ON REGULARITY OF LANGUAGES GENERATED BY COPYING SYSTEMS

by

A. Ehrenfeucht*

and

G. Rozenberg**

CU-CS-234-82 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.

THIS MATERIAL IS BASED UPON WORK SUPPORTED
IN PART BY THE NATIONAL SCIENCE FOUNDATION
UNDER GRANT NO. MCS 79-03838.

ABSTRACT

Let Σ be an arbitrary fixed alphabet. The direct *copying relation* (over Σ^+) is a binary relation defined by: x *copy* y if and only if $x = x_1 u x_2$ and $y = x_1 u u x_2$ for some words x_1, x_2, u where u is nonempty. The *copying* relation *copy** is defined as the reflexive and transitive closure of *copy*. A *copying system* is an ordered pair $G = (\Sigma, w)$ where $w \in \Sigma^+$; its language is $L(G) = \{z \in \Sigma^+: w \text{ copy}^* z\}$ - it is referred to as a *copy language*. This note provides a sufficient condition for a copy language to be regular; an application of this condition is demonstrated.

INTRODUCTION

The investigation of repetitions of subwords in words is a research area initiated by A. Thue ([*T*]) and since then very active in various ares of mathematics and formal language theory (see, e.g., [*BEM*], [*C*], [*D*], [*MH*], [*P*] *and* [*S1*]). The recent attention to this topic in formal language theory was brought up by J. Berstel, when he demonstrated in [*B*] an intimate connection between this area and modern formal language theory; a number of recent papers followed the observation by Berstel (see, e.g., [*Br*], [*Cr*], [*ER*], [*K*] *and* [*S2*]).

A way to understand repetitive properties (of subwords in words) of formal languages is to consider repetitions in their "pure grammatical form", that is to consider grammatical systems that explicitly use repetitions as the way of language generation.

In this note such systems are introduced and then the question of the regularity of the languages they generate is studied.

PRELIMINARIES

We use mostly standard language theoretic notation and terminology.

For a finite set A, $\#A$ denotes its cardinality; $\emptyset$ denotes the empty set.

For a positive real r, $\lceil r \rceil$ denotes the smallest positive integer n such that $n \geq r$.

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

For words x and y, x is a *subword* of y (written x *sub* y) if $y = y_1 x y_2$ for some words y_1 and y_2. (Subwords are often referred to also as *segments* or *factors).*

A nonempty word y is called *square free* if it cannot be written in the form $y_1 xx y_2$ where x is a nonempty word; the set of square free words over an alphabet Σ is denoted by $SF(\Sigma^+)$.

To simplify the notation of this paper we assume that Σ is an arbitrary but fixed finite alphabet.

The *direct copying relation*, denoted *copy*, is the binary relation in $\Sigma^+ \times \Sigma^+$ defined as follows:

x *copy* y if and only if there exist words x_1, u, x_2 with u nonempty such that $x = x_1 u x_2$ and $y = x_1 uu x_2$.

The *copying relation*, denoted *copy**, is defined to be the reflexive and the transitive closure of *copy*.

A *copying system* (*over* Σ) is an ordered pair $G = (\Sigma, w)$ where $w \in \Sigma^+$. The *language* of G is $L(G) = \{z \in \Sigma^+ : w \ copy^* \ z\}$; it is referred to as a *copy language*.

Clearly *copy** is a partial order on Σ^+; hence for a language $X \subseteq \Sigma^+$ we define the set of its *minimal elements* by $MIN(X) = \{ x \in X : y \ copy^* \ x$ implies $y = x$ for each $y \in X\}$. A language $X \subseteq \Sigma^+$ is *upward closed under copying*, written X is uc-*closed*, if $X = \{ y \in \Sigma^+ : x \ copy^* \ y$ for some $x \in X\}$.

RESULTS

In this section we provide a criterium allowing to demonstrate that certain copy languages are not regular. Then this result is used to demonstrate that a specific copying system generates a nonregular language.

Theorem 1. Let X be a uc-closed regular language such that X contains a word w with the property $alph(w) \geq 3$. Then $MIN(K)$ is infinite.

Proof:

The proof of this theorem goes through a sequence of lemmas.

Lemma 1. $MIN(\Sigma^+) = SF(\Sigma^+)$.

Proof of Lemma 1:

Obvious. ∎

Lemma 2. Let $X \subseteq \Sigma^+$ be uc-closed. Then for each $w \in X$ and each $u \in (alph(w))^*$ there exists a $z \in (alph(w))^*$ such that $w\ u\ z \in X$.

Proof of Lemma 2:

If $X = \emptyset$, then the lemma trivially holds. Hence let us assume that $X \neq \emptyset$.

The proof of this lemma goes by induction on $|u|$.

$|u| = 0$.

Then $z = \Lambda$ satisfies the statement of the lemma.

Assume that the lemma holds for all u such that

$|u| \leq k$ for some $k \geq 0$.

Consider now a word $u' \in (alph(w))^+$ such that $|u'| = k + 1$.

Hence $u' = u\ a$ where $u \in (alph(w))^*$ and $a \in alph(w)$.....(1)

By the inductive assumption there exists a word $z \in (alph(w))^*$ such that

$w\ u\ z \in X$.....(2)

By (1), w can be written in the form $w = w_1\ a\ w_2$ for some $w_1, w_2 \in (alph(w))^*$.

