ON THE SIZE OF THE ALPHABET AND THE SUBWORD COMPLEXITY OF SQUARE-FREE DOL LANGUAGES

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-207-81 June 1981

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

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

All correspondence to the second author.

ANY OPINIONS, FINDINGS, AND CONCLUSIONS OR RECOMMENDATIONS EXPRESSED IN THIS PUBLICATION ARE THOSE OF THE AUTHOR AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE NATIONAL SCIENCE FOUNDATION.

THIS MATERIAL IS BASED UPON WORK SUPPORTED BY THE NATIONAL SCIENCE FOUNDATION UNDER GRANT NO. MCS 79-Ø3838.

ABSTRACT

A word is called square-free if it does not contain a subword of the form $\alpha\alpha$ where α is a nonempty word. A language is called square-free if it consists of square-free words only. The subword complexity of a language K, denoted π_K, is a function of positive integers which for a positive integer n assigns the number of different subwords of length n occurring in words of K. It is known that if a DOL language K is square free then, for all n, $\pi_K(n) \leq r\, n \log_2 n$ for some positive integer r. We demonstrate that there exists a square-free DOL language K on four letters such that, for all n, $\pi_K(n) \geq p\, n \log_2 n$ for some positive real p. This turns out to be the best lower bound on the size of the alphabet needed for a square-free DOL language to have the number of subwords of order $n \log_n n$.

INTRODUCTION

In order to understand the structure of a language one may investigate the set of its subwords. As a first step in this direction one may take a numerical approach and simply count the number of subwords of a given length in the language. For a language K, let π_K be the function of positive integers such that $\pi_K(n)$ is the number of different subwords of length n occurring in words of K; π_K is referred to as the *subword complexity* of K. The subword complexity of DOL languages was quite extensively investigated (see, e.g., [ER1], [L] and [RS]). Among others it was demonstrated that the subword complexity of a DOL language is sensitive to various "local" restrictions on a DOL system that generates it; local restrictions mean restrictions on the set of productions - e.g., one can require that the length of the right-hand side of every production is longer than 1.

Another approach to investigate the set of subwords of a language is to consider structural restrictions on their distribution in words. Thus following [T] one calls a word *square-free* if it does not contain a subword of the form $\alpha\alpha$ where α is a nonempty word; a language is called square-free if it consists of square-free words only. Square-free DOL languages are a subject of active investigation, see, e.g., [B1], [B2], [S1] and [S2]. It was demonstrated ([ER1]) that if K is a square-free DOL language then, for all n, $\pi_K(n) \leq r\, n \log_2 n$ where r is a positive integer (one should contrast this with the fact that there exist DOL languages which have the subword complexity function of order n^2). In the same paper it was demonstrated that there exists a DOL language K such that, for all n, $\pi_K(n) \geq p\, n \log_2 n$ for a positive real p. However, this particular language is over 9 letters. Hence the question arises whether the "$n \log_2 n$" remains "reachable" in square-free DOL languages using less than 9 letters. It was shown in [ER2] that if a square-free DOL language K is over a three letter alphabet, then for all n, $\pi_K(n) \leq r\,n$ for a positive integer n. In this paper we show that four letters suffice to achieve the order of $n \log_2 n$ subwords of length n in a DOL square-free language. In this sense this paper establishes the precise boundary between order n and order $n \log_2 n$ square-free DOL languages.

We assume the reader to be familiar with the basic theory of DOL systems and languages - see, e.g., [RS].

1. PRELIMINARIES

We use mostly standard language-theoretic notation and terminology (see, e.g., [RS]). Perhaps the following points require an additional explanation.

