ON THE SUBWORD COMPLEXITY OF
LOCALLY CATENATIVE DOL LANGUAGES

by
A. Ehrenfeucht*

and
G. Rozenberg**

CU-CS-231-82 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.
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 IN PART BY THE NATIONAL SCIENCE FOUNDATION UNDER GRANT NO. MCS 74-03838.
ABSTRACT

The subword complexity of language $K$, denoted $\pi_K$, is the function of positive integers such that $\pi_K(n)$ equals the number of subwords of length $n$ that occur in (words of) $K$. It is proved that if $K$ is a locally catenative DOL language then $\pi_K$ is bounded by a linear function.

INTRODUCTION

The investigation of the structure of subwords in words (of a formal language) constitutes "an access" to the understanding of the structure of a language. In particular by counting the number of subwords of each length in a language one gets a measure of "subword complexity" of the language (see, e.g., [L], [ER1], [R]).

The subword complexity approach has turned out to be very useful in the investigation of (deterministic variants) of L languages (see, e.g., [L]). In particular it was demonstrated that subword complexity is sensitive to various "local" and "global" restrictions on DOL systems (see, e.g., [R], [ER2] and [ER1]). Thus in [ER1] it was shown that the classical (global) Thue-restriction to square-free languages (see [T]) reflects in restricting the subword complexity of a (square-free) DOL language to no more than order of $n \log n$ (where $n$ is the length of subwords considered).

In this note we consider one of the first (classic?) global restrictions considered in the theory of DOL systems: local catenativity (see, e.g., [RS]). We demonstrate that this restriction reflects itself in a rather drastic restriction on the subword
complexity of DOL systems satisfying it: the subword complexity of a locally catenative DOL system is bounded by an order \( n \) function and so it is "as low as possible".

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

PRELIMINARIES AND BASIC DEFINITIONS

We use the standard notation and terminology concerning DOL systems (see [RS]).

For the purpose of this note it is convenient to use the following terminology: if \( G \) is a \((i_1, \ldots, i_k)\)-locally catenative DOL system such that \( i_1, \ldots, i_k \) are relatively prime (that is \( \gcd(i_1, \ldots, i_k) = 1 \)), then we say that \( G \) is a \textit{relatively prime locally catenative DOL system} (and \( L(G) \) is a \textit{relatively prime locally catenative DOL language}).

Let \( C \) be a positive integer and let \( K \) be a language. We say that \( K \) has a \textit{C-distribution} ([ER2]) if there exists an alphabet \( \Delta \) such that the set of letters occurring in every subword (of a word in \( K \)) of length \( C \) equals \( \Delta \). If \( K \) has a C-distribution for some \( C \), then we say that \( K \) has a \textit{constant distribution}.

Let us recall that for a language \( K \) its \textit{subword complexity}, denoted \( \pi_K \), is a function of positive integers such that \( \pi_K(n) \) is the number of different subwords of length \( n \) occurring in words of \( K \).

The following result was proved in [ER2].
Proposition 1. Let $K$ be a DOL language that has a constant distribution. Then there exists a positive integer $q$ such that $\pi_K(n) \leq qn$ for each positive integer $n$. 

To simplify the notation, in the rest of this paper we will consider an arbitrary but fixed alphabet $E$; all languages considered are over $E$.

Also, since problems considered become trivial otherwise, unless stated otherwise we consider only infinite DOL systems (and so only infinite DOL languages).

The following technical notion will be a useful tool in proving our result.

Definition. A language $K$ is simple if the following holds: there exist words $x_1, \ldots, x_\ell$, $\ell \geq 1$, such that $\text{alph}(x_i) = \text{alph}(x_j)$ for all $1 \leq i, j \leq \ell$ and $K \subseteq \{x_1, \ldots, x_\ell\}^*$. If $K, x_1, \ldots, x_\ell$ are as above then we also say that $K$ is $\{x_1, \ldots, x_\ell\}$-simple.

Note that each singleton language is simple.

RESULT

Theorem 1. If $K$ is a locally catenative DOL language, then there exists a positive integer $q$ such that $\pi_K(n) \leq qn$ for every positive integer $n$.

Proof.

