ON THE SUBWORD COMPLEXITY
OF SQUARE-FREE DOL LANGUAGES

by

A. Ehrenfeucht
and
G. Rozenberg

CU-CS-174-80 April, 1980

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

All correspondence to second author.

ON THE SUBWORD COMPLEXITY OF SQUARE-FREE DOL LANGUAGES

by

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

and

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

April, 1980

All correspondence to the second author.

ABSTRACT

The subword complexity of a language K is the function which to every positive integer n assigns the number of different subwords of length n occurring in words of K. A language K is square-free if no word in it contains a subword of the form xx where x is a nonempty word. The (best) upper and lower bounds on the subword complexity of infinite square-free DOL languages are established.

INTRODUCTION

The problems of repetitions of subwords in words (and in infinite words) were first studied by A. Thue (see [11] and [12]). Since then those problems were investigated (and rediscovered) by quite a number of authors with quite different motivations. In particular results of Thue were also used in various constructions in formal language theory (see, e.g., [3]). Recently one notices a revival of interest in Thue problems among formal language theorists (see, e.g., [1], [2], [7], [8] and [10]). In particular (see [1] and [10]) it was discovered that the theory of nonrepetitive sequences of Thue is very strongly related to the theory of DOL sequences. For example, Thue's original examples of square-free sequences were constructed using DOL systems and indeed, as pointed out in [1], most (if not all) examples of nonrepetitive sequences known in the literature are either DOL sequences or codings of DOL sequences. In this way a quite significant connection is established between the theory of nonrepetitive sequences and the theory of DOL systems. The theory of nonrepetitive sequences originates a new and very interesting research area within the theory of DOL systems while the theory of DOL systems provides a better insight into the theory of nonrepetitive sequences (see, e.g., [1] and [10]).

In this paper we investigate DOL systems which generate nonrepetitive words only. In particular we investigate the upper and the lower bounds on the subword complexity of languages generated by such systems and we establish that those languages are quite "poor" as far as number of subwords is concerned. (For a language K its subword complexity is a function assigning to each positive integer n the number of different subwords of length n occurring in words of K). In a sense this result is quite counter intuitive: one is inclined to think that to construct an infinite language consisting of nonrepetitive words one needs a lot of different subwords to avoid repetitions. (This aspect of the problem was pointed to us by J. Berstel who suggested to investigate the subword complexity of DOL systems generating square-free words only. Actually J. Berstel conjectured that the subword complexity of such languages is bounded by a linear function; we prove that the number of subwords of length n in such languages is of order $n \log_2 n$). We believe that this paper sheds a new light on the theory of square-free languages (sequences) and that it demonstrates how known results and techniques of the theory of DOL systems contribute to the theory of nonrepetitive languages (sequences).

We assume the reader to be familiar with basic aspects of DOL systems (see, e.g., [9]).

PRELIMINARIES

We will use standard notation and terminology concerning DOL systems (see, e.g., [9]). Thus 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. A letter a is *erasing* if, for some $m \geq 1$, $h^m(a) = \Lambda$ (where Λ is the empty word), otherwise a is *nonerasing*; *maxr* G denotes $\max\{|x| : x = h(a)$ for some $a \in \Sigma\}$. Since the problems considered become trivial otherwise, we consider only DOL systems which generate infinite languages.

It turns out that the notion of the rank of a letter in a DOL system (see [5]) will be quite useful in our investigation.

Definition. Let $G = (\Sigma, h, \omega)$ be a DOL system and let, for a letter $a \in \Sigma$, $G_a = (\Sigma, h, a)$. We say that a letter $a \in \Sigma$ is of *rank* 0 (*in* G) if $L(G_a)$ is finite. Let, for $i \geq 0$, Σ_i denote the set of all letters of rank i and let, for $j \geq 1$, $G_{(j)} = (\Sigma_{(j)}, h_{(j)}, \omega_{(j)})$ where

