Fall 9-1-1982

On the Subword Complexity of m-Free DOL Languages ; CU-CS-232-82

Andrzej Ehrenfeucht
University of Colorado Boulder

Grzegorz Rozenberg
University of Leiden

ON THE SUBWORD COMPLEXITY OF m-FREE DOL LANGUAGES

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-232-82 September, 1982

*Department of Computer Science University of Colorado at Boulder, Boulder, Colorado 80309

**Institute of Applied Mathematics and Computer Science, University of Leiden, Leiden, The Netherlands

All correspondence to second author.

This research was supported by NSF grant MCS 79-03838.

ON THE SUBWORD COMPLEXITY OF m-FREE DOL LANGUAGES

by

A. Ehrenfeucht*

and

G. Rozenberg**

*A. Ehrenfeucht
Computer Science Dept.
Campus Box 430
University of Colorado
Boulder, Colorado 80309 USA

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

All correspondence to the second author.

ABSTRACT

A word is called m-*free* ($m \geq 2$) if it does not contain a subword of the form x^m where x is a nonempty word. A language is called m-*free* if it consists of m-free words only. The *subword complexity* of a language K, denoted π_K, is a function of positive integers which to each 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 m-free for some $m \geq 2$, then, for all n, $\pi_K(n) \leq q\, n \log_2 n$ for some positive integer q. We demonstrate that there exists a 3-free DOL language K on three letters such that, for all $n \geq n_0$, $\pi_K(n) \geq r\, n \log_2 n$ for some positive real r and a positive integer n_0. We also demonstrate that if $m \geq 3$ and K is a m-free DOL language on two letters, then, for all n, $\pi_K(n) \leq p\, n$ for some positive integer p.

INTRODUCTION

The investigation of subwords of (words in a) formal language constitutes an important aspect of the investigation of the combinatorial structure of formal languages. Such an investigation may concern more structural aspects (as initiated in [T], see also [B] and [S]) or more numerical aspects (see, e.g., [L] and [ER1]) of this topic.

Recently these two trends were combined togehter in the framework of DOL languages (see, e.g., [ER2], [ER3] and [R]). It was demonstrated that the subword complexity of a DOL language (which was designed as a "numerical measure") is very sensitive to various structural restrictions, in particular to Thue-type restrictions on the repetition of subwords in words of a language.

This note is concerned with the influence of the size of the alphabet of a so-called m-free DOL language on its subword complexity. In particular we prove that the transition from the two-letter alphabet to the three-letter alphabet corresponds to the transition from order n to order $n \log_2 n$ subword complexity of m-free languages DOL where $m \geq 3$. Since the boundary for such a transition is already known in the case of 2-free (square free) DOL languages, this node rounds off this particular aspect of research concerning the structure of subwords in DOL languages.

PRELIMINARIES

We assume the reader to be familiar with the basic theory of DOL systems (see, e.g., [RS]). We will use the standard notation and terminology concerning DOL systems (as used in [RS]).

Perhaps recalling the following notation and terminology will facilitate the reading of this note.

N^+ denotes the set of positive integers. For a set A, #A denotes its cardinality.

For a word x, $|x|$ denotes its length and $alph(x)$ denotes the set of letters occurring in x.

For words x, y, x is a *subword* of y, written x *sub* y, if $y = y_1 \times y_2$ for some words y_1, y_2; $sub(y)$ denotes the set of subwords of y and, for $n \geq 0$, $sub_n(y)$ denotes the set of subwords of length n occurring in y. (Subwords are sometimes referred to as *segments* or *factors*). For a word y and $m \in N^+$, $m \geq 2$, we say that y is m-*free* if for each nonempty word x, $x^m \notin sub(y)$. 2-free words are referred to also as *square free* and 3-free words are referred toalso as *cube free*.

For a language K, $alph(K) = \bigcup_{x \in K} alph(x)$, $sub(K) = \bigcup_{x \in X} sub(x)$ and, for each $n \geq 0$, $sub_n(K) = \bigcup_{x \in X} sub_n(x)$. For $m \geq 2$, a language K is said to be

m-*free* if it consists of m-free words only.

The *subword complexity* of a language K, denoted π_K, is a function of N^+ such that, for each $n \in N^+$, $\pi_K(n) = \#sub_n(K)$.

We close this section by recalling the following result from [T] (see also [S]); it will be useful in the proof of our main result.

Proposition 1. Let $G_0 = (\{a,b\}, g_0, a)$ where $g_0(a) = ab$ and $g_0(b) = ba$. Then $L(G_0)$ is cube free. □

RESULTS

Let $H = (\{a,b,c\}, h, c)$ be the DOL system where $h(a) = ab$, $h(b) = ba$ and $h(c) = cacbc$.

Lemma 1. L(H) is cube free.

Proof.

A word $y \in sub(L(H))$ is called a *block* if $y = c\,x\,c$ for some $x \in \{a,b\}^*$. For a block y its *age*, denoted *age*(y), is defined by $age(y) = \log_2(|y| - 2)$. If y_1, y_2 are blocks such that $age(y_1) > age(y_2)$ then we say that y_1 is *older* than y_2.