$\emptyset$ denotes the empty set, N^+ denotes the set of positive integers and, for a finite set A, #A denotes the cardinality of A. We consider finite alphabets only. Λ denotes the empty word, $|w|$ denotes the length of a word w, $alph(w)$ the set of letters occurring in w and, for a letter x, $\#_x w$ denotes the number of occurrences of x in w. For $n \in N^+$ and a word w the prefix of w of length n, denoted $pref_n(w)$, is defined by

$$pref_n(w) = \begin{cases} t_1 \dots t_n & \text{if } w = t_1 \dots t_r,\ r \geq n, \text{ where } t_1, \dots, t_r \text{ are letters,} \\ w & \text{if } |w| < n, \end{cases}$$

similarly the suffix of w of length n, denoted $suf_n(w)$, is defined by

$$suf_n(w) = \begin{cases} t_n \dots t_1 & \text{if } w = t_r \dots t_1,\ r \geq n, \text{ where } t_1, \dots, t_r \text{ are letters,} \\ w & \text{if } |w| < n. \end{cases}$$

We will also use the notation $first(w)$ to denote $pref_1(w)$ and $last(w)$ to denote $suf_1(w)$. If a word w is a subword of a word z then we write $w \sqsubseteq z$; $sub(z)$ denotes the set of all subwords of z and for a language K,

$$sub(K) = \bigcup_{z \in K} sub(z).$$

The *subword complexity of a language* K, denoted as π_K, is the function from N^+ into N^+ defined by $\pi_K(n) = \#\{w \in sub(K) : |w| = n\}$.

A word w is called *square-free* if, for no nonempty word α, $\alpha\alpha$ is a subword of w. The following obvious to prove result will be useful in the sequel. First, we need the following notion.

Let w, z be nonempty words such that $w \sqsubseteq z$. We say that w is *unique in* z if for all words z_1, z_2, z_3, z_4, if $z = z_1 w z_2$ and $z = z_3 w z_4$ then $z_1 = z_3$ and $z_2 = z_4$.

Lemma 1.1. Let w and z be nonempty words such that w is unique in z. Let α be a nonempty word such that $\alpha\alpha \sqsubseteq z$. Then w is not a subword of α. □

For a homomorphism $h : \Sigma^* \to \Sigma^*$, $minr(h) = \min\{|h(x)| : x \in \Sigma\}$ and $maxr(h) = \max\{|h(x)| : x \in \Sigma\}$.

If $\Delta \subseteq \Sigma$ where Σ is an alphabet then $pres_{\Delta,\Sigma}$, or simply $pres_\Delta$ if Σ is understood, denotes the homomorphism defined by $pres_{\Delta,\Sigma}(x) = x$ for $x \in \Delta$ and $pres_{\Delta,\Sigma}(x) = \Lambda$ for $x \in \Sigma \setminus \Delta$.

We say that h is *square-free* if h(z) is square-free for every square-free $z \in \Sigma^*$. The following result from [BEM] will be useful in the sequel.

Theorem 1.1. Let Σ be an alphabet and let h be a homomorphism of Σ^*. If

(i). h(z) is square-free for every square-free word $z \in \Sigma^*$ such that $|z| \leq 3$, and

(ii). if $h(x) \sqsubseteq h(y)$ implies $x = y$ for all $x, y \in \Sigma$,

then h is square-free. □

A DOL *system* will be specified as a triplet $G = (\Sigma, g, w)$ where Σ is its alphabet, g is its homomorphism and w is the axiom of G. Then E(G) denotes the sequence of G and L(G) denotes the language of G.

2. RESULTS

In this section we investigate the subword complexity of square-free DOL languages over a four letter alphabet. Our first result provides a method to construct a square-free DOL language such that the number of subwords of length n in it is of order $n \log_2 n$.

Theorem 2.1. Let Δ and Σ be alphabets where $\Delta = \{a, b, c,\}$ and $\Sigma = \Delta \cup \{d\}$ with $d \notin \Delta$. Let $h : \Delta^* \to \Delta^*$ be a square-free homomorphism and let $w \in \Delta^+$ be such that

(C1). $\mathit{minr}(h) \geq 3$,

(C2). for every $x \in \Delta$, $\mathit{first}(h(x)) = a$ and $\mathit{last}(h(x)) = b$,

(C3). for every $x, y \in \Delta$, $h(x) \sqsubseteq h(y)$ implies $x = y$,

(C4). the word bcwca is square free and

(C5). $|c\,w\,c| \geq \mathit{maxr}(h)$.

