_{n_0}K$ is a finite set.

For each $\rho(w)$ we denote by $p_{\rho(w)}$ a fixed word α (called the *period of* $\rho(w)$) such that $\rho(w)=\alpha^{\infty}$. Let $W=\{\rho(w): w \in dsub_{n_0}K\}$. Now we write every word $x \in K$ in the form $x=F(x)\,D(x)\,T(x)$ where $D(x)$ is the leftmost occurrence among all subwords y of x such that y is a subword of some $\rho \in W$ and among all subwords of x that are subwords of (an element of) W none is longer than y. Now we partition K into sublanguages as follows.

(1) If $x \in K$ is such that $D(x)$ is a subword of a ρ in W of the form $D(x)=\alpha_1 p_\rho^n a_2$ for some $n \geqq 1$, α_1, $\alpha_2 \in \Delta^*$ where for no $m>n$, $D(x)$ can be written in the form $D(x)=\beta_1 p_\rho^m \beta_2$ for some β_1, $\beta_2 \in \Delta^*$, then we say that $x \in K(\rho, \alpha_1, \alpha_2)$.

(2) U_0 consists of all words x in K that cannot be written in the form $F(x)\,D(x)\,T(x)$ where $D(x)=\alpha_1 p_\rho^n \alpha_2$ for some $n \geqq 1$, $\rho \in W$ and α_1, $\alpha_2 \in \Delta^*$.

(viii) U_0 is finite.

Otherwise, by (i), U_0 contains infinitely many deep subwords and so it contains words that can be written in the form indicated in (1) above.

(ix) Consider an infinite sublanguage $M=K(\rho, \alpha_1, \alpha_2)$. Let $exp\ M=\{n \geqq 1 : D(x)=\alpha_1 p_\rho^n \alpha_2$ for some $x \in M\}$. Then $exp\ M$ is infinite.

This is proved by contradiction as follows.

Assume that $exp\ M$ is finite. Then, clearly, the set $\{D(x): x \in M\}$ must be finite. Hence M must be finite [because (1) implies that an infinite language contains arbitrarily long deep subwords]; a contradiction.

(x) Consider an infinite sublanguage $M=K(\rho, \alpha_1, \alpha_2)$. Let $beg\ M=\{F(x): x \in M\}$. Then $beg\ M$ is finite.

This is proved by contradiction as follows.

Assume that $beg\ M$ is an infinite language. Hence there exists a letter in Δ, say b, such that for infinitely many words y in $beg\ M$ we have $last\ y=b$; let $Z_b=\{x \in M : last\ F(x)=b\}$. Now let $exp\ Z_b=\{n \geqq 1 : D(x)=\alpha_1 p_\rho^n \alpha_2$ for some $x \in Z_b\}$. Again [*see* the proof of (ix)] it is easily seen that $exp\ Z_b$ is infinite. But this implies that $b\,pref_{n_0}(\alpha_1 p_\rho^{n_0})$ is a deep subword and so $b=L(pref_{n_0}(\alpha_1 p_\rho^{n_0}))$. Consequently if we take a word x in M such that $last\ F(x)=b$ and $D(x)=\alpha_1 p_\rho^m \alpha_2$ with $m \geqq n_0$ then the subword of x starting on the last letter of $F(x)$ and ending on the last letter of $D(x)$ is also a subword of ρ; a contradiction.

(xi) Consider an infinite sublanguage $M=K(\rho, \alpha_1, \alpha_2)$. Let $end\, M=\{T(x): x\in M\}$. Then $end\, M$ is finite.

This can be proved analogously to (x).

But (viii) through (xi) implies that $K\subseteq K_0\cup \bigcup_{i=1} K_i$ where $r\geqq 1$, K_0 is a finite language and, for every $i\in\{1, \ldots, r\}$, there exist words $x_{i,1}$, $x_{i,2}$, and y_i such that $K_i=x_{i,1}y_i^+x_{i,2}$. Then clearly there exists a positive integer C such that $\pi_K(n)\leqq C$ for every positive integer.

This completes the proof of the theorem. □

It turns out that in the most general case, that is when h is an arbitrary homomorphism, nothing meaningful can be said about the relationship between $\pi_{h(K)}$ and π_K.

