RAIRO. INFORMATIQUE THÉORIQUE

A. EHRENFEUCHT
G. ROZENBERG

On the separating power of EOL systems

RAIRO. Informatique théorique, tome 17, nº 1 (1983), p. 13-22
<http://www.numdam.org/item?id=ITA_1983__17_1_13_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 17, n° 1, 1983, p. 13 à 22)

ON THE SEPARATING POWER OF EOL SYSTEMS (*)

by A. EHRENFEUCHT [1] and G. ROZENBERG [2]

Communicated by J. BERSTEL

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 $K \cap K_2 = \emptyset$. 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 Autebert, Beauquier, Boasson and Nivat.*

Résumé. — *Un mot est appelé un* carré pur *s'il est de la forme yy avec y non vide ; il est appelé un* carré *s'il contient un carré pur — sinon il est appelé* sans carré. *Un langage K sépare les langages K_1 et K_2 si $K_1 \subseteq K$ et $K \cap K_2 = \emptyset$. On démontre qu'aucun langage EOL (*a fortiori *aucun langage algébrique) ne peut séparer l'ensemble de tous les carrés purs de l'ensemble de tous les mots sans carrés sur un alphabet Δ ayant au moins trois lettres. Par conséquent, l'ensemble de tous les carrés sur Δ n'est pas EOL, donc il n'est pas algébrique. Ceci résout un problème ouvert posé par Audebert, Beauquier, Boasson et Nivat.*

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. [5]). 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

(*) Received in October 1980, revised in February 1981.
[1] Department of Computer Science, University of Colorado, Boulder.
[2] Institute of Applied Mathematics and Computer Science, University of Leiden, 2300 RA Leiden, The Netherlands.

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 [10] 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. [9]). Also recently they form an active research topic within formal language theory (*see*, e. g. [2, 4, 8, 9].

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 in [1]).

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. [7]) 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 contains (strictly) 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 [7].

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 $alph K = \bigcup_{x \in K} alph(x)$; $K_1 \backslash K_2$ denotes the set theoretic difference of languages K_1 and K_2.

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

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

In this paper we consider finite alphabets only.

We will follow [7] 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

$$\max \{ |\alpha| : \alpha \in h(a) \quad \text{for some} \quad \alpha \in \Sigma \}.$$

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) of the main path with the property that the length of its contribution to w is not longer than $\frac{|w|}{2}$.

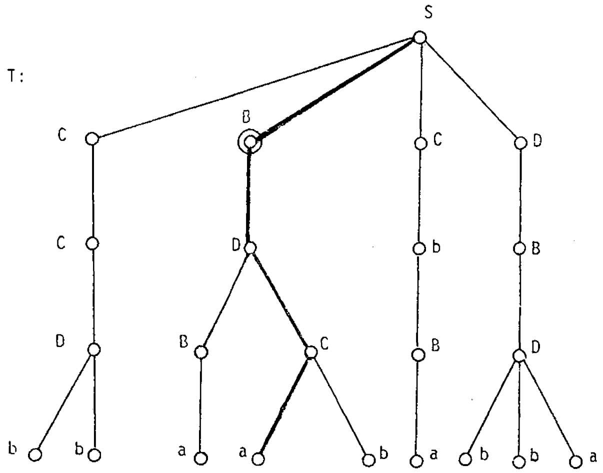
(3) The *type of* T, denoted by $type(T)$, is the vector (A, k, l, d) such that:

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:



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 $\frac{|w|}{2maxr(G)}$.

Proof: Assume to the contrary that this contribution is not longer than $\frac{|w|}{2maxr(G)}$. 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* [6]) 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* [7]) 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 [10].

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 [3] concerning propagating square-free homomorphisms will be useful in our considerations.

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

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 $PSQ(\Delta)$ from $SQF(\Delta)$ when $\# \Delta > 2$. 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 $\# \Delta \geq 7$, then the cardinality of the alphabet of G grows (fast!) with the growth of n.

LEMMA 5: Let Δ be a finite alphabet with $\# \Delta \geq 7$ and let n be a positive even integer. Let G be a fast EOL system such that

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

Proof: Let $G=(\Sigma, h, S, \Delta)$ be a fast EOL system such that

$$PSQ_n(\Delta) - L(G) - SQF_n(\Delta).$$

Let $\# \Sigma = m$ 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 $\Theta = \{ a, b, c \}$ 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 $i \in \{ 0, 1 \}$.

