ON THE SEPARATING POWER OF EOL SYSTEMS

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-186-80 September, 1980

* 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.

ABSTRACT

A word is called a *pure square* if it is of the form yy where y is a nonempty word; it is called a *square* if it contains a pure square — otherwise it is called *square-free*. A language K *separates* languages K_1 and K_2 if $K_1 \subseteq K$ and $K_1 \cap K_2 = \emptyset$. It is demonstrated that no EOL language (and hence no context-free language) can separate the set of all pure squares over an alphabet Δ from the set of all square-free words over Δ, where Δ has at least three letters. Thus the set of all square words over Δ is not an EOL language (and so it is not a context-free language). This settles an open problem posed by J. Berstel and L. Boasson.

INTRODUCTION

Let L be a class of languages. A way to investigate the structure of languages in L is to aim at results of the form: "If $K \in L$ and K contains some words, then K must contain some other words". A classical result in this direction is the pumping-lemma for context-free languages (see, e.g., [H]). In the pumping lemma "some words" are distinguished by certain minimal length. In general one would like to have a result of the form: "If $K \in L$ and K contains words satisfying property P then K must contain some other words (e.g., not satisfying P)" where P is a combinatorial property of words. Such a result can be formulated as follows. We say that K *separates* languages K_1 and K_2 if $K_1 \subseteq K$ and $K \cap K_2 = \emptyset$. Then we set K_1 to be equal to the set of words satisfying the property P (or to its subset) and we set K_2 to be equal to the set of words satisfying a property R (or to its subset) and we get the following formulation of the desired result: "If $K \in L$ then K does not separate K_1 from K_2".

A very basic combinatorial property of a word is a structure of repetitions of its subwords. Following [T] we say that a word is *square-free* if it does not contain a subword of the form yy where y is a nonempty word; otherwise we say that the word is a *square*. A word is a *pure square* if it is of the form yy where y is a nonempty word. Then a language is called square-free (square, pure square) if it consists of square-free (square, pure square) words only. Square-free languages (and sequences) have a large number of interesting mathematical applications and interpretations (see, e.g., [S2]). Also recently they form an active research topic within formal language theory (see, e.g., [B], [ER], [S1] and [S2]).

Because of the pumping lemma it is clear that given an alphabet Δ with at least 3 letters (there exist only six square-free words over an alphabet of two letters!) no context-free language can be equal to (the infinite subset of) the set of all square-free words over Δ. However, pumping is a mechanism generating repetitions of words and so it is quite natural to ask whether a context-free grammar can generate the set of all squares over Δ. (This question was posed by J. Berstel and L. Boasson from Paris).

In this paper we answer this question in negative. As a matter of fact, we prove a quite stronger result: no EOL language (see, e.g., [RS]) can separate the set of all pure squares over Δ from the set of all square free words over Δ. This settles the original problem because the class of EOL languages (strictly) contains the class of context-free languages. We believe that our result contributes to the understanding of the combinatorial structure of EOL (and hence also context-free) languages.

We assume the reader to be familiar with basic theory of EOL languages, e.g., in the scope of [RS].

PRELIMINARIES

We will use mostly standard formal language-theoretic notation and terminology. Perhaps only the following points require an additional comment.

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

For a language K, #K denotes its cardinality and *alph* $K = \bigcup_{x \in K}$ *alph*(x);

$K_1 \setminus K_2$ denotes the set theoretic difference of languages K_1 and K_2.

For a finite set K, #K denotes its cardinality.

A homomorphism $h: \Sigma^* \to \Delta^*$ is termed *propagating* if $h(a) \neq \Lambda$ for all $a \in \Sigma$.

In this paper we consider finite alphabets only.

We will follow [RS] in our notation and terminology concerning L systems. In particular we denote an EOL *system* by $G = (\Sigma, h, S, \Delta)$ where Σ is the alphabet of G, h its finite substitution, S its axiom and Δ the terminal alphabet of G. We will also use *al*(G) to denote Σ and *maxr*(G) to denote $\max\{|\alpha| : \alpha \in h(a)$ for some $\alpha \in \Sigma\}$.

The analysis of derivations trees in an EOL system plays an important role in this paper. We will use somewhat informally the notion of a contribution of a node in a derivation tree of T to the result of T. We also need the following notions concerning derivation trees.