THEOREM 2: *For every positive integer e there exist a language K, $K\subseteq \Delta^*$, a positive integer constant C and a homomorphism $h:\Delta^*\to\Sigma^*$, where $\#\Sigma=e$, such that, for every positive integer n, $\pi_K(n)\leqq Cn$ and $\pi_{h(K)}(n)=e^n$.*

Proof: Let e be a positive integer and let $\tau_e=i_0 i_1\ldots$ be an ω-word (that is a one way infinite word) over the alphabet $\theta=\{1, \ldots, e\}$ such that every word over θ is a subword of τ_e. Let $\Sigma=\{b_1, \ldots, b_e\}$, $\Delta=\Sigma\cup\{a\}$, $a\notin\Sigma$ and let $K=\{b_{i_0}ab_{i_1}a^3\ldots b_{i_r}a^{3^r}: r\geqq 0$ and $i_0 i_1\ldots i_r$ is a prefix of $\tau_2\}$.

Let us estimate π_K first. Notice that if $x\in sub_n K$ for $n\geqq 1$ then either (1), $x=a^n$, or (2), $x=a^s b_j a^t$ for $j\in\{1, \ldots, e\}$ where $s+t=n-1$, or (3), $|pres_\Sigma x|\geqq 2$.

Clearly there is one word x satisfying (1) and en words satisfying (2). To estimate the number of words x satisfying (3) we proceed as follows.

If x satisfies (3) then x is of the form $x=yb_{j_1}a^{k(x)}b_{j_2}a^{l(x)}$ for $y\in\Delta^*$, b_{j_1}, $b_{j_2}\in\Sigma$, $l(x)$ a nonnegative integer and $k(x)$ a positive integer.

(i) Let $x\in sub_n K$, $x=yb_{j_1}a^{k(x)}b_{j_2}a^{l(x)}$ where $y\in\Delta^*$, b_{j_1}, $b_{j_2}\in\Sigma$, $l(x)\geqq 0$ and $k(x)\geqq 1$. Then there exist two positive integers q_1, q_2 such that either $k(x)=q_1$ or $k(x)=q_2$.

This is proved as follows.

Let q_1 be the maximal among all $k(x)$ (for all $x\in sub_n K$). Clearly for some $u\geqq 0$, $q_1=3^u$ and $n>3^u$. Then for every word $x=yb_{j_1}a^{k(x)}b_{j_2}a^{l(x)}$, $|y|<3^u<n$. Consequently if $k(x)\neq q_1$ then $k(x)=q_2=3^{u-1}$. Thus (i) holds.

But every word $x=yb_{j_1}a^{k(x)}b_{j_2}a^{l(x)}$ is uniquely determined by its suffix $a^{k(x)}b_{j_2}a^{l(x)}$ and so (i) implies that there are no more than $2en$ different word $x\in sub_n K$ satisfying (3).

Altogether $\pi_K(n) \leqq C n$ where $C = e + 3$.

Now let h be the homomorphism on Δ^* defined by $h(a) = \Lambda$ and $h(b_i) = b_i$ for $1 \leqq i \leqq e$. Clearly, for every nonnegative integer n, $\pi_{h(K)}(n) = e^n$.

Thus the theorem holds. $\square$

Notice that the language K used in the proof of Theorem 2 is such that π_K is a nondecreasing function. It turns out that when one considers an analogous situation for Λ-free homomorphisms then the jump in the subword complexity is rather limited.

THEOREM 3: *Let $K \subseteq \Delta^*$ be a language such that π_K is a nondecreasing function. Let h be a Λ-free homomorphism on Δ^*. Then there exists a positive integer constant C such that, for every positive integer n, $\pi_{h(K)}(n) \leqq C n \pi_K(n)$.*