Hence by (2) we have $w\ u\ z = w_1\ a\ w_2\ u\ z$ and so the fact that X is uc-closed

implies that $w_1\ a\ w_2\ u\ a\ w_2\ u\ z = w\ u'\ w_2\ u\ z \in X$.....(3)

Clearly (3) completes the proof of the inductive step, and so the lemma holds. ■

Lemma 3. Let $u, x \in \Sigma^*, w \in SF(\Sigma^+)$ and let $z \in \Sigma^+$ be such that $z\ copy^*\ uwx$. Then $|z| \geq \frac{|w|}{2^{|ux|}}$.

Proof of Lemma 3:

The lemma is proved by induction on $|ux|$.

$|ux| = 0$.

Then $u\ w\ x = w \in SF(\Sigma^+)$ and so, by Lemma 1, $u\ w\ x \in MIN(\Sigma^+)$. Hence $z\ copy^*\ u\ w\ x$ implies that $z = u\ w\ x$. But clearly $|z| \geq \frac{|z|}{2^0} = \frac{|z|}{2^{|ux|}}$ and so the lemma holds. Assume now that the lemma holds whenever

$|ux| \leq k$ for some $k \geq 0$.

Consider now the case when

$|ux| = k + 1$.

Since $z\ copy^*\ u\ w\ x$, there exists a $n \geq 1$ and words $z_0, z_1, \ldots, z_n$ such that $z_0 = z$, $z_n = uwx$, $z_0\ copy\ z_1$, $z_1\ copy\ z_2, \ldots, z_{n-1}\ copy\ z_n$. Consider z_{n-1}. It must be that $z_{n-1} = y_1\ v\ y_2$ for some $v \neq \Lambda$, where $z_n = u\ w\ x = y_1\ vv\ y_2$. Since $w \in SF(\Sigma^+)$, $vv\ sub\ w$ cannot hold and so

either y_2 is a proper suffix of x,

or y_1 is a proper prefix of u.

Since these two cases are symmetric we will consider only the former case leaving the proof of the latter case to the reader.

Thus we assume that

y_2 is a proper suffix of x.....(4)

We consider separately two cases.

Case 1. Let $w = w_1w_2$ where $|w_1| \geq \lceil \frac{|w|}{2} \rceil$ and $y_1 v = u\, w_1\, t$ for some $t \in \Sigma^*$, where $t = \Lambda$ if $|w_1| < |w|$ and t is a prefix of x otherwise.

Then $z_{n-1} = y_1\, v\, y_2 = u\, w_1\, t\, y_2$. Since $z\ copy^*\ z_{n-1}$, $w_1 \in SF(\Sigma^+)$ and $|u\ t\ y_2| < |u\ x| \leq k$ (because of (4)), the inductive hypothesis implies that $|z| \geq \frac{|w_1|}{2^{|uty_2|}} \geq \frac{|w_1|}{2^k}$. From this and the fact that $|w_1| \geq \lceil \frac{|w|}{2} \rceil$ it follows that $|z| \geq \frac{|w|}{2^{k+1}}$.

Thus the induction is completed and the lemma holds if Case 1 holds.

Case 2. Let $w = w_1w_2$ where $|w_2| \geq \lceil \frac{|w|}{2} \rceil$ and $v\, y_2 = t\, w_2\, x$ for some $t \in \Sigma^*$, where $t = \Lambda$ if $|w_2| < |w|$ and t is a suffix of u otherwise.

Then $z_{n-1} = y_1\, v\, y_2 = y_1\, t\, w_2\, x$. Since $z\ copy^*\ z_{n-1}$, $w_2 \in SF(\Sigma^+)$ and $|y_1\ t\ x| < |u\ x| \leq k$ (this follows from (4) and from the observation that in Case 2 y_1 is a proper prefix of u), the inductive hypothesis implies that $|z| \geq \frac{|w_2|}{2^{|y_1tx|}} \geq \frac{|w_2|}{2^k}$. From this and from the fact that $|w_2| \geq \lceil \frac{|w|}{2} \rceil$ it follows that $|z| \geq \frac{|w|}{2^{k+1}}$.

Thus the induction is completed and the lemma holds also if Case 2 holds.

Consequently Lemma 3 holds. ■

As a direct corollary of Lemma 3 we get the following result.

Lemma 4. Let $X \subseteq \Sigma^+$ be uc-closed. Let $u, x \in \Sigma^*$, $w \in SF(\Sigma^+)$ and $z \in X$ be such that $z\ copy^*\ u\ w\ x$. Then $|z| \geq \frac{|w|}{2^{|ux|}}$. ■

The following is an obvious observation concerning regular languages.

Lemma 5. Let $X \subseteq \Sigma^*$ be a regular language. There exists a positive integer constant $q(X)$ such that, for all $u, w, x \in \Sigma^*$, $u\ w\ x \in X$ implies that $u\ w\ y \in X$ for some y such that $|y| \leq q(X)$. ∎

Now we complete the proof of the theorem as follows.

Let X be a uc-closed regular language and let w_0 be a fixed word in X such that $\#alph(w_0) \geq 3$ (by the assumption of the theorem such a word w_0 exists). For each m let $w(m) \in SF(\Sigma^+)$ be a fixed word such that $|w(m)| > m$ (since $\#alph(w_0) \geq 3$ such a word always exists, [*T*]).

Lemma 6. Let $m \geq 1$. There exists a word $z \in MIN(X)$ and there exists a word $v \in \Sigma^*$ such that $|v| \leq q(X)$ and $z\ copy^*\ w_0\ w(m)\ v$ where $q(X)$ is a constant satisfying the statement of Lemma 5.