Definition. Let G be an EOL system and let T be a derivation tree of a word w in G, where $|w| \geq 2$.

(1) The *main path* of T, denoted by *main*(T), is the path defined by:

(i). the first node of *main*(T) is the root of T,

(ii). if v is the i'th node of *main*(T), $i \geq 1$, and it is not the leaf then the (i+1)'st node of *main*(T) is the leftmost among all those descendants of v that have the contributions to w not shorter than the length of the contribution to w of any of the successors of v,

(iii). the last node of main(T) is a leaf of T.

(2). The *special node of* T, denoted by *spec*(T), is the first node (counted from the root) with the property that the length of its contribution to w is not longer than $\frac{|w|}{2}$.

(3). The *type of* T, denoted by *type*(T), is the vector (A, k, ℓ, d) such that:

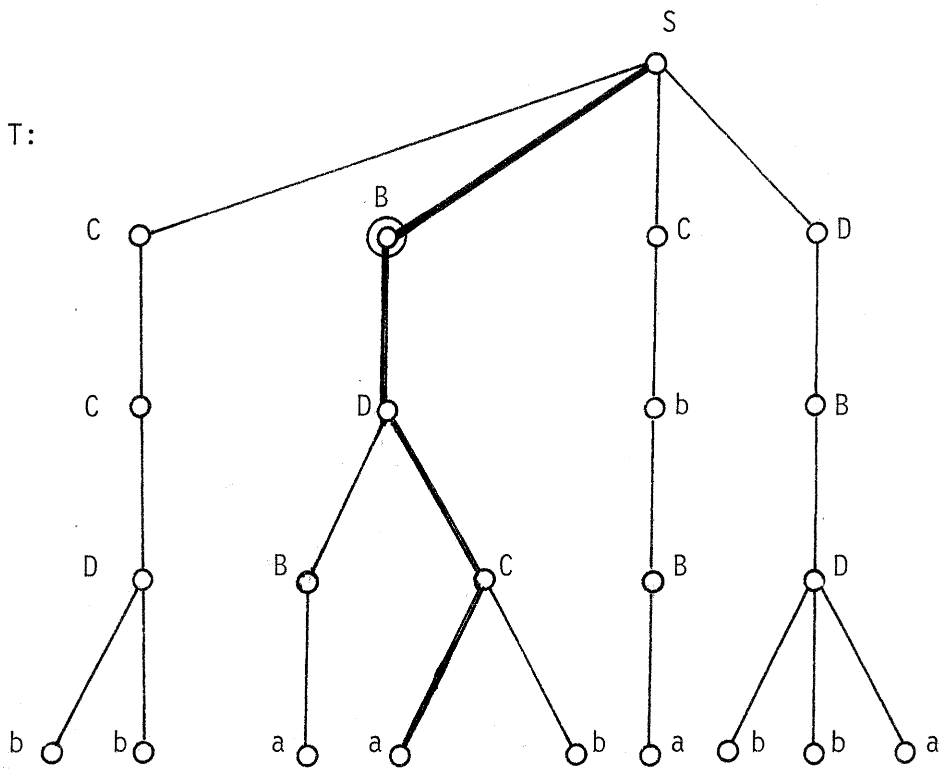
A is the label of spec(T),

the contribution of *spec*(T) to w starts on the k'th letter of w and ends on the ℓ'th letter of w,

the distance of *spec*(T) to the last node of *main*(T) equals d. □

Example. In the picture of the following derivation tree T in an EOL system the main path is in bold face and the special node is double circled:

T:



The type of T is (B, 3, 5, 3,). □

Lemma 1. Let G be an EOL system and let T be a derivation tree of a word w in G. The length of the contribution of *spec*(T) to w is longer than $\frac{|w|}{2\ maxr(G)}$.

Proof.

Assume to the contrary that this contribution is not longer than $\frac{|w|}{2\ maxr(G)}$. Then (because clearly *spec*(T) is different from the root of T) *spec*(T) has an ancestor in T such that the length of his contribution to w is not longer than $\frac{|w|}{2}$. This, however, contradicts the definition of the special node of T; thus the lemma holds. □

The following class of EOL systems will be considered in this paper.

