ON THE SUBWORD COMPLEXITY OF DOL LANGUAGES WITH A CONSTANT DISTRIBUTION

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-206-81 May 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.

ON THE SUBWORD COMPLEXITY OF DOL LANGUAGES WITH A CONSTANT DISTRIBUTION

by

A. Ehrenfeucht
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado 80309
U.S.A.

G. Rozenberg
Institute of Applied Mathematics and Computer Science,
University of Leiden
Wassenaarseweg 80, Leiden
The Netherlands.

ABSTRACT

We say that a language $K \subseteq \Sigma^*$ has a constant distribution if there exist a positive integer C and an alphabet $\Delta \subseteq \Sigma$ such that the set of letters occurring in every subword of K of length C equals Δ. The subword complexity function of K, denoted π_K, is the function of positive integers such that $\pi_K(n)$ equals the number of subwords of length n that occur in (words of) K. We prove that if K is a DOL language with a constant distribution then π_K is bounded by a linear function.

ON THE SUBWORD COMPLEXITY OF DOL LANGUAGES

INTRODUCTION

A way to understand the structure of a language K is to investigate the set of all subwords that occur in words of K (denoted as sub(K)). In particular one can count the number of words in sub(K) that are of a given length n; this number is denoted by $\pi_K(n)$. Hence π_K is the "subword complexity" function of K: for each n it yields the number of subwords of length n occurring in words of K (see, e.g., [2], [4]).

The subword complexity function turned out to be a useful tool in investigating TOL systems and its subclasses (see, e.g., [4]). The subword complexity function can be considered to measure a global feature of a TOL system: the number of subwords in a TOL language K is well defined without knowing an actual TOL system generating K. Still, it was demonstrated that the subword complexity function can "detect" local restrictions on TOL systems that is restrictions on the set of productions in a given system. In particular it turned out that generatively deterministic TOL systems generate languages with a "limited" number of subwords. Then within the subclass of generatively deterministic TOL systems (the so called DOL systems) restrictions on the length of the right-hand side of productions are reflected in further limitations on the maximal number of subwords that can be generated (see, e.g., [3]).

The number of subwords in DOL languages is also sensitive to some global restrictions. For example if (following [7] and [8]) one requires that every word in a DOL language K is square-free, i.e., it does not contain a subword of the form x x where x is a nonempty word, then for every positive integer n, $\pi_K(n) \leq D\ n \log_2 n$ for some positive integer D (see [3]).

In this paper we continue the investigation of global restrictions on a DOL language that have an effect on its subword complexity. We say that a language $K \subseteq \Sigma^*$ has a constant distribution if there exist a positive integer C and an alphabet $\Delta \subseteq \Sigma$ such that the set of letters occurring in every element

of sub(K) of length C equals Δ. We demonstrate that if K is a DOL language of constant distribution then $\pi_K(n)$ is bounded by a linear function of n. We also show an application of this result in estimating the subword complexity function of a square-free (m-free in general) DOL language over an alphabet of a limited size.

I. PRELIMINARIES AND BASIC NOTIONS

We assume the reader to be familiar with the rudiments of the theory of DOL systems (see, e.g., [4]). A DOL system G will be specified in the form $G = (\Sigma,h,\omega)$, its sequence $\omega_0,\omega_1,\ldots$ is denoted by E(G), its language is denoted by L(G) and $\underline{maxr}(G)$ denotes the maximal length of the right-hand side of a production in G.

Since the problems we consider are trivial otherwise, in this paper we deal with infinite DOL languages only.

For a word α, $\underline{alph}(\alpha)$ denotes the set of all letters occurring in α; for a language K, $\underline{alph}(K) = \bigcup_{\alpha\in K} \underline{alph}(\alpha)$. For a word α and a letter x, $\#_x(\alpha)$ denotes the number of occurrences of x in α. If K is a language then $\underline{sub}(K)$ denotes the set of all subwords (occurring in the words) of K; for a positive integer n, $\underline{sub}_n(K)$ denotes the set of all subwords of length n of K. For a language K its subword complexity, denoted π_K, is a function of positive integers such that $\pi_K(n) = \#\underline{sub}_n(K)$ (for a finite set Z, #Z denotes its cardinality).