$\Sigma_{(j)} = \Sigma \setminus \bigcup_{i=0}^{j-1} \Sigma_i$, $\omega_{(j)} = g_j(\omega)$ and, for $b \in \Sigma_{(j)}$, $h_{(j)}(b) = g_{(j)}h(b)$ where $g_{(j)}$ is the homomorphism on Σ^* defined by $g_{(j)}(a) = a$ for $a \in \Sigma_{(j)}$, and $g_{(j)}(a) = \Lambda$ for $a \in \bigcup_{i=0}^{j-1} \Sigma_i$. If a letter $a \in \Sigma_{(j)}$ is of rank 0 in $G_{(j)}$ then we say that it is of *rank* j (*in* G). If $a \in \Sigma$ is of rank j for some $j \geq 0$ then we say that a has *rank in* G; otherwise we say that a is *without a rank*. □

For a word x, $|x|$ denotes its lengths while (if x is nonempty) *first*x denotes the first letter of x. For a finite set A, #A denotes its cardinality. For a language K and a positive integer n, $sub_n K$ denotes the set of subwords of length n of K while *sub*K denotes the set of all subwords of K. Given an alphabet Σ and $\Delta \subseteq \Sigma$, $pres_\Delta$ denotes the homomorphism on Σ^* defined by $pres_\Delta(a) = \Lambda$ if $a \in \Sigma \setminus \Delta$ and $pres_\Delta(a) = a$ if $a \in \Delta$.

We need the following notions concerning repetitions of subwords in a word.

Definition. A word is called *square-free* if it does not contain a subword of the form x^2 where x is a nonempty word. A word is called *strongly cube-free* if it does not contain a subword of the form $x^2\, first\, x$ where x is a nonempty word. A *language* is called *square-free* (resp. *strongly cube-free*) if it does not contain a square-free (resp. strongly cube-free) word. □

√

Clearly, every square-free word (language) is also strongly cube-free. Actually strongly cube-free words (languages) can be viewed also differently.

Definition. A word y is said to have an *overlap* if there exist words y_1, y_2, x_1, x_2, x_3 and x such that $y = y_1 x_1 x_2 x_3 y_2$, $x = x_1 x_2 = x_2 x_3$ where x_1, x_2, x_3 are nonempty words. Otherwise we say that y is *overlap-free*. A *language* is called *overlap-free* if each word in it is overlap-free. □

Theorem 1. A word is overlap-free if and only if it is strongly cube-free.

Proof.

(i). Let u be a word containing two overlapping occurrences of the same word. Hence $u = u_1x_1x_2x_3u_2$ where for some word x, $x_1x_2 = x_2x_3 = x$ where x_1, x_2, x_3 are all nonempty words; thus u has two different occurences of x "overlapping on" x_2. But then x_1x_1 *first*x_1 is a subword of u and so u is not strongly cube-free.

(ii). Let u be a word which can be written in the form $u = u_1 x x (\mathit{first}\, x) u_2$ where x is a nonempty word; hence u is not strongly cube-free. Then $\mathbf{u} = u_1 x (\mathit{first}\, x) y (\mathit{first}\, x) u_2$ where $x = (\mathit{first}\, x) y$. But then u can be written in the form $u = u_1z_1z_2z_3u_2$ where $z_1 = x$, $z_2 = \mathit{first}\, x$ and $z_3 = y\ \mathit{first}\, x$. Consequently u has two different occurrences of $z = z_1z_2 = z_2z_3$ "overlapping on" z_2. But then u is not overlap-free. □

RESULTS

In this section the subword complexity of square-free DOL languages is investigated. We begin by establishing an upper bound for this complexity.

Theorem 2. If K is a square-free DOL language then, for every positive integer n, $\#sub_n K \le C\, n \log_2 n$ for some positive integer constant C.

Proof. Let $G = (\Sigma, h, \omega)$ be a DOL system generating K.