Note that it follows directly from the definition of h that each block is of length at least 3 and the age of each block is a nonnegative integer.

A block y is called an a-*block* if $y = cay_1$ for some y_1; otherwise y is called a b-*block*. If y_1, y_2 are blocks such that either both y_1 and y_2 are a-blocks or both y_1 and y_2 are b-blocks, then y_1, y_2 are *similar* (blocks).

Claim 1. If $uy_1 \, v \, y_2 \, z \in L(G)$ where y_1, y_2 are similar blocks such that $age(y_1) = age(y_2)$, then v contains a block older than y_1 (and y_2).

Proof of Claim 1.

Consider $E(G) = \omega_0, \omega_1, \ldots$ and let e be such that $\omega_e = u\, y_1\, v\, y_2\, z$. An occurrence of c in ω_k, for some $0 \leq k < e$, is called the *real ancestor of* the given occurrence of y_1 (y_2 respectively) if

(1) this occurrence of c contributes the given occurrence of y_1 (y_2 respectively) and moreover

(2) if an occurrence of c in ω_j, for some $0 \leq j \leq e$, contributes the given occurrence of y_1 (y_2 respectively), then $j < k$.

Since y_1, y_2 are similar blocks of the same age, from the definition of h it follows that, for some $1 \leq \ell < e$, ω_ℓ contains two different occurrences of e such that one of them is the real ancestor of the given occurrence of y_1 and the other one is the real ancestor of the given occurrence of y_2. Since $|h(a)| = |h(b)| = 2$, this implies that v contains a block older than y_1 (and y_2).

Thus Claim 1 holds. □

Now we continue the proof of Lemma 1 as follows. Assume that for some $i \geq 1$ and some $x \in \{a,b,c\}^+$, xxx *sub* ω_i. There are two possibilities.

(i) $c \notin alph(x)$. Since this directly contradicts Proposition 1, this case is impossible.

(ii) $c \in alph(x)$. Clearly, in this case xxx must contain at least 3 occurrences of c and consequently xxx must contain a block. Let

$y = c\bar{y}c$ for some $\bar{y} \in \{a,b\}^+$ be a block contained in xxx such that no other block of xxx is older than y.

Thus xxx = uyz for some $u, z \in \{a,b,c\}^*$. Clearly, it cannot be that x *sub* $\bar{y}$ (as otherwise $c \notin alph(x)$). Thus y *sub* xx and consequently xxx contains two different (disjoint) occurrences of y. Hence by Claim 1, xxx contains a block older than y; a contradiction.

Thus all words of L(H) are cube-free and Lemma 1 holds. □

Theorem 1. There exists a cube-free DOL language such that $alph(K) = 3$ and there exists a positive real q and a $r \in N^+$ such that $\pi_K(n) \geq q\, n \log_2 n$ for all positive integers $n \geq r$.

Proof.

Let G = ({a,b,c}, g, cac) be the DOL system where g(a) = ab, g(b) = ba and g(c) = cacbc. Let K = L(G). Obviously $alph(K) = 3$.

(1) K is cube free. This follows directly from Lemma 1.

(2) $\pi_K(n) \geq q\, n \log_2 n$ for all positive integers $n \geq r$ for some $r \in N^+$.

To prove this we proceed as follows.

Let for a word $z \in \{a,b,c\}^+$:

$max_{a,b}(z) = \max\{|u| : u \in sub(z) \cap \{a,b\}^+\}$,

$MAX_{a,b}(z) = \{u : u \in sub(z) \cap \{a,b\}^+ \text{ and } |u| = max_{a,b}(z)\}$ and let

$$tag(z) = \begin{cases} max_{a,b}(z) & \text{if z contains } \textit{precisely one} \text{ occurrence} \\ & \text{of } \textit{precisely one} \text{ subword from } MAX_{a,b}(z), \\ 0 & \text{otherwise.} \end{cases}$$

Let $E(G) = \rho_0, \rho_1, \ldots$; clearly, for each $i \geq 1$, $tag(\rho_i) = 2^i$.

Let for each $n \in N^+$, $Q_n = \{k \in N^+ : 2^k \leq \frac{n}{2} \text{ and } 3^k \geq n\}$.

Claim 2. For each $n \in N^+$, $\#Q_n \geq \left(1 - \frac{1}{\log_2 3}\right) \log_2 n - 3$.

Proof of Claim 2.

Let $n \in N^+$. For each $k \in Q_n$, $k \leq \log_2 n - 1$ and $k \geq \frac{\log_2 n}{\log_2 3}$. Hence

$$\#Q_n \geq (\log_2 n - 1) - \frac{\log_2 n}{\log_2 3} - 2 = \left(1 - \frac{1}{\log_2 3}\right) \log_2 n - 3.$$

Thus Claim 2 holds. □

Claim 3. For each $n \in N^+$ and each $k \in Q_n$, $sub_n(K)$ contains at least $\frac{n}{2}$ words z such that $tag(z) = 2^k$.