The following is the basic notion of this paper.

Let $K \subseteq \Sigma^*$ be a language and C a positive integer constant. We say that K has a C-distribution if there exists an alphabet $\Delta \subseteq \Sigma$ such that every word $\alpha \in \underline{sub}(K)$ with $|\alpha| \geq C$ satisfies $\underline{alph}(\alpha) = \Delta$. We say that K has a constant distribution if there exists a positive integer constant C such that K has a C-distribution.

II. RESULTS

In this section we prove the main result of this paper (Theorem 1) and show some applications of it.

Theorem 1. Let L be a DOL language that has a constant distribution. Then there exists a positive integer constant Q such that $\pi_L(n) \leq Qn$ for every positive integer n.

Proof.

Let L be generated by a DOL system $H = (\Sigma, f, \rho)$ with $E(G) = \rho_0, \rho_1, \ldots$. Since L has a constant distribution, there exists a positive integer constant C such that L has a C-distribution for an alphabet $\Delta \subseteq \Sigma$. Clearly we can assume that $\Delta = \Sigma$. We simply have to start from a new axiom which equals ρ_t where t is the smallest integer i such that $|\rho_j| \geq C$ for all $j \geq i$; in this way we may loose a finite number of subwords only.

Then we slice-up (see [4]) our system H in such a way that we obtain v component DOL systems $H_1, \ldots, H_v$ where each component system $G = (\Sigma, h, \omega)$ satisfies the following three conditions. For every $x \in \Sigma$,

(i). $\underline{alph}(h(x)) = \underline{alph}(h^m(x))$ for every positive integer m,

(ii). either, for every positive integer m, $h(x) = h^m(x)$

or, for every positive integer m, $|h^{m+1}(x)| > |h^m(x)|$,

(iii). for every $y \in \Sigma$,

either $\#_y(h^m(x)) = 0$ for every positive integer m

or $\#_y(h^m(x)) = 1$ for every positive integer m

or $\#_y(h^m(x)) > 1$ for every positive integer m.

It is easily seen that such a slice-up is always possible.

Hence $L = L(H) = \bigcup_{i=1}^{m} L(H_i)$. Clearly to prove the theorem it suffices to prove that its statement holds for every component language $L(H_i)$. Thus let $G = (\Sigma, h, \omega)$ be a component system in the above slice-up of H. Let $L(G) = K$ and $E(G) = \omega_0, \omega_1, \ldots$. Let us divide all letters from Σ into stationary letters,

that is letters that satisfy the "either part" of the condition (ii) above, and growing letters, that is letters satisfying the "or part" of the condition (ii) above. Let Σ_s and Σ_g denote the subsets of Σ consisting of stationary and growing letters respectively.

The proof of the theorem goes now through a sequence of lemmas.

Lemma 1. $\Sigma_g \neq \emptyset$.

Proof.

Otherwise K would be finite. □

Corollary 1. If $\alpha \in \Sigma_s^* \cap \underline{sub}(K)$ then $|\alpha| < C$.

Proof.

Immediate from Lemma 1 and from the fact that L(G) has a C-distribution. □

Lemma 2. For every $x \in \Sigma$, either $h(x) = h^m(x)$ for every positive integer m or for every $y \in \Sigma$, $\#_y(h^m(x)) > 1$ and $|h^{m+1}(x)| > |h^m(x)|$ for every positive integer m.

Proof.

Clearly it suffices to show that for every letter $x \in \Sigma_g$ the "or" part of the statement of the lemma holds. To this aim choose m_0 to be such that $|h^{m_0}(x)| > 2C$. Thus for every $y \in \Sigma$, $\#_y(h^{m_0}(x)) > 1$ because otherwise $\underline{sub}_C(K)$ would contain a word α such that $y \notin \underline{alph}(\alpha)$; a contradiction.

The lemma follows now from the condition (iii) on the slice-up of H. □

Lemma 3. For every $x,y \in \Sigma_g$ and for every positive integer m, $|h^m(x)| < |h^m(y)|$ $\underline{maxr}$ G.

Proof.