The proof of this theorem goes through a sequence of lemmas as follows.
Lemma 1. If $K$ is simple then $K$ has a constant distribution.

Proof of Lemma 1.

Suppose that $K$ is $\{x_1, \ldots, x_\ell\}$-simple. Then it is easy to see that $K$ has a $C$-distribution where $C = 2\max\{|w_i| : 1 \leq i \leq \ell\}$. □

Lemma 2. If $K$ is a simple DOL language, then $\pi_K$ is bounded by a linear function.

Proof of Lemma 2.

Lemma 2 follows directly from Lemma 1 and Proposition 1. □

Lemma 3. Each relatively prime locally catenative DOL language is a finite union of simple DOL languages.

Proof of Lemma 3.

Let $K$ be a relatively prime locally catenative DOL language and let $G = (\Sigma, h, \omega)$ be a relatively prime locally catenative DOL system generating $K$, that is $L(G) = K$. Thus $G$ is $(i_1, \ldots, i_m)$-locally catenative with threshold $r_0$ where $\gcd(i_1, \ldots, i_m) = 1$. Let $E(G) = \omega_0, \omega_1, \ldots$

It is well known that the sequence $\{\alpha_x(\omega_i)\}_{i \geq 0}$ is ultimately periodic; let $n_0$ be a threshold and $p$ a period of this sequence. Let $q_0 = \max\{r_0, n_0\}$.

Since $\gcd(i_1, \ldots, i_m) = 1$ there exist nonnegative integers $k_1, \ldots, k_m$ such that $k_1i_1 + k_2i_2 + \ldots + k_mi_m = 1 \pmod{p}$. \ldots . \ldots (1)

Let $t = k_1i_1 + k_2i_2 + \ldots + k_mi_m$. From (1) it follows that $\omega_{n+t} = \omega_n + ps + 1$ for some $s \geq 0$; since $p$ is a period of the
the sequence \( \{\alpha \phi(n_i)\}_{i \geq 0} \), this implies that for each \( n \geq q_0 \),
\[
\alpha \phi(n_{n+t}) = \alpha \phi(n_{n+1})
\]
.................................(2).

On the other hand it is obvious that, for each \( n \geq q_0 \), \( n \) is a
subword of \( n_{n+t} \) and consequently we have
for each \( n \geq q_0 \), \( \alpha \phi(n) \subseteq \alpha \phi(n_{n+t}) \)
.................................(3).

From (2) and (3) it follows that for each \( n \geq q_0 \),
\[
\alpha \phi(n) \subseteq \alpha \phi(n_{n+1})
\]
.................................(4).

From (4) it follows that for some \( e \geq 1 \) the language
\( \{\omega_e, \omega_{e+1}, \ldots\} \) is a simple DOL language and consequently \( K \) is
a finite union of simple DOL languages.

Thus Lemma 3 holds.

Now we complete the proof of the theorem as follows.

Let \( K \) be a locally catenative DOL language.

If \( K \) is relatively prime, then the theorem follows from
Lemma 2 and Lemma 3.

Let us assume then that \( K \) is not relatively prime. Let \( H \)
be a locally catenative DOL system such that \( L(H) = K \). Hence \( H \) is
\( (i_1, \ldots, i_n) \)-locally catenative for some \( i_1, \ldots, i_n \) such that
\( \gcd(i_1, \ldots, i_n) = d > 1 \). Clearly by the \( d \)-speed up of \( H \) we
obtain \( d \) relatively prime locally catenative systems \( H_1, \ldots, H_d \)
such that \( K = L(H) = L(H_1) \cup \ldots \cup L(H_d) \).

Hence, by Lemma 2 and Lemma 3, the theorem holds also in
this case. □
To put the above result in a proper perspective let us recall the following result (see [R]).

**Proposition 2.** Let $K$ be a language. Either

1. $\pi_k(n) \geq n + 1$ for every positive integer $n$, or
2. there exists a positive integer $C$ such that $\pi_k(n) \leq C$ for every positive integer $n$.

Hence (except for a trivial case) the subword complexity of a locally catenative DOL language is "as low as possible."

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of NSF grant number MCS79-03838.

REFERENCES