Proof: Let $h: \Delta^* \to \Sigma^*$. Let $n \geqq 1$ and let $z \in sub_n h(K)$. Since h is Λ-free there exist $a \in \Delta \cup \{\Lambda\}$, $b \in \Delta \cup \{\Lambda\}$, $y \in sub\, K$ with $|y| \leqq n$ such that z is a subword of $h(ayb)$ where if $a \neq \Lambda$ then z is not a subword of $h(yb)$ and if $b \neq \Lambda$ then z is not a subword of $h(ay)$. Hence $z = z_1 z_2 z_3$ where z_1 is a suffix of $h(a)$, $z_2 = h(y)$ and z_3 is a prefix of $h(b)$. Consequently:

$$\pi_{h(K)}(n) \leqq (\# \Sigma)^{(2\, maxr\, h) - 2} \sum_{l=1}^{n} \pi_K(l).$$

Since π_K is a nondecreasing function:

$$\pi_{h(K)}(n) \leqq (\# \Sigma)^{(2\, maxr\, h) - 2}\, n\, \pi_K(n).$$

Hence if we set $C = (\# \Sigma)^{(2\, maxr\, h) - 2}$, the theorem holds. $\square$

Comparing Theorem 2 (and its proof) and Theorem 3 one sees a big difference between arbitrary and Λ-free homomorphisms as far as their effect on the subword complexity is concerned. However, it turns out that when one considers the case when π_K does not satisfy the "nondecreasing" restriction, the situation is quite different. First of all we demonstrate that in such a case there is no meaningful lower bound for the ratio $\pi_{h(K)}(n)/\pi_K(n)$.

THEOREM 4: *Let K be a language, $K \subseteq \Delta^*$ and let h be a Λ-free homomorphism on Δ^*. Let f be a function of positive integers such that $\lim_{n \to \infty} f(n) = \infty$. Then there exist an infinite set M of positive integers such that, for every $m \in M$, $f(m)\pi_K(m) > \pi_{h(K)}(m)$.*

Proof: Let $h: \Delta^* \to \Sigma^*$ and let $C = (\# \Sigma)^{(2\, maxr\, h) - 2}$.

Since a subword of $h(K)$ of length not exceeding n is "obtained" from a subword of K of length not exceeding n, reasoning as in the proof of Theorem 3 we obtain that, for every positive integer n,

$$\sum_{l=1}^{n} \pi_{h(K)}(l) \leqq C \sum_{l=1}^{n} \pi_K(l). \tag{1}$$

We prove the theorem by contradiction as follows. Assume that there exists a positive integer n_0 such that, for every $n > n_0$,

$$f(n)\,\pi_K(n) \leqq \pi_{h(K)}(n). \tag{2}$$

Consequently, for every $n > n_0$, we have :

$$\sum_{l=1}^{n} \pi_{h(K)}(l) = \sum_{l=1}^{n_0} \pi_{h(K)}(l) + \sum_{l=n_0+1}^{n} \pi_{h(K)}(l)$$

$$\geqslant \sum_{l=1}^{n_0} \pi_{h(K)}(l) + \sum_{l=n_0+1}^{n} f(l)\,\pi_K(l).$$

But then (1) implies that:

$$\sum_{l=1}^{n_0} \pi_{h(K)}(l) + \sum_{l=n_0+1}^{n} f(l)\,\pi_K(l) \leqq C \sum_{l=1}^{n} \pi_K(l).$$

Consequently:

$$\sum_{l=1}^{n_0} \pi_{h(K)}(l) \leqq C \sum_{l=1}^{n_0} \pi_K(l) + \sum_{l=n_0+1}^{n} (C - f(l))\,\pi_K(l).$$

Since $f(n)$ is ultimately growing, there exists a positive integer $n_1 > n_0$ such that for every $n \geqq n_1$, $f(n) > C$.

Consequently, for every $n > n_1$, we have:

$$\sum_{l=1}^{n_0} \pi_{h(K)}(l) \leqq C \sum_{l=1}^{n_0} \pi_K(l) + \sum_{l=n_0+1}^{n_1-1} (C - f(l))\,\pi_K(l) + \sum_{l=n_1}^{n} (C - f(l))\,\pi_K(l). \tag{3}$$

But for:

$$n > n_1 + C \sum_{l=1}^{n_0} \pi_K(l) + abs\left(\sum_{l=n_0+1}^{n_1-1} (C - f(l))\,\pi_K(l) \right),$$

we have:

$$\sum_{l=n_1}^{n} (C - f(l))\,\pi_K(l) < -C \sum_{l=1}^{n_0} \pi_K(l) - abs\left(\sum_{l=n_0+1}^{n_1-1} (C - f(l))\,\pi_K(l) \right)$$

and consequently (3) implies that:

$$\sum_{l=1}^{n_0} \pi_{h(K)}(l) < 0;$$

a contradiction.