By Lemma 2, $x \in \underline{alph}(h(y))$ and consequently $|h^m(x)| < |h^{m+1}(y)|$. But $|h^{m+1}(y)| \leq |h^m(y)|$ $\underline{maxr}$ G and the lemma follows. □

Let $\beta \in \underline{sub}(K)$ and let $i(\beta)$ be the smallest j such that ω_j contains an

occurrence of β. Then by $\underline{occ}(\beta)$ we denote the leftmost occurrence of β in $\omega_{i(\beta)}$. A subword $\alpha \in \underline{sub}(K)$ is called an $\underline{ancestor}$ of β if α has an occurrence $o(\alpha)$ in some ω_t, $0 \leq t < i(\beta)$, such that every element of $\underline{occ}(\beta)$ is contributed by an element of $o(\alpha)$ and if we omit either the leftmost or the rightmost occurrence in $o(\alpha)$ then this property does not hold anymore. An ancestor α of β is called $\underline{special}$ if it has an occurrence, referred to as the $\underline{special\ occurrence}$ (with respect to β), in some ω_t, $0 \leq t < i(\beta)$, such that

it contains an occurrence of a growing letter the contribution of which to $\omega_{i(\beta)}$ lies totally within $\underline{occ}(\beta)$ .....(1)

and moreover no occurrence of an ancestor of β in some ω_u, $0 \leq u < t$, satisfies (1).

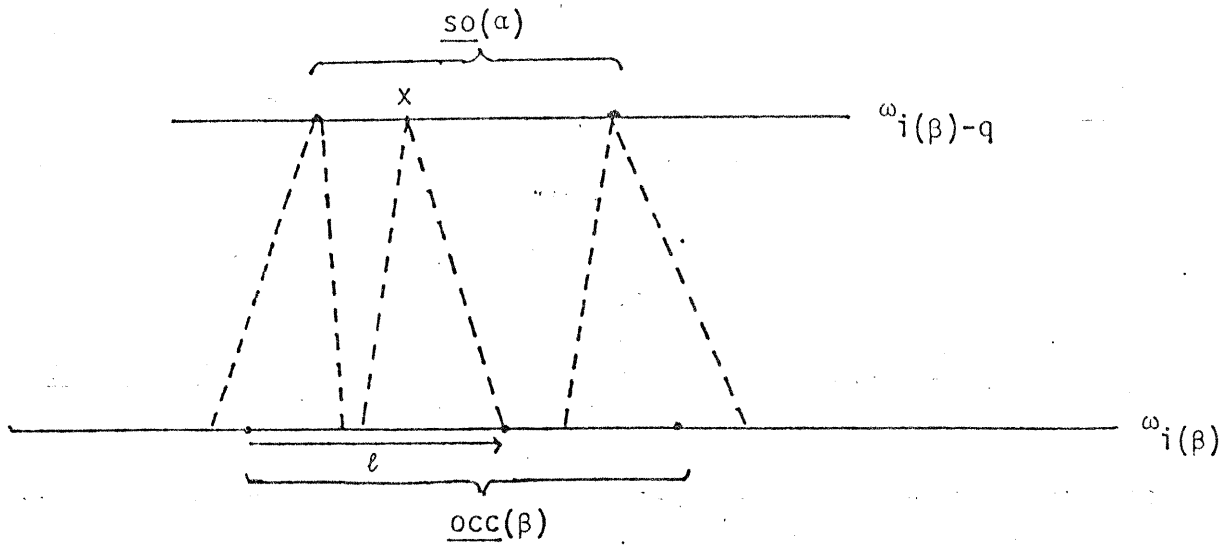
$\underline{Lemma}$ 4. There exists a positive integer n_0 such that if $\alpha \in \underline{sub}(K)$ and $|\alpha| \geq n_0$ then α has a special ancestor.

$\underline{Proof}$.

Immediate from Corollary 1. □

Let n_0 be a fixed (e.g., the smallest) constant satisfying the statement of Lemma 4 and such that $n_0 > C+1$. Let $n \geq n_0$. We will analyse now words from $\underline{sub}_n(K)$.