Let $g : \Sigma^* \to \Sigma^*$ be the homomorphism defined by: $g(x) = h(x)$ for $x \in \Delta$ and $g(d) = d\,c\,d\,t_1\,d\,t_2 \ldots d\,t_\ell\,d\,cd$ where $w = t_1 \ldots t_\ell$, $\ell \geq 1$ and $t_1, \ldots, t_\ell \in \Delta$. Let $G = (\Sigma, g, d\,a\,b\,c\,d)$.

Then $L(G)$ is square-free and there exists a positive real p such that $\pi_k(n) \geq p\,n \log_2 n$ for every $n \in N^+$.

Proof:

The proof of this theorem goes through a sequence of lemmas.

Lemma 2.1. If $z \in \Sigma^*$, z is square-free and z is such that $\#_d(z) = 1$ then $g(z)$ is square-free.

Proof of Lemma 2.1:

Let $z = z_1\,d\,z_2$ where $z_1, z_2 \in \Delta^*$ and let $\beta = g(z_1\,d\,z_2)$. Assume to the contrary that, for some $\alpha \neq \Lambda$, $\alpha\alpha \sqsubseteq \beta$. Since h is square-free and $g(z_1) = h(z_1)$, $g(z_2) = h(z_2)$ it must be that $d \in \mathit{alph}(\alpha)$.

Clearly, (see the definition of $g(d)$), if $|g(z_1)| \geq 2$ then $\mathit{suf}_2(g(z_1))d$

is unique in β and if $|g(z_2)| \geq 2$ then $d\,pref_2(g(z_2))$ is unique in β. Consequently, by Lemma 1.1, $\alpha\alpha \sqsubseteq last(g(z_1))\, g(d)\, first(g(z_2))$. Since d^2 is not a subword of $g(d)$ this implies that $pres_\Delta(last(g(z_1))\, g(d)\, first(g(z_2)))$ is not square-free. Since (C.2) implies that $last(g(z_1)) = b$ if $z_1 \neq \Lambda$ and $first(g(z_2)) = a$ if $z_2 \neq \Lambda$, $pres_\Delta(last(g(z_1))\, g(d)\, first(g(z_2)))$ is a subword of $b\,c\,w\,c\,a$. Thus $b\,c\,w\,c\,a$ is not square-free which contradicts the assumption (C4). Consequently, $\beta = g(z)$ is square-free and Lemma 2.1 holds. □

Lemma 2.2. For every $x \in \Delta$, $g(d\,x\,d)$ is square-free.

Proof of Lemma 2.2:

Assume to the contrary that, for some $\alpha \neq \Lambda$, $\alpha\alpha \sqsubseteq \beta$ where $\beta = g(d\,x\,d)$. Then Lemma 2.1 implies that neither $\alpha\alpha \sqsubseteq g(d\,x)$ nor $\alpha\alpha \sqsubseteq g(x\,d)$. However, (C1) implies that $|g(x)| = |h(x)| \geq 3$ and both, $d\,pref_2(g(x))$ and $suf_2(g(x))d$ are unique in β. Thus, by Lemma 1.1 we get a contradiction. Hence β must be square-free which concludes the proof of Lemma 2.2. □

Lemma 2.3. For all $x, y \in \Sigma$, if $g(x) \sqsubseteq g(y)$ then $x = y$.

Proof of Lemma 2.3:

If $x, y \in \Delta$ then $g(x) = h(x)$ and $g(y) = h(y)$ and so the lemma follows from condition (C3).

If $x \in \Delta$ and $y = d$ then (C1) and the definition of g imply that $g(x)$ is not a subword of $g(y)$. If $x = d$ and $y \in \Delta$ then $g(x)$ is not a subword of $g(y)$ because $d \in alph\, g(x)$ and $d \notin alph\, g(y)$. Hence Lemma 2.3 holds. □

Lemma 2.4. g is square-free.

Proof of Lemma 2.4:

Let $z \in \Sigma^*$ be such that $|z| \leq 3$ and z is square-free. Consider $g(z)$.

If $\#_d(z) = 0$ then $g(z) = h(z)$ and so $g(z)$ is square-free.

If $\#_d(z) = 1$ then Lemma 2.1 implies that $g(z)$ is square-free.

If $\#_d(z) = 2$ then z must be of the form $d\,x\,d$, where $x \in \Delta$. Hence Lemma 2.2 implies that $g(z)$ is square-free.