Definition. Let G be an EOL system, $w \in L(G)$ and let D be a derivation of w in G. We say that D is a *fast derivation* if its length is not bigger than $|w|$. We say that G is a *fast* EOL *system* if for every word w in L(G) there exists a fast derivation of w in G. □

Lemma 2. For every EOL language K there exists a fast EOL system G such that L(G) = K.

Proof.

It is well-known (see [vL]) that for every EOL language K there exists an EOL system H generating K such that for every word w in L(H) there exists a derivation of w in H such that the length of this derivation is bounded by $C|w|$ where C is a constant dependent on H only. Applying the C speed-up to H (see [RS]) one obtains the EOL system $G = speed_C\ H$ which is fast. □

The following notions concerning repetitions of subwords in a word will be considered in the sequel.

Definition. (1). A word is called a *pure square* if it is of the form yy where y is a nonempty word. (2). A word is called a *square* if it contains a subword that is a pure square; otherwise we say that the word is *square-free.* □

Given an alphabet Δ and a positive integer n we let

$PSQ_n(\Delta)$ to denote the set of all words of length n over Δ which are pure squares,

$PSQ(\Delta)$ to denote the set of all pure square words over Δ,

$SQ(\Delta)$ to denote the set of all square words over Δ,

$SQF_n(\Delta)$ to denote the set of all square-free words over Δ of length n, and

$SQF(\Delta)$ to denote the set of all square-free words over Δ.

The following basic result is from [T].

Lemma 3. If Δ is an alphabet such that $\#\Delta \geq 3$ then there exists an infinite square-free word over Δ. □

Definition. Let h be a homomorphism, $h : \Sigma^* \to \Delta^*$. We say that h is *square-free* if, for every $w \in SQF(\Sigma)$, $h(w) \in SQF(\Delta)$. □

The following result from [BEN] concerning propagating square-free homomorphisms will be useful in our considerations.

Lemma 4. For every positive integers $k \geq 2$, $\ell \geq 3$ there exist alphabets Σ, Δ and a propagating square-free homomorphism $h : \Sigma^* \to \Delta^*$ where $\#\Sigma = k$ and $\#\Delta = \ell$. □

RESULTS

The following notion is the basic notion of this paper.

Definition. Let K, K_1, K_2 be languages. We say that K *separates* K_1 *from* K_2 if $K_1 \subseteq K$ and $K \cap K_2 = \emptyset$; this is denoted by writing $K_1 - K - K_2$. □

We will demonstrate that no EOL language can separate PSQ(Δ) from SQF(Δ) when #Δ > 2. We start by showing that if G is a fast EOL system such that L(G) separates $PSQ_n(\Delta)$ from $SQF_n(\Delta)$, where n is even and #Δ ≥ 7, then the cardinality of the alphabet of G grows (fast!) with the growth of n.

Lemma 5. Let Δ be a finite alphabet with #Δ ≥ 7 and let n be a positive even integer. Let G be a fast EOL system such that

$PSQ_n(\Delta) - L(G) - SQF_n(\Delta)$. Then $\#al(G) > \dfrac{2^{\frac{n}{2maxr(G)}}}{n^3}$.

Proof.

Let G = (Σ, h, S, Δ) be a fast EOL system such that $PSQ_n(\Delta) - L(G) - SQF_n(\Delta)$. Let #Σ = m and *maxr*(G) = t. Let Δ_1 be a fixed subset of Δ consisting of 7 symbols, say $\Delta_1 = \{a_0,a_1,b_0,b_1,c_0,c_1,\$\}$ and let α be a fixed square-free word over the alphabet Θ = {a,b,c} where $|\alpha| = \frac{n}{2} - 1$ (the existence of such an α is guaranteed by Lemma 3). Let $\Delta_2 = \Delta_1 \backslash \{\$\}$ and let g be the homomorphism from Δ_2^* onto Θ^* defined by: $g(a_i) = a$, $g(b_i) = b$ and $g(c_i) = c$ for $i \in \{1,2\}$.

Let $Z(\alpha,g) = \{\beta\$\beta\$: \beta \in \Delta_2^* \text{ and } g(\beta) = \alpha\}$.