For a word $\beta \in \underline{sub}_n(K)$ its $\underline{category}$, denoted as $\underline{cat}(\beta)$, is a triplet (α,ℓ,q) such that α is the special ancestor of β, (its existence is guaranteed by Lemma 4), q is such that the special occurrence of α, with respect to β, denoted $\underline{so}(\alpha)$, is in $\omega_{i(\beta)-q}$ and $1 \leq \ell \leq n$ is the length of the prefix of β ending on the last occurrence in $\underline{occ}(\beta)$ contributed by the leftmost occurrence in $\underline{so}(\alpha)$ among all occurrences of growing letters. The situation can be illustrated as follows:



where x is the leftmost among all occurrences of growing letters in $\underline{so}(\alpha)$.

Lemma 5. If $\beta_1, \beta_2 \in \underline{sub}_n(K)$ and $\underline{cat}(\beta_1) = \underline{cat}(\beta_2)$ then $\beta_1 = \beta_2$.

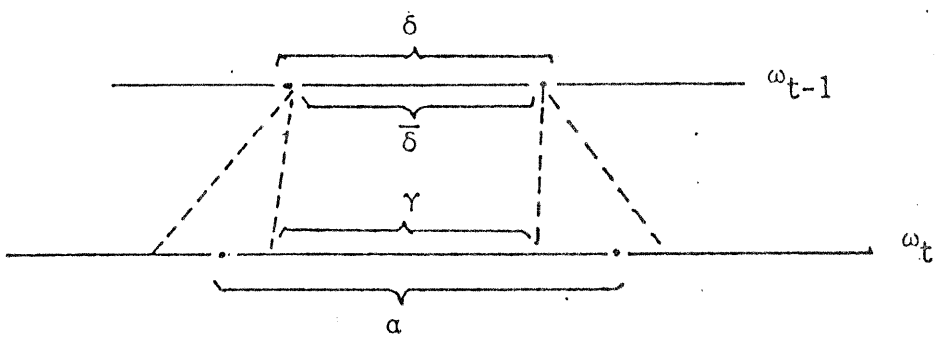
Proof.

Obvious. □

Lemma 6. There exists a positive integer constant D_1 such that the number of all special ancestors of words in $\underline{sub}_n(K)$ is not bigger than D_1.

Proof.

Let $\beta \in \underline{sub}_n(K)$ and let α be the special ancestor of β such that the special occurrence of α is in ω_t where $t > 0$. Then $|\alpha| < 3\ \underline{maxr}(G) + C$. To see this assume to the contrary that $|\alpha| \geq 3\ \underline{maxr}(G) + C$ and consider the direct ancestor δ of α (since $t > 0$ such a δ exists). We have the following situation:



Then $|\gamma| \geq \underline{maxr}(G) + C$ and so, by Corollary 1, γ contains an occurrence of a growing letter. Consequently $\bar{\delta}$ contains an occurrence of a growing letter and so δ (rather than α) must be the special ancestor of β; a contradiction.

Hence the number of all special ancestors of words in $\underline{sub}_n(K)$ bounded by D_1 which equals $\#\Sigma^{3\underline{maxr}(G)+C} + \bar{D}_1$ where $\bar{D}_1$ is the number of special ancestors with special occurrences in ω. □

Remark. Note that from the above proof it follows that if α is a special ancestor then $|\alpha| < \max\{3\ \underline{maxr}(G) + C, |\omega|\}$. We will use A to denote the constant $\max\{3\ \underline{maxr}(G) + C, |\omega|\}$.

Lemma 7. There exists a positive integer F such that for every special ancestor α (of a word from $\underline{sub}_n(K)$) and for every $1 \leq \ell \leq n$,

$\#\{q : \underline{cat}(\beta) = (\alpha,\ell,q) \text{ for some } \beta \in \underline{sub}_n(K)\} \leq F$.

Proof.