Thus (2) cannot be true and consequently the theorem holds. □

Although we cannot prove the analogue of Theorem 2 for Λ-free homomorphisms we can show that, in general, no polynomial upper bound exists for the ratio $\pi_{h(K)}(n)/\pi_K(n)$.

THEOREM 5: *There exists a language K and a Λ-free homomorphism h such that for no polynomial f, $\pi_{h(K)}(n) \leqq f(n)\, \pi_K(n)$ for all positive integers n.*

Proof: Let $\tau = m_1, m_2, \ldots$ be an infinite sequence of positive integers such that, for each $i \geqq 1$, $m_{i+1} > m_i^2$. Then let K be the language over the alphabet $\{¢, a, b\}$ defined by $K = \bigcup_{i=1}^{\infty} ¢\,[a^{m_i}, b^{m_i}\}^{m_i}$ and let h be the homomorphism on $\{¢, a, b\}^*$ defined by $h(¢) = ¢^2$, $h(a) = a$ and $h(b) = b$.

Let us consider K first. Let us fix a positive integer i, let $n_i = m_1^2 + 2$ and let us compute $\pi_K(n_i)$. Clearly if $x \in sub_{n_i} K$ then $x \in sub\left(\bigcup_{r=i+1}^{\infty} ¢\,\{a^{m_r}, b^{m_r}\}^{m_r}\right)$

But then x is in one of the following forms:

(1) $x \in pref(¢a^+)$;

(2) $x \in pref(¢b^+)$;

(3) $x \in sub\, a^+$;

(4) $x \in sub\, b^+$;

(5) $x \in sub(a^+ b^+)$ where $alph\, x = \{a, b\}$;

(6) $x \in sub(b^+ a^+)$ where $alph\, x = \{a, b\}$.

But (remember that $|x| = m_i^2 + 2$) there is one x only satisfying (1), one x only satisfying (2), one x only satisfying (3), one x only satisfying (4), $m_i^2 + 1$ words x satisfying (5) and $m_i^2 + 1$ words x satisfying (6). Consequently $\pi_K(n_i) = 2 m_i^2 + 6$.

Consider now $h(K)$. Since by replacing $¢$ by $¢^2$ we have "padded" the length of words in $¢^2\{a^{m_i}, b^{m_i}\}^{m_i}$ to $m_i^2 + 2 = n_i$, if $x \in sub_{n_i} K$ then:

$$x \in sub\left(\bigcup_{r=i}^{\infty} ¢^2\,\{a^{m_r}, b^{m_r}\}^{m_r}\right)$$

It is obvious then that $\pi_{h(K)}(n_i) \geqq 2^{m_i}$. Consequently:

$$\frac{\pi_{h(K)}(n_i)}{\pi_K(n_i)} \geqq \frac{2^{\sqrt{n_i - 2}}}{2(n_i - 2) + 6}.$$

Since there are infinitely many n_i of the term m_i^2+2 where $i \geqq 1$ and since $2^{\sqrt{n_i-2}}/(2(n_i-2)+6)$ grows faster than any polynomial, the theorem holds. □

III. DOL LANGUAGES

In this section we consider the effect of homomorphisms on the subword complexities of DOL languages. We start by considering arbitrary DOL languages.

THEOREM 6: *Let K be a DOL language, $K \subseteq \Delta^*$, and let h be a homomorphism on Δ^*. There exists a positive integer constant C such that, for every positive integer n, $\pi_{h(K)}(n) \leqq C n^2$.*

Proof: Let $K=L(G)$ where $G=(\Delta, g, \omega)$ is a DOL system with $\# \Delta = m$. *We assume that* G satisfies the following condition: for every $n \geqq 1$ and every $a \in \Delta$,

$$alph\, g^n(a) = alph\, g(a). \tag{4}$$

(If G does not satisfy this condition then we can speed it up, *see*, e. g., [4], and deal with a finite number of DOL systems each of which satisfies this condition).