Obviously

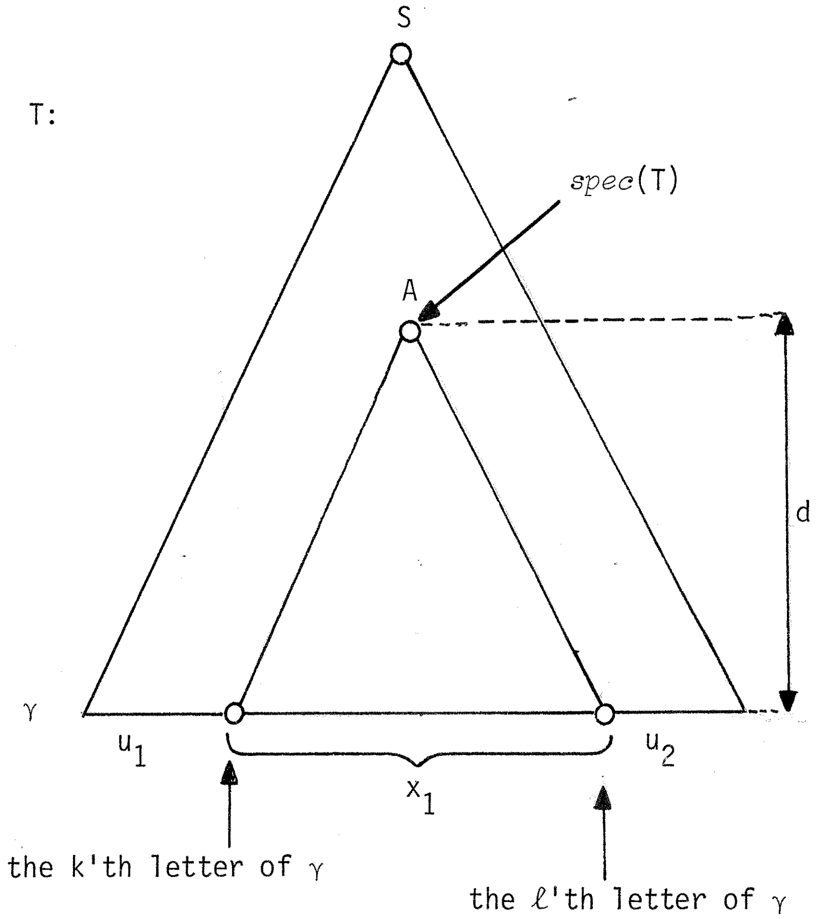
$Z(\alpha,g) \subseteq PSQ_n(\Delta)$ and $\#Z(\alpha,g) = 2^{\frac{n-2}{2}}$ ....(1)

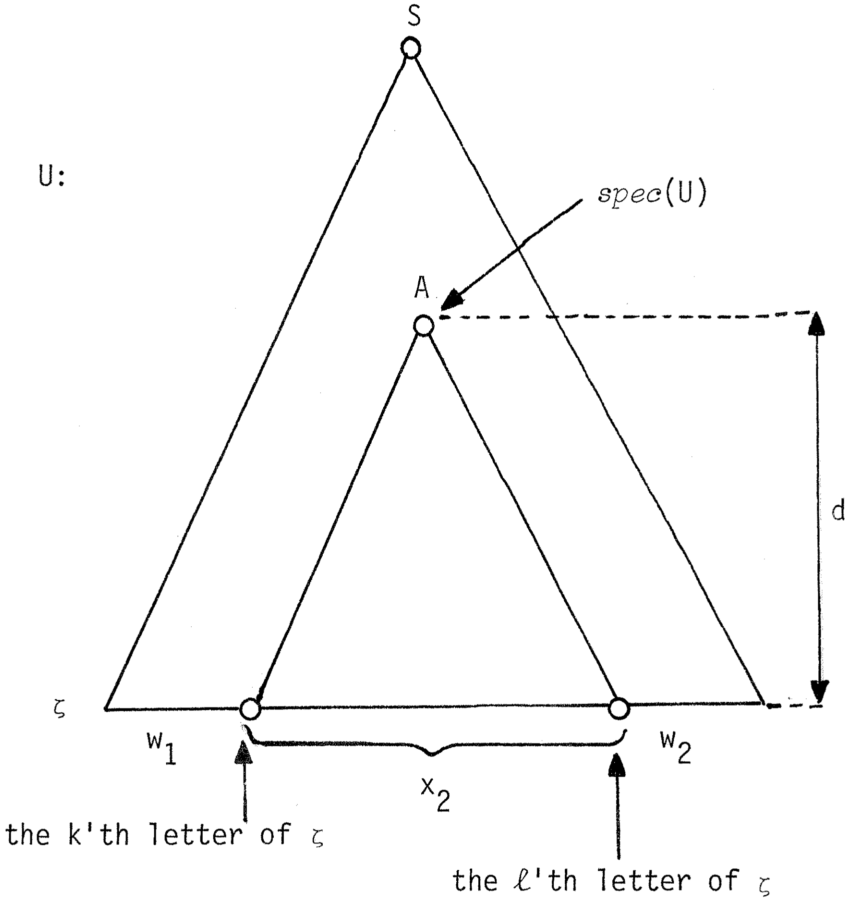
We define a *description of* $Z(\alpha,g)$ *in* G to be a set of ordered pairs (γ,T), where $\gamma \in Z(\alpha,g)$ and T is a derivation tree corresponding to a fast derivation of γ in G, such that for each γ in $Z(\alpha,g)$ only one element of the form (γ,T) is in the set. Let D be an arbitrary but fixed description of $Z(\alpha,g)$ in G.