Consequently, $g(z)$ is always square-free. Consequently Lemma 2.3 and Theorem 1.1 imply that g is square-free. Hence Lemma 2.4 holds. □

Since $d\,a\,b\,c\,d$ is square-free, Lemma 2.4 implies that $L(G)$ is square-free and so the first part of the conclusion of Theorem 2.1 holds.

Now we proceed to estimate the subword complexity of $L(G)$.

Let $maxr\,(h) = r$ and $\#_d g(d) = s$.

Lemma 2.5 $s > r$.

Proof of Lemma 2.5:

From the definition of $g(d)$ it follows that $\#_d g(d) = |c\,w\,c| + 1$ and (C5) implies that $|c\,w\,c| \geq r$. Hence the result holds. □

Let $E(G) = \omega_0, \omega_1, \ldots$. Clearly for $k \geq 0$ $\omega_k = g^k(d)\,g^k(a\,b\,c)\,g^k(d)$. Obviously the following result holds.

Lemma 2.6. For every $k \geq 1$, $|g^k(d)| > s^k$ and $|g^k(a\,b\,c)| \leq 3\,r^k$. □

Let for $n \geq 1$,

$Z_n = \{k : |g^k(a\,b\,c)| \leq \frac{n}{2}$ and $|g^k(d)| \geq n\}$ and

$Z_n' = \{k : 3\,r^k \leq \frac{n}{2}$ and $s^k \geq n\}$.

Lemma 2.7. For every $n \geq 1$, $Z_n' \subseteq Z_n$ and if $k \geq 1$ is such that

$$\frac{\log_2 n}{\log_2 s} \leq k \leq \frac{\log_2 n - \log_2 6}{\log_2 r}$$

then $k \in Z_n'$.

Proof of Lemma 2.7:

The first part of the statement follows from Lemma 2.6. The second part of the statement follows from the definition of Z_n'. □

Lemma 2.8. For every $n \in N^+$, $\pi_{L(G)}(n) \geq \frac{n}{2} \# Z_n'$.

Proof of Lemma 2.8:

For $k \in Z_n'$ let P_k be the set of all these subwords of length n of ω_k that contain $g^k(a\,b\,c)$. From the definition of Z_n', from Lemma 2.6 and from the fact that $last(g^k(d)) = d = first(g^k(d))$ while $g^k(a\,b\,c) \in \Delta^*$ it follows that $\#P_k \geq \frac{n}{2}$. On the other hand, because $g^k(a\,b\,c)$ is strictly growing (with the growth of k) it is clear that $P_k \cap P_\ell = \emptyset$ if $k \neq \ell$. Hence the lemma follows. □

Now we complete the proof of the theorem as follows. Clearly from *Lemma* 2.7 it follows that

$$\#Z_n' \geq \frac{\log_2 n - \log_2 6}{\log_2 r} - \frac{\log_2 n}{\log_2 s} - 2 = e \log_2 n - m,$$

where $e = \frac{1}{\log_2 r} - \frac{1}{\log_2 s}$ and $m = \frac{\log_2 6}{\log_2 r} + 2$.

Note that from Lemma 2.5 it follows that $e > 0$.

Thus Lemma 2.8 implies that

$$\pi_{L(G)}(n) \geq \frac{n}{2}(e \log_2 n - m) \quad \ldots\ldots (1)$$

Note that

$$\frac{e}{2}\log_2 n - m \geq 0 \text{ for every } n \geq n_0 = 2^{\frac{2m}{e}} \quad \ldots\ldots (2)$$

and consequently (add $\frac{e}{2}\log_2 n$ to both sides of inequality (2))

$$e \log_2 n - m \geq \frac{e}{2}\log_2 n \text{ for every } n \geq n_0. \quad \ldots\ldots (3)$$

From (3) it follows that

$\pi_{L(G)}(n) \geq \frac{e}{4} n \log_2 n$ for every $n \geq n_0$ ....(4)

On the other hand $\dfrac{n \log_2 n}{n_0 \log_2 n_0} < 1$ for $n < n_0$ and so, note that $e < 1$, we have

$\pi_{L(G)}(n) \geq \dfrac{e}{4 n_0 \log_2 n_0} n \log_2 n$ for every $n < n_0$ ....(5)

Then (4), (5) and the definition of n_0 yield

$\pi_{L(G)}(n) \geq p\, n \log n$ for every $n \in N^+$,

where $p = \dfrac{e^2}{8m 2^{\frac{2m}{e}}}$.