Proof of Claim 3.

Let $n \in N^+$ and let $k \in Q_n$. Consider ρ_k. Clearly $\rho_k = h^k(c)h^k(a)h^k(c)$. Clearly $h^k(b) > 3^k$ and $h^k(a) = 2^k$. Thus ρ_k contains at least $\frac{n}{2}$ subwords containing $h^k(a)$ as a subword.

Thus Claim 3 holds. □

From Claim 3 it follows that, for each $n \in N^+$, $\pi_K(n) \geq \frac{n}{2} \#Q_n$. Thus by Claim 2 we get

$$\pi_K(n) \geq \frac{n}{2} \#Q_n \geq \frac{n}{2} \left[\left(1 - \frac{1}{\log_2 3}\right) \log_2 n - 3\right] .$$

It is easy to calculate that for $s = \frac{3}{1 - \frac{1}{\log_2 3}}$

$$\pi_K(n) \geq \frac{1}{4} \left(1 - \frac{1}{\log_2 3}\right) n \log_2 n$$

for all $n \geq 2^{2s}$.

Hence the theorem holds. □

Also over a two-letter alphabet one can have infinite DOL languages that are n-free for any $n \geq 3$ (see [T] and [S]). We will

demonstrate now that these languages have a "poorer" subword complexity than their counterparts over a three-letter alphabet.

Let K be a language and let $C \in N^+$. K has a C-*distribution* if there exists an alphabet Δ such that $alph(x) = \Delta$ for each $x \in sub_C(K)$. If K has a C-distribution for some $C \in N^+$ then we say that K has a *constant distribution*.

The following result is given in [ER3].

Proposition 2. If a DOL language K has a constant distribution, then there exists a $q \in N^+$ such that $\pi_K(n) \leq qn$ for every $n \in N^+$. □

Using the above result we can prove the following theorem.

Theorem 2. Let $m \geq 3$ and let K be a m-free DOL language such that $alph(K) = 2$. Then there exists a $q \in N^+$ such that $\pi_K(n) \leq qn$ for all $n \in N^+$.

Proof.

Since K is m-free, $alph(x) = alph(K)$ for each $x \in sub_m(K)$. Thus K has a m-distribution and so the theorem follows from Proposition 2. □

Hence, theorems 1 and 2 provide the precise boundary between order n and order $n \log_2 n$ m-free DOL languages.

To put these results in a proper perspective let us recall now results establishing such a boundary in the case of square free DOL languages. (The first result is from [ER3] and the second from [ER4]).

Proposition 3. Let K be a square free DOL language such that $alph(K) = 3$. Then there exists a $q \in N^+$ such that $\pi_K(n) \leq qn$ for all $n \in N^+$. □

Proposition 4. There exists a square free DOL language such that $alph(K) = 4$ and there exists a positive real q such that $\pi_K(n) \geq q\ n \log_2 n$ for all $n \in N^+$. □

Thus Theorem 1, Theorem 2, Proposition 1 and Proposition 2 provide a full picture of the influence of the size of an alphabet on the subword complexity of m-free DOL languages for all $m \geq 2$.

Finally let us recall (see [ER2]) that $n \log_2 n$ constitutes an upper bound (on the subword complexity) for all m-free DOL languages.

Proposition 5. Let $m \geq 2$ and let K be a m-free DOL language. There exists a positive integer q such that $\pi_K(n) \leq q\ n \log_2 n$ for all $n \in N^+$. □

ACKNOWLEDGMENTS

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

REFERENCES

[B] J. Berstel, Sur les mots sans carre definis par un morphisme, 1979, *Springer Lecture Notes in Computer Science*, v. 71, 16-25.

[ER1] A. Ehrenfeucht and G. Rozenberg, On subword complexities of homomorphic images of languages, *RAIRO Informatique Theorique*, to appear.

[ER2] A. Ehrenfeucht and G. Rozenberg, On the subword complexity of square free DOL languages, 1981, *Theoretical Computer Science*, v. 16, 25-32.

[ER3] A. Ehrenfeucht and G. Rozenberg, On the subword complexity of DOL languages with a constant distribution, 1981, *Information Processing Letters*, v. 13, 108-114.

[ER4] A. Ehrenfeucht and G. Rozenberg, On the size of the alphabet and the subword complexity of square free DOL languages, *Semigroup Forum*, to appear.

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

[R] G. Rozenberg, On subwords of formal languages, 1981, *Lecture Notes in Computer Science*, v. 117, 328-333.

[RS] G. Rozenberg and A. Salomaa, *The mathematical theory of L systems*, 1981, Academic Press, London, New York.

[S] A. Salomaa, *Jewels of formal language theory*, 1981, Computer Science Press.

[T] A. Thue, Uber unendliche Zeichenreihen, 1906, *Norske Vid., Selsk., Skr., I Mat. Nat. Kl. Christiania*, v. 7, 1-22.