Given a positive integer t, let $\underline{min}(t) = \min\{|h^t(x)| : x \in \Sigma_g\}$ and $\underline{max}(t) = \max\{|h^t(x)| : x \in \Sigma_g\}$. Note that Lemma 3 implies that

$$\underline{max}(t) < \underline{min}(t)\ \underline{maxr}(G) \quad \ldots\ldots(2)$$

Consider now a word $\beta \in \underline{sub}_n(K)$ and let $\underline{cat}(\beta) = (\alpha,\ell,q)$. Since every growing letter in α generates in q steps a word not longer than $\underline{max}(q)$ and every stationary letter in α generates in q steps a word not longer than $\underline{maxr}(G)$, we get

$$|h^q(\alpha)| \leq |\alpha|\ \max\{\underline{max}(q), \underline{maxr}(G)\} \quad \ldots\ldots(3)$$

Since $|\beta| = n$, we also have

$$n \leq |h^q(\alpha)| \quad \ldots\ldots(4)$$

Then (3) and (4) yield

$$n \leq |h^q(\alpha)| \leq |\alpha|\ \max\{\underline{max}(q), \underline{maxr}(G)\}.$$

This combined with (2) (see also the remark following the proof of Lemma 6) yields

$$n \leq |h^q(\alpha)| < A \max\{\underline{min}(q)\ \underline{maxr}(G), \underline{maxr}(G)\} =$$

$$= A\ \underline{min}(q)\ \underline{maxr}(G) \quad \ldots\ldots(5).$$

Since α contains at least one occurrence of a growing letter, the definition of a special occurrence yields $\underline{\min}(q) \leq n$ and hence from (5) we get

$$n \leq |h^q(\alpha)| < n A \underline{\max r}(G) \quad \ldots\ldots(6)$$

Now let for $1 \leq \ell \leq n$ and a special ancestor α,

$Z_{\ell,\alpha} = \{q : \underline{cat}(\beta) = (\alpha, \ell, q)$ for some $\beta \in \underline{sub}_n(K)\}$.

Consider $m \in Z_{\ell,\alpha}$ and $m + a \in Z_{\ell,\alpha}$ where $a \geq 0$.

By Corollary 1, each subword of $|h^m(\alpha)|$ of length (C+1) contains at least one growing letter. Consequently

$$|h^{m+a}(\alpha)| > \left(\frac{|h^m(\alpha)|}{C+1} - 1\right) 2^a \quad \ldots\ldots(7)$$

From (6) and (7) we get

$$\left(\frac{|h^m(\alpha)|}{C+1} - 1\right) 2^a < |h^{m+a}(\alpha)| < n A \underline{\max r}(G)$$

and consequently

$$\left(\frac{|h^m(\alpha)|}{C+1} - 1\right) 2^a < n A \underline{\max r}(G) \quad \ldots\ldots(8)$$

From (6) and (8) we get

$$\left(\frac{n}{C+1} - 1\right) 2^a < n A \underline{\max r}(G)$$

and consequently

$$2^a < \frac{n}{\frac{n}{C+1} - 1} A \underline{\max r}(G) = \frac{1}{\frac{1}{C+1} - \frac{1}{n}} A \underline{\max r}(G) \quad \ldots\ldots(9)$$

Note that $\dfrac{1}{\left(\frac{1}{C+1} - \frac{1}{n}\right)}$ is a monotically decreasing function of n and so, because $n \geq n_0$, we get $2^a < \dfrac{1}{\frac{1}{C+1} - \frac{1}{n_0}} A \underline{\max r}(G)$

Thus $a < \log_2 \left[\left(\dfrac{1}{\frac{1}{C+1} - \frac{1}{n_0}} \right) A \underline{\max r}(G) \right]$

Consequently

$$\#Z_{\ell,\alpha} \leq \log_2 \left[\left(\frac{1}{\frac{1}{C+1} - \frac{1}{n_0}} \right) A \underline{\max r}(G) \right] + 1 \quad \ldots\ldots(10)$$

Thus if we set F equal the right-hand side of the inequality (10), the lemma holds. □

Lemma 8. Let $CAT_n = \#\{\underline{cat}(\beta) : \beta \in \underline{sub}_n(K)\}$. Then $\#CAT_n \leq B\,n$ where B is a positive integer constant.

Proof.