This concludes the proof of the second part of the conclusion of the theorem. □

Now using Theorem 2.1 we can exhibit a square-free DOL language over a four letter alphabet which has the number of subwords of length n of order $n \log_2 n$.

Theorem 2.2. There exists an infinite DOL language $K \subseteq \Sigma^*$ such that $\#\Sigma = 4$, K is square-free and there exists a positive real p such that $\pi_K(n) \geq p\, n \log_2 n$ for all $n \in N^+$.

Proof.

Let $h : \{a, b, c\}^* \to \{a, b, c\}^*$ be the homomorphism defined by $h(a) = a b c a b$, $h(b) = a c a b c b$ and $h(c) = a c b c a c b$. It is proved in [T] that h is square-free (see also Corollary 1.1 in [BEM]).

Let $w = a\,b\,a\,c\,b$ and let $g : \{a, b, c, d\}^* \to \{a, b, c, d\}^*$ be the homomorphism defined by $g(x) = h(x)$ for $x \in \{a, b, c\}$ and $g(d) = d\,c\,d\,a\,d\,b\,d\,a\,d\,c\,d\,b\,d\,c\,d$. It is easily seen that h, w, g satisfy the assumptions of Theorem 2.1. Consequently, by Theorem 1.1, $K = L(G)$ where $G = (\{a, b, c, d\}, g, d\,a\,b\,c\,d)$ satisfies the statement of the theorem. □

To put the above result in a proper perspective we recall now two results (the first one is from [ER1] and the second one is from [ER2].

Theorem 2.3. If K is a square-free DOL language then there exists an $r \in N^+$ such that, for all $n \in N^+$, $\pi_K(n) \leq r\,n \log_2 n$. □

Theorem 2.4. If K is a square-free DOL language, $K \subseteq \Sigma^*$ where $\#\Sigma = 3$ then there exists an $r \in N^+$ such that, for all $n \in N^+$, $\pi_K(n) \leq r\,n$. □

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NSF grant MCS 79-03838.

REFERENCES

[BEM] Bean, D. R., Ehrenfeucht, A. and McNulty, G. F., Avoidable patterns in strings of symbols, *Pacific Journal of Mathematics*, 85, 261-294.

[B1] Berstel, J., Sur les mots sans carré défins par un morphisme, *Lecture Notes in Computer Science*, 71, 16-25, 1979.

[B2] Berstel, J., Mots sans carré et morphismes itérés, Univ. Paris 7, Institut de Programmation, Techm. Rep. 78-42, 1978.

[ER1] Ehrenfeucht, A. and Rozenberg, G., On the subword complexity of square-free DOL languages, *Theoretical Computer Science*, to appear.

[ER2] Ehrenfeucht, A. and Rozenberg, G., On the subword complexity of DOL languages with a constant distribution, Dept. of Computer Science, University of Colorado at Boulder, Tech. Rpt. CU-CS-206-81, 1981.

[L] Lee, K. P., Subwords of developmental languages, Ph.D. thesis, State University of New York at Buffalo, 1975.

[RS] Rozenberg, G. and Salomaa, A., *The mathematical theory of L systems*, Academic Press, London-New York, 1980.

[S1] Salomaa, A., *Jewels of formal languages*, Computer Science Press, to appear.

[S2] Salomaa, A., Morphisms and language theory, in: R. Book (ed.), *Formal Languages*, Academic Press, New York, London, 1980.

[T] Thue, A., Über die gegenseitigen Lage gleicher Teile gewisser Zeichenreilren, *Norske Vid. Selsk. Skr., I Mat. Nat. KL., Christiania*, 1, 1-67, 1912.

/

ON THE SIZE OF THE ALPHABET AND THE SUBWORD COMPLEXITY OF SQUARE-FREE DOL LANGUAGES

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-207-81 June 1981

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

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

All correspondence to the second author.