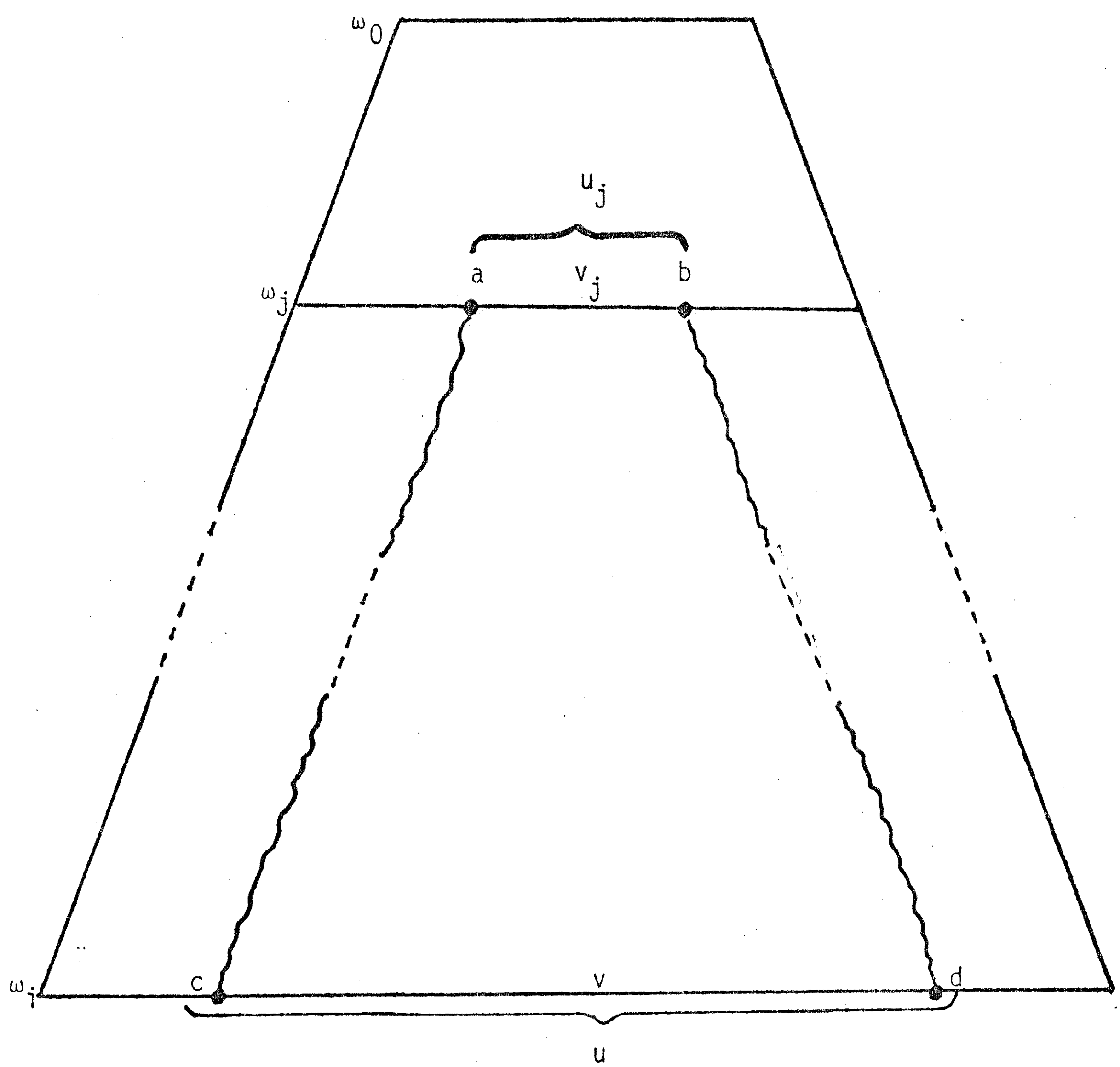
(i). If $a \in \Sigma$, then either a is of rank 0 or a does not have a rank.

This is established as follows. If a has a rank greater than 0 then G must contain a letter b of rank 1 such that, for some $m \ge 1$, $h^m(b) = u\,b\,v$ where $u, v \in \Sigma^*$, $uv \ne \Lambda$ and u and v consist of letters of rank 0 only. Since both $u, h^m(u), h^{2m}(u), \ldots$ and $v, h^m(v), h^{2m}(v), \ldots$ are infinite ultimately periodic sequences, L(G) cannot be square-free; a contradiction.

(ii). There exists a positive integer constant q such that if u is a subword of K consisting of letters of rank 0 only, then $|u| < q$.