If $(\alpha,\ell,q) \in CAT_n$ then $1 \leq \ell \leq n$ and so, by Lemma 6 and Lemma 7, $\#CAT_n \leq D_1 F n$, where D_1 and F are the constants from the statements of Lemma 6 and Lemma 7 respectively. Thus the lemma holds. □

To complete the proof of the theorem we proceed as follows. First of all we note that we have considered $n > n_0$ only. This, however, can be done without the loss of generality because we loose a finite number of (sub)words only. Thus Lemma 5 and Lemma 8 imply that the statement of the theorem holds with L replaced by K.
Since L is a finite union of component languages, the theorem holds. □

To put Theorem 1 in a proper perspective we would like to remark that (it is easily seen that) the theorem is not true if one replaces DOL languages by e.g. context-free languages.

We end this paper by demonstrating two applications of Theorem 1. We need some terminology first.

Already around 1910 A. Thue has introduced a way of investigating a pattern of subwords in languages and (infinite) sequences (see [7] and [8]). Given a word x and a positive integer $m \geq 2$ we say that x is m-free if it does not contain a subword of the form y^m where y is a nonempty word. (If x is 2-free we say that it is square-free). A language is called m-free if it consists of m-free words only. Investigation of m-free languages and sequences forms recently a vivid research area within formal language theory (see, e.g. [1], [3], [5], [6]).

It was proved in [3] that if K is a square-free DOL language then $\pi_K(n)$ is bounded by a quadratic function of n. Moreover it was demonstrated that there exists a square-free DOL language in which $\pi_K(n)$ is of order n^2. The example given in [3] used a DOL system with 9 letters. We will show now how using Theorem 1 one can easily prove that $\pi_K(n)$ is bounded by a linear function of n if K is a square-free DOL language using 3 letters.

Corollary 2. Let K be a square-free DOL language such that #alph(K) = 3. Then there exists a positive integer constant D such that $\pi_K(n) \leq D\,n$ for every positive integer n.

Proof.

Clearly K must be a language with a 4-distribution (otherwise K cannot be square-free). Thus the result follows from Theorem 1. □

On the other hand it was proved in [3] that if K is an infinite square-free language then $\pi_K(n) \geq n$ for every positive integer n. Since it is well-known (see, e.g., [1]) that there exist square-free DOL languages over a three letter alphabet (and by the argument from the proof above such a language must have 4-distribution) we get the following ramification of Theorem 1.

Theorem 2. There exists a DOL language K that has a constant distribution and is such that $\pi_K(n) \geq n$ for every positive integer n. □

For m-free DOL languages we can prove the following result.

Corollary 3. Let K be a DOL language such that #alph(K) = 2. If there exists a positive integer constant m such that K is m-free then there exists a positive integer D such that $\pi_K(n) \leq D\,n$ for every positive integer n.

Proof.

If K is m-free then every element β of $\underline{sub}_m(K)$ is such that alph(β) = alph(K). Hence the result follows from Theorem 1. □

Acknowledgements. The authors are indebted to H.C.M. Kleijn for useful comments on the first version of this paper. The authors gratefully acknowledge the support of NSF grant MCS 79-03838.

REFERENCES

[1] Berstel, J., Sur les mots sans carré définis par un morphisme, Lecture Notes in Computer Science, Springer-Verlag, v. 71, 16-25, 1979.

[2] Ehrenfeucht, A. and Rozenberg, G., On subword complexities of homomorphic images of languages, Rev. Fr. Automat. Inform. Rech. Opér., Sér. Rouge, to appear.

[3] Ehrenfeucht, A. and Rozenberg, G., On the subword complexity of square-free DOL languages, Theoretical Computer Science, to appear.

[4] Rozenberg, G. and Salomaa, A., The mathematical theory of L systems, Academic Press, London, 1980.

[5] Salomaa, A., Morphisms on free monoids and language theory, in R. Book (ed.), Formal language theory, Academic Press, London, 1980.

[6] Salomaa, A., Jewels of language theory, Computer Science Press, to appear.

[7] Thue, A., Über unendliche Zeichenreihen, Norsk. Vid. Selsk. Skr. I Mat.-Nat. Kl., nr. 7, 1-22, 1906.

[8] Thue, A., Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, Norsk. Vid. Selsk. Skr. I Math.-Nat., nr. 1, 1-67, 1912.