Let $n \geqq 1$ and let $z \in sub_n h(K)$. Then let s be the smallest integer t such that z is a subword of $h(\omega_t)$, where $E(G)=\omega_0, \omega_1, \ldots$ Let y be (a fixed occurrence of) the smallest subword of ω_s such that z is a subword of $h(y)$. The situation can be represented in figure 1.

Clearly in each ω_i, $0 \leqq i \leqq s-1$, we can distinguish (the occurrence of) the smallest subword that is the ancestor of y in ω_i; let us denote it by y_i. Let r be the smallest integer t such that the ancestor of y in ω_t consists of at least two letters, let this ancestor be α.

Let now, for each $i \in \{r, r+1, \ldots, s\}$, $\gamma(i)$ denote the number of occurrences of letters in ω_i that yield (through ω_s and then h) a nonempty contribution to z.

(i) $\gamma(i+(m+1)^2+1) > \gamma(i)$ for $r \leqq i < s-(m+1)^2-1$.

We prove it by a contradiction. Assume that (i) is not true, meaning that:

$$\gamma(i) = \gamma(j) \quad \text{for} \quad r \leqq i \leqq j \leqq i+(m+1)^2+1. \tag{5}$$

Let $c_{f,i}$ be the leftmost occurrence in y_i contributing a nonempty subword to z and let $c_{l,i}$ be the rightmost occurrence in y_i contributing a nomempty subword to z. Clearly (5) together with (4) implies that every occurrence c in y_i which contributes to z but is different from both $c_{f,i}$ and $c_{l,i}$ is such that it has only one propagating descendant on each level $i+1, \ldots, i+(m+1)^2+1$ and moreover all of those descendants are occurrences of the same letter.

On the other hand there must exist integers j_1, j_2 such that:

$i<j_1<j_2 \leqq i+(m+1)^2+1$, $c_{f,j_1}=c_{f,j_2}$ and $c_{l,j_1}=c_{l,j_2}$.

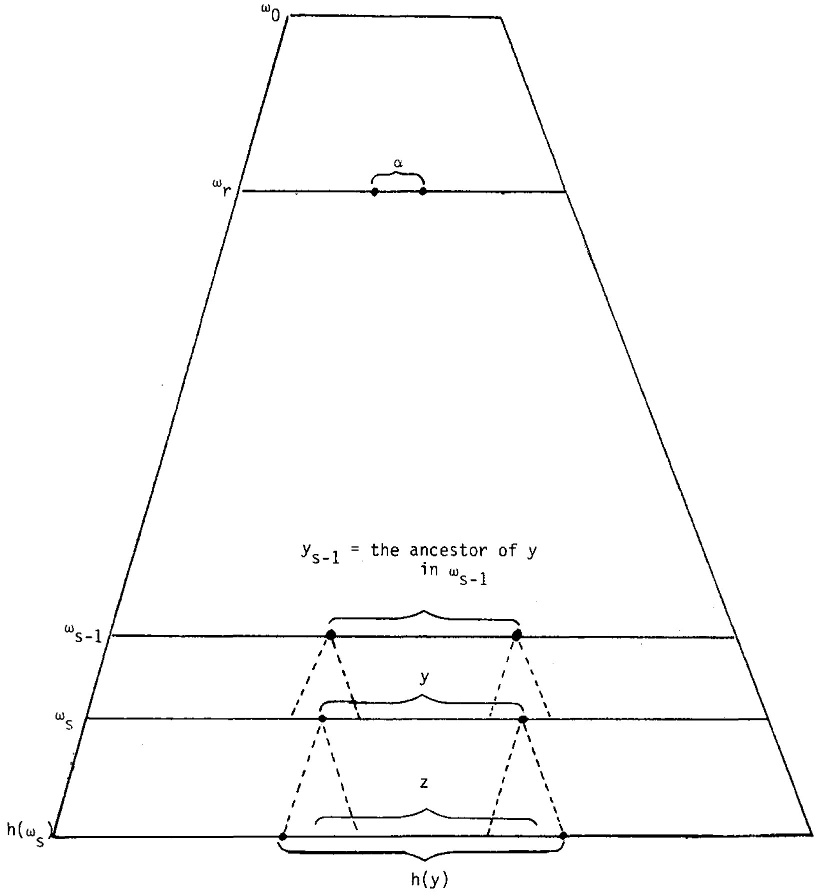

Fig. 1

But then z is a subword of $h(\omega_{s-(j_2-j_1)})$ which contradicts the minimality of s.