This is proved by contradiction as follows. Let u be "an arbitrarily long" subword of K consisting of letters of rank 0 only. Since it is well known (see, e.g., [9]) that subwords consisting of erasing letters only are shorter than certain constant, u must contain "arbitrarily many" nonerasing letters. Let $E(G) = \omega_0, \omega_1, \omega_2, \ldots$ where for some $i \ge 1$, $\omega_i = x\,u\,y$. Notice that in words $\omega_0, \omega_1, \ldots, \omega_{i-1}$ we can distinguish (occurrences of) subwords $u_0, u_1, \ldots, u_{i-1}$ respectively which are the shortest subwords which are ancestors of u. Let j be the minimal integer such that $|u_j| \ge 2$. So let $u_j = a\,v_j\,b$ where $a, b \in \Sigma$, $v_j \in \Sigma^*$.

Clearly $|u_j| \leq \max\{|\omega|, maxr\, G\}$ and v_j, if nonempty, consists of letters of rank 0 only (because its contribution to ω_i is either empty or it consists of letters of rank 0 only). Let $u = c\,v\,d$ where $c, d \in \Sigma$ and $v \in \Sigma^*$. The situation can be best illustrated as follows:



Since the length of v_j is limited and u is arbitrarily long either on the path from a to c or on the path from b to d there must be a symbol, say e, repeating at least twice which contributes to v a subword which contains a nonerasing letter; since both cases are symmetric assume that e occurs on the path from a to c. Hence for some $m \geq 1$ $h^m(e) = z_1 e z_2$ where z_2 is nonempty and consists of type 0 letters only with at least one of them being nonerasing. Since, clearly, z_2, $h^m(z_2)$, $h^{2m}(z_2)$, ... is an infinite ultimately periodic sequence of nonempty words, L(G) must contain a word which is not square-free; a contradiction. Hence there exists a positive integer constant q such that each subword of L(G) consisting of letters of rank 0 only must be shorter than q.

(iii). Now let $\overline{G} = (\overline{\Sigma}, \overline{h}, \overline{\omega})$ be the DOL system defined as follows:

$\overline{\Sigma} = \{[u,a,v] : u,v \in \Sigma_0^*, |u| < q, |v| < q \text{ and } a \in \Sigma \setminus \Sigma_0\}$,

$\overline{\omega} = [u_1,a_1,\Lambda][u_2,a_2,\Lambda] \ldots [u_\ell,a_\ell,u_{\ell+1}]$ where

$u_1,u_2,u_3, \ldots,u_{\ell+1} \in \Sigma_0^*$, $a_1,\ldots,a_\ell \in \Sigma \setminus \Sigma_0$, $\ell \geq 1$ and

$\omega = u_1 a_1 u_2 a_2 \ldots u_\ell a_\ell u_{\ell+1}$,

for $[u,a,v] \in \overline{\Sigma}$, $\overline{h}([u,a,v]) = [z_0,b_1,\Lambda] \ldots [z_{k-1},b_k,z_k]$

where $k \geq 1$, $h(a) = x_0 b_1 x_1 b_2 \ldots b_k x_k$, $x_0, \ldots, x_k \in \Sigma_0^*$,

$b_1, \ldots, b_k \in \Sigma \setminus \Sigma_0$, $z_0 = h(u) x_0$, $z_1 = x_1$, $z_2 = x_2$, .., $z_{k-1} = x_{k-1}$

and $z_k = x_k h(v)$.

We can clearly assume that $\overline{G}$ is an everywhere growing DOL system (i.e., for every $a \in \Sigma$, $|h(a)| \geq 2$); if $\overline{G}$ is not such a system then we can speed it up (see, e.g., [8]) and then deal with a finite number of DOL systems $\overline{G}_1, \ldots, \overline{G}_m$ each of which is everywhere growing. From the construction of $\overline{G}$ it directly follows that $L(G) = g(L(\overline{G}))$ where g is the homomorphism on $\overline{\Sigma}^*$ defined by $g([u, a, v]) = u\,a\,v$. It is proved in [6] that if H is an everywhere growing DOL system and f is a nonerasing homomorphism then, for every positive integer n, $\#sub_n\, f(L(H)) \leq D\, n \log_2 n$ for some positive integer D.

Thus the theorem holds. □

We demonstrate now that the above established upper bound $(n \log_2 n)$ is the best possible.

Theorem 3. There exist a square-free DOL language K and a positive constant D such that for every $n \geq 1$, $\#sub_n K \geq D\, n \log_2 n$.