Claim 1. Let (γ,T) and (ζ,U) be elements of D such that $\gamma \neq \zeta$ and *type*(T) = *type*(U). Then the subword contributed by *spec*(T) in T equals the subword contributed by *spec*(U) in U.

Proof of Claim 1.

The situation is best illustrated as follows:


T:
S
spec(T)
A
d
γ
u1
u2
x1
the k'th letter of γ
the ℓ'th letter of γ


S
U:
spec(U)
A
d
ζ
w_1
w_2
x_2
the k'th letter of ζ
the ℓ'th letter of ζ

where *type*(T) = *type*(U) = (A,k,ℓ,d).

Consequently $u_1\ x_2\ u_2 \in L(G)$.

Assume now, to the contrary, that the subword contributed by *spec*(T) in T is not equal to the subword contributed by *spec*(U) in U, hence $x_1 \neq x_2$. Then we observe the following.

(i). $u_1\ x_2\ u_2 \notin PSQ_n(\Delta)$.

This follows from the definition of the special node and the simple observation that if in a word from $PSQ_n(\Delta)$ one replaces a subword no longer than $\frac{n}{2}$ by a different subword of the same length than the resulting word is no longer in $PSQ_n(\Delta)$.

(ii). $u_1\ x_2\ u_2 \in SQF_n(\Delta)$.

This is proved as follows.

Assume that $u_1\ x_2\ u_2$ contains a square yy where y is a nonempty word. If $\$ \in$ *alph*(y) then $u_1\ x_2\ u_2 = yy$ which contradicts (i) above. Hence the definition of $Z(\alpha,g)$ implies that $u_1\ x_2\ u_2 = \beta\$\beta\$$ for some $\beta \in g^{-1}(\alpha)$ where yy is a subword of β. Consequently α is not square-free; a contradiction.

Thus, indeed, $u_1\ x_2\ u_2 \in SQF_n(\Delta)$ and (ii) is proved.

However (ii) contradicts the fact that $PSQ_n(\Delta) - L(G) - SQF_n(\Delta)$ and consequently it must be that $x_1 = x_2$. Hence Claim 1 holds.

We say that elements (γ_1,T_1), (γ_2,T_2), of D are *similar* if *type*(T_1) = *type*(T_2).

Claim 2. If W is a subset of $Z(\alpha,g)$ such that all words in W are similar, then $\#W \leq 2^{\frac{n}{2}(1-\frac{1}{t})}$.

Proof of Claim 2.

Assume that the type "shared by" all words in W is (A,k,ℓ,d). Hence if $k \le j \le \ell$ and $x, y \in W$ then the j'th occurrence in x is identical to the j'th occurrence in y. In other words, x and y can differ only by 0, 1-indices attached to occurrences of a, b, c outside of occurrences k through ℓ. Thus Lemma 1 implies that

$$\#W \le 2^{\frac{n-2}{2} - (\frac{n}{2t} - 1)} = 2^{\frac{n}{2}(1-\frac{1}{t})}.$$

Consequently Claim 2 holds.