Hence (i) holds.

(ii) $l=(s-r) \leqq ((m+1)^2+1)n$.

This follows directly from (i).

Now let $\rho(z)$ be the prefix of z ending on the rightmost occurrence of a letter contributed by the leftmost occurrence of a letter in α, and let $q=|\rho(z)|$. Let the *description of* z be the triplet *des* $z=(\alpha, l, q)$.

(iii) If z_1, $z_2 \in sub_n h(K)$ and $des\, z_1 = des\, z_2$ then $z_1 = z_2$.

This is obvious.

(iv) $|\alpha| \leqq \max\{|\omega|, maxr\, g\} = p$.

This is obvious.

Now (ii), (iii) and (iv) imply that:

$$\pi_{h(K)}(n) \leqq m^p((m+1)^2+1)\, nn = m^p((m+1)^2+1)\, n^2.$$

Thus the theorem holds. □

The above theorem strengthens considerably the result from [2] which says that for a DOL language K there exists a positive integer constant C such that $\pi_K(n) \leqq C n^2$ for every nonnegative integer n. It is also shown in [2] that there exists a DOL language K and a positive real D such that $\pi_K(n) \geqq D n^2$ for every nonnegative integer n. Hence Theorem 6 presents the best possible upper bound on the subword complexity of an HDOL language.

In general a homomorphism can increase the subword complexity of a DOL language quite considerably (as a matter of fact from the "lowest possible" to the "highest possible" — compare Theorem 7 with Theorems 1 and 6).

THEOREM 7: *There exist a DOL language K, $K \subseteq \Delta^*$, a positive real C, a positive integer D and a homomorphism $h: \Delta^* \to \Sigma^*$ such that, for every positive integer n, $\pi_{h(K)}(n) \geqq C n^2$ and $\pi_K(n) < D n$.*

Proof: It is well-known, *see* [2], that there exist a DOL system $G = (\Sigma, g, \omega)$ and a positive real C such that, for every nonnegative integer n, $\pi_{L(G)}(n) \geqq C n^2$.

Let $G' = (\Delta, g', \omega)$ be the DOL system where $\Delta = \Sigma \cup \{a\}$, $a \notin \Sigma$, and g' is defined by $g'(x) = g(x)\, a^{maxr\, G - |g(x)|}$ for $x \in \Sigma$, and $g'(a) = a^{maxr\, G}$.

Notice that G' is a uniformly growing DOL system and so, *see* [2], there exist a positive integer D such that, for every nonnegative integer n, $\pi_{L(G')}(n) < D n$.

Let h be the homomorphism on Δ^* defined by $h(x) = x$ for $x \in \Sigma$ and $h(a) = \Lambda$. Then clearly $h(L(G')) = L(G)$ and consequently the theorem holds. □

Note that the DOL language K used in the proof of Theorem 7 is a uniformly growing DOL language and so one cannot have a theory of subword complexity of HDOL languages analogous to the case of DOL languages (*see* [2]) where considering everywhere growing and then uniformly growing DOL languages gave rise to the drop of subword complexity to the levels of the order of $n \log_2 n$ and n respectively.

However one obtains this kind of theory when one considers H_Λ DOL languages.

THEOREM 8: *Let K be an everywhere growing DOL language,* $K \subseteq \Delta^*$, *and let h be a* Λ*-free homomorphism on* Δ^*. *There exists a positive integer constant C such that, for every positive integer n,* $\pi_{h(K)}(n) \leq C n \log_2 n$.