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

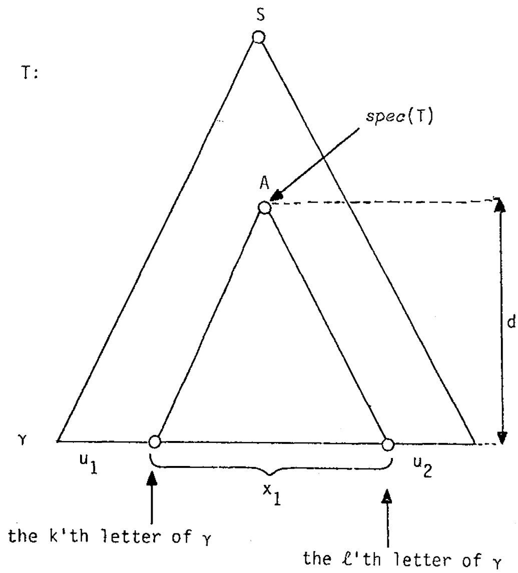
Obviously

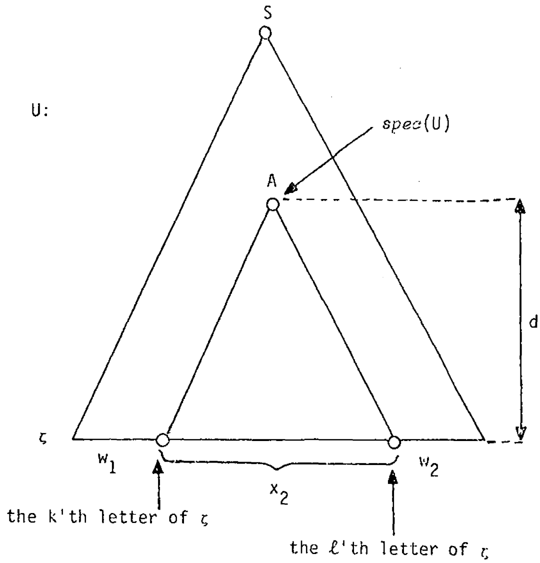
$$Z(\alpha, g) \subseteq PSQ_n(\Delta) \quad \text{and} \quad \# Z(\alpha, g) = 2^{\frac{n-2}{2}} \quad \ldots (1)$$

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:





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 $\# W \leq 2^{\frac{n}{2}\left(1-\frac{1}{t}\right)}$.

Proof of Claim 2: Assume that the type "shared by" all words in W is (A, k, l, d). Hence if $k \leq j \leq l$ 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

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

Consequently Claim 2 holds. □

CLAIM 3: Let $T_D = \{ T : (\gamma, T) \in D \text{ for some } \gamma \in Z(\alpha, g) \}$. Then

$$\# \{ type(T) : T \in T_D \} \leq \frac{n^3}{2} \# al(G).$$

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

$$\# \{ type(T) : T \in T_D \} \leq \frac{n^3}{2} \# al(G) = \frac{mn^3}{2}. \quad \square$$

Now we complete the proof of Lemma 5 as follows.

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

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

and consequently (because $\# Z(\alpha, g) = 2^{\frac{n}{2} - 1}$)

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

Thus the lemma holds. □

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

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

$$PSQ(\Theta) \subseteq f^{-1}(PSQ(\Delta)) \quad \text{and} \quad SQF(\Theta)) \subseteq f^{-1}(SQF(\Delta)).$$

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

$$PSQ(\Theta) - f^{-1}(K) - SQF(\Theta),$$

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 $SQ(\Delta)$ from $SQF(\Delta)$.

Proof: Directly from Theorem 1. □

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

Proof: Directly from Corollary 1 and from the fact that energy context-free language is an EOL language (*see*, e. g. [7]). □

We conclude this paper by the following remark. Originally the problem of separating $SQ(\Delta)$ from $SQF(\Delta)$ 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 $\# al(G) > \frac{2^{\frac{n}{4}}}{n^2}$."

ACKNOWLEDGMENTS

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

REFERENCES

1. J. M. AUTEBERT, J. BEAUQUIER, L. BOASSON and M. NIVAT, *Quelques problèmes ouverts en théorie des langages algébriques*, RAIRO Informatique Théorique, vol. 13, 1979, p. 363-379.
2. J. BERSTEL, *Sur les mots sans carré définis par un morphisme*, Lecture Notes in Computer Science, Springer-Verlag, vol. 71, 1979, p. 16-25.
3. D. R. BEAN, A. EHRENFEUCHT and G. F. MCNULTY, *Avoidable patterns in strings of symbols*, Pacific Journal of Mathematics, vol. 85, n° 2, 1979, p. 261-294.
4. A. EHRENFEUCHT and G. ROZENBERG, *On the subword complexity of square-free DOL languages*, Theoretical Computer Science, to appear.
5. M. HARRISON, *Introduction to formal language theory*, Addison-Wesley, Reading, Massachusetts, 1978.
6. J. VAN LEEUWEN, *The tape complexity of context independent developmental languages*, Journal of Computer and System Sciences, vol. 11, 1975, p. 203-211.
7. G. ROZENBERG and A. SALOMAA, *The mathematical theory of* L *systems*, Academic Press, London, New York, 1980.
8. A. SALOMAA, *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.
9. A. SALOMAA, *Jewels of formal language theory*, Computer Press, Potomac, Md., to appear.
10. A. THUE, *Ueber unendliche Zeichenreihen*, Norsk. Vid. Selsk. Skr. I Mat.-Nat. Kl., n° 7, 1906, p. 1-22.