Claim 3. Let $T_D = \{T : (\gamma,T) \in D \text{ for some } \gamma \in Z(\alpha,g)\}$. Then $\#\{\mathit{type}(T) : T \in T_D\} \le \frac{n^3}{2}\ \#\mathit{al}(G)$.

Proof of Claim 3.

Let $(A,k,\ell,d) \in \{\mathit{type}(T) : T \in T_D\}$. Since, for every $\gamma \in Z(\alpha,g)$, $|\gamma| = n$ (and so $d \le n$) and the number of possible pairs (k,ℓ) that can be chosen is bounded by $\binom{n}{2} \le \frac{n^2}{2}$, we have indeed that $\#\{\mathit{type}(T) : T \in T_D\} \le \frac{n^3}{2}\#\mathit{al}(G)$. □

Now we complete the proof of Lemma 5 as follows.

Clearly $\#Z(\alpha,g)$ is not bigger than the product of $\#\{\mathit{type}(T) : T \in T_D\}$ by the maximal number of words from $Z(\alpha,g)$ that can be similar. Thus Claim 2 and Claim 3 imply that

$$\#Z(\alpha,g) \le m\ \frac{n^3}{2}\ 2^{\frac{n}{2}(1-\frac{1}{t})}$$

and consequently

$$m \ge \frac{2^{\frac{n}{2t}}}{n^3}.$$

Thus the lemma holds. □

Theorem 1. Let $\#\Delta > 2$. Then no EOL language separates PSQ(Δ) from SQF(Δ).

Proof.

(i). The theorem holds when $\#\Delta \geq 7$.

This follows directly from Lemma 2 and Lemma 5.

(ii). The theorem holds when $2 < \#\Delta < 7$.

This is proved by contradiction as follows.

Assume that $2 < \#\Delta < 7$ and that K is an EOL language such that PSQ(Δ) - K - SQF(Δ). Let Θ be an alphabet such that $\#\Theta = 7$ and let f be a propogating square-free homomorphism from Θ^* into Δ^*; Lemma 4 guarantees the existence of such a homomorphism. Clearly

PSQ(Θ) $\subseteq f^{-1}$(PSQ(Δ)) and SQF(Θ)) $\subseteq f^{-1}$(SQF(Δ)).

Since it is easily seen that the inverse homomorphic image of an EOL language is an EOL language whenever the homomorphism involved is propagating, we get that

PSQ(Θ) - f^{-1}(K) - SQF(Δ),

where f^{-1}(K) is an EOL language.

This, however, contradicts (i), and consequently (ii) holds.

Thus the theorem holds. □

Corollary 1. Let Δ be an alphabet such that $\#\Delta > 2$. Then no EOL language can separate SQ(Δ) from SQF(Δ).

Proof.

Directly from Theorem 1. □

Corollary 2. Let Δ be an alphabet such that $\#\Delta > 2$. Then no context-free language can separate SQ(Δ) from SQF(Δ).

Proof.

Directly from Corollary 1 and from the fact that every context-free language is an EOL language (see, e.g., [RS]). □

We conclude this paper by the following remark. Originally the problem of separating SQ(Δ) from SQF(Δ) was posed for context-free languages. If one considers this original problem then the proof of the theorem goes in the same way except that now context-free grammars in Chomsky Normal Form play the same role as fast EOL systems played in our proof. In this case the formulation of Lemma 5 (which may be of interest on its own) becomes: "Let Δ be a finite alphabet with $\#\Delta \geq 7$ and let n be a positive even integer. Let G be a context-free grammar in Chomsky Normal Form such that $PSQ_n(\Delta)$ - L(G) - $SQF_n(\Delta)$. Then $\#\mathit{al}(G) > \frac{2^{\frac{n}{4}}}{n^2}$."

REFERENCES

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

[BEN] Bean, D. R., Ehrenfeucht, A. and McNulty, G. F., Avoidable patterns in strings of symbols, *Pacific Journal of Mathematics*, v. 85, n. 2, 261-294, 1979.

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

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

[vL] van Leeuwen, J., The tape complexity of context independent developmental languages, *Journal of Computer and System Sciences*, 11, 203-211, 1975.

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

[S1] 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.

[S2] Salomaa, A., *Jewels of formal language theory*, Computer Press, Potomac, Md., to appear.

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

ACKNOWLEDGMENTS

The authors gratefully acknowledge support under National Science Foundation grant number MCS 79-04038.