Proof: Let $K = L(G)$ where $G = (\Delta, g, \omega)$ is an everywhere growing DOL system. Let $E(G) = \omega_0, \omega_1, \ldots$

Let $n \geq 1$ and let $z \in sub_n h(K)$. Let $s \geq 1$ be such that z is a subword of $h(\omega_s)$ and let us fix an occurrence of z in $h(\omega_s)$. Then let y be (a fixed occurrence of) the longest subword in ω_s which is mapped by h into a subword of the given occurrence of z. Finally let r be the smallest integer t such that ω_t contains a subword β whose constribution to ω_s is included in y; then let α be the longest such subword in ω_r.

(i) $|\alpha| \leq \max\{maxr\, g, |\omega|\} = p$.

This is obvious.

Now let $\bar{\alpha}$ be α extended by two letters immediately to its left and two letters immediately to its right (if such letters to the left of α do not exist then we extend α to the left taking all remaining, if any, letters to the left of α, we proceed analogously in extending α to the right). Let $\bar{y}$ denote the contribution of $\bar{\alpha}$ to ω_s.

(ii) z is included in the image of $\bar{y}$ under h.

This is obvious.

(iii) $l = (s - r) \leq \log_2 n$.

This is obvious.

Now let u be the leftmost occurrence in z contributed (through h and ω_s) by the leftmost occurrence in α and let q be the length of the longest prefix of z that does not contain u. Then let the description of z be the triplet $des\, z = (\bar{\alpha}, l, q)$.

(iv) If $z_1, z_2 \in sub_n h(K)$ and $des\, z_1 = des\, z_2$ then $z_1 = z_2$.

This is obvious.

Now (i) through (iv) imply that $\pi_{h(K)}(n) \leq C p (\log_2 n) n$ for a positive integer C and so the theorem holds. □

Since in [2] an everywhere growing DOL language K was given such that there exists a positive real D such that $\pi_K(n) \geq D n \log_2 n$, Theorem 8 represents "the best possible" bound.

THEOREM 9: *Let K be a uniformly growing DOL language,* $K \subseteq \Delta^*$, *and let h be a* Λ*-free homomorphism on* Δ^*. *There exists a positive integer constant C such that, for every positive integer n,* $\pi_{h(K)}(n) \leq C n$.

Proof: The proof of this theorem is analogous to the proof of Theorem 8 except that we get a different upper bound for the value of l (we will use the notation from the proof of Theorem 8). Let v be such that for every $a \in \Sigma$, $|h(a)| = v \geqq 2$.

Thus $n \geqq v^l$.

On the other hand, because $|\bar{\alpha}| \leqq p+4$, we have $n \leqq (p+4) v^l \, maxr \, h$. Hence:

$$l \leqq \log_v n \leqq l + \log_v \bar{p} \qquad \text{where} \quad \bar{p} = (p+4) \, maxr \, h.$$

Thus $\log_v n - \log_v p \leqq l \leqq \log_v n$.

Consequently the set of all possible values of l for all subwords $z \in sub_n \, h(K)$ is of cardinality not greater than $(1 + \log_v \bar{p})$. Thus $\pi_{h(K)}(n) \leqq p (1 + \log_v \bar{p}) \, n$ which implies the theorem. $\square$

Since in [2] a uniformly growing DOL language K was given such that there exists a positive real D such that $\pi_K(n) \geqq D \, n$, Theorem 9 represents "the best possible" bound.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of NSF grant number MCS 79-03838. The authors are indebted to J. Karhumaki and to the referee for useful comments concerning the first version of the paper.

REFERENCES

1. A. Ehrenfeucht and G. Rozenberg, *A Limit Theorem for Sets of Subwords in Deterministic TOL Systems*, Information Processing Letters, Vol. 2, 1973, pp. 70-73.
2. A. Ehrenfeucht, K. P. Lee, and G. Rozenberg, *Subword Complexities of Various Classes of Deterministic Developmental Languages without Interactions*, Theoretical Computer Science, Vol. 1, 1975, pp. 59-76.
3. K. P. Lee, *Subwords of Developmental Languages*, Ph. D. Thesis, State University of New York at Buffalo, 1975.
4. G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems*, Academic Press, London - New York, 1980.