Proof. Consider the DOL system $G = (\Sigma, h, \omega)$ with $\Sigma = \{0,1,2\}$, $h(0) = 012$, $h(1) = 02$, $h(2) = 1$ and $\omega = 0$ from [8]. It is shown in [8] (see also [1]) that L(G) is square-free. Let $G_{(3)} = (\Sigma, h_{(3)}, 0)$ where for $a \in \Sigma$, $h_{(3)}(a) = h^3(a)$; thus $G_{(3)}$ results from G by starting with the axiom 0 and then taking only each third word of G. Clearly also $L(G_{(3)})$ is square-free. Notice that if f_G and $f_{G_{(3)}}$ denote the growth functions of G and $G_{(3)}$ respectively then

for $n \geq 0$, $f_G(n) \leq 3^n$ and $f_{G_{(3)}}(n) > 4^n$ ....(1).

Now let $H = (\Theta, g, 0\overline{0}\overline{\overline{0}})$ be the DOL system where $\Theta = \Sigma \cup \overline{\Sigma} \cup \overline{\overline{\Sigma}}$ with $\overline{\Sigma} = \{\overline{a} : a \in \Sigma\}$ and $\overline{\overline{\Sigma}} = \{\overline{\overline{a}} : a \in \Sigma\}$, $g(a) = h_{(3)}(a)$, $g(\overline{a}) = \overline{h(a)}$ and $g(\overline{\overline{a}}) = \overline{\overline{h_{(3)}(a)}}$ for $a \in \Sigma$ (where for a word $\alpha \in \Sigma^+$, $\overline{\alpha}$ results from α by replacing every letter a in it by $\overline{a}$ and $\overline{\overline{\alpha}}$ results from α by replacing every letter a in it by $\overline{\overline{a}}$).

Clearly also L(H) is square-free. Let $n \geq 1$ and let us estimate a lower bound for $\#sub_{3n} L(H)$. To this aim consider the word

$z = g^m(0\bar{0}\bar{\bar{0}})$ where $m = \lceil \log_4 2n \rceil$. Then $z = z_1 z_2 z_3$ where $z_1 \in \Sigma^+$, $z_2 \in \bar{\Sigma}^+$ and $z_3 \in \bar{\bar{\Sigma}}^+$. Notice that it follows from (1) that $|z_3| \geq 2n$. Let y be the prefix of z_3 of length 2n. Since L(H) is square-free (and so by Theorem 1 also overlap-free) all subwords of y of length n are different. Let u be one fixed subword out of these n subwords. Note that E(G) has the strong prefix property (that is $h^{n+1}(\omega) = h^n(\omega)\alpha_n$ for each $n \geq 0$ where $\alpha_n \in \Sigma^+$) hence we can talk about the "fixed occurrence of u" in z_3 and in all suffixes of all consecutive words of L(H) where we consider the longest suffixes which are over the alphabet $\bar{\bar{\Sigma}}$. Now let us estimate the lower bound for the number of all those subwords of L(H) that end on this fixed occurrence of u and are of length 3n.

Note that if t and t' are such two different subwords where $|pres_{\bar{\Sigma}}\, t| \leq n$ and $|pres_{\bar{\Sigma}}\, t'| \leq n$ then $t \neq t'$ (because f_G is a monotonically growing function). Hence, let us estimate a bound on a positive integer p having the property that if $x = g^{m+p}(0\bar{0}\bar{\bar{0}})$ then $|pres_{\bar{\Sigma}}\, x| \leq n$. First of all, as long as $3^{m+p} \leq n$, then (by (1)) p has the desired property.

Thus $(m+p) \log_4 3 \leq \log_4 n$

and consequently $p \leq C \log_4 n - 0.5$, where $C = \dfrac{1 - \log_4 3}{\log_4 3}$.

Since we have n possible choices for u we get that

$\#sub_{3n}\ L(H) \geq n\ (C \log_4 n - 0.5)$.

Consequently there exists a positive constant C_1 such that for all $n \geq 4$

$\#sub_{3n}\ L(H) \geq C_1\ n \log_4 n$

(any C_1 such that $C_1 \leq C - 0.5$ will do).

Then it is rather easy to see that there exists a positive constant D such that $\#sub_n\ L(H) \geq D\, n \log_2 n$ for every $n \geq 1$.

Hence the theorem holds. □

We turn now to the lower bound on the subword complexity of square-free DOL languages.

Theorem 4. If K is an infinite square-free language then $\#sub_n\ K \geq n$ for every positive integer n.

Proof.

Let n be a positive integer. If $n = 1$ then clearly $\#sub_n\ K \geq n$. So let $n \geq 2$ and let z in K be such that $|z| \geq 2n-1$. Let z_1, z_2, ..., z_{n-1} be words resulting from z be erasing from it the first, the two first, ..., and the (n-1) first letters respectively. Now let $y, y_1, \ldots, y_{n-1}$ be prefixes of length n of words $z, z_1, \ldots, z_{n-1}$ respectively. Note that all those words $y, y_1, \ldots, y_{n-1}$ appear as subwords of z in such a way that any two of them overlap in z. Since K is square-free, Theorem 1 implies that K is overlap-free and consequently $y, y_1, \ldots, y_{n-1}$ are all different subwords of z. Thus $\#sub_n\ K \geq n$. □

Finally we demonstrate that the linear bound on the subword complexity of square-free DOL languages is the best possible.

Theorem 5. There exist a square-free DOL language K and a positive integer constant C such that for every positive integer n, $\#sub_n K \leq Cn$.

Proof. It is well-known (see, e.g., [1]) that there exists a square-free DOL language defined by a uniformly growing DOL system. (A DOL system $G = (\Sigma, h, \omega)$ is called uniformly growing if there exists a positive integer constant t such that, for every $a \in \Sigma$, $|h(a)| = t$.) However, if G is a uniformly growing DOL system then (see [4]) there exists a positive integer constant C such that, for all $n \geq 0$, $\#sub_n L(G) \leq Cn$. □

We conclude this paper with the following two remarks.

(1). In this paper we have established lower and upper bounds on the subword complexity of square-free DOL languages. Thue's original interest (as well as the interest of the most of his followers) was in square-free infinite words. For this reason [1] and [9] consider DOL systems (Σ, h, ω) with the property that ω is a prefix of $h(\omega)$; each DOL system of this kind defines a unique infinite word.It is easy to see that al results we have presented in this paper are also valid for DOL systems of this particular kind.

(2). Analogously to the notion of a square-free word (language), for every $k \geq 2$ we can consider the notion of a k-repetitions-free word (language); Thue considered 3-repetitions-free words which he called <u>cube-free</u>.It is easy to see that our lower and upper bounds for the subword complexity remain valid also in the general case of k-repetitions-free DOL languages.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of NSF grant number MCS 79-03838. The authors are indebted to J. Berstel for bringing to their attention the topic of the subword complexity of square-free DOL languages.

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] Berstel, J., Mots sans carré et morphismes iterés, *Discrete Mathematics,* v.29, n.3, 235-244, 1979.

[3] Brzozowski, J. A., Culik II, K. and Gabrielian A., Classification of noncounting events, *Journal of Computer and System Science,* v.5, 41-53, 1971.

[4] Ehrenfeucht, A., Lee, K. P. and Rozenberg, G., Subword complexities of various classes of deterministic developmental languages without interactions, *Theoretical Computer Science,* v.1, 59-76, 1975.

[5] Ehrenfeucht, A. and Rozenberg, G., On the structure of polynomially bounded DOL systems, *Fundamenta Informaticae,* v.2, 187-197, 1979.

[6] Ehrenfeucht, A. and Rozenberg, G., On subword complexities of homomorphic images of languages, *Dept. of Comp. Science, University of Colorado at Boulder, Technical Report* CU-CS-173-80, 1980.

[7] Harrison, M., *Introduction to formal language theory,* Addison-Wesley, 1978.

[8] Istrail, S., On irreducible languages and nonrational numbers, *Bull. Soc. Math. Roumanie,* v.21, 301-308, 1977.

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

[10] Salomaa, A., Morphisms on free monoids and language theory, in Book, R.(ed.), *Formal language theory: perspectives and open problems,* Academic Press, London - New York, to appear.

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

[12] Thue, A., Über die gegenseitige lage gleicher teile gewisser zeichenreichen, *Norsk. Vid. Selsk. Skr. I Math. - Nat. Kl.,* nr. 1, 1-67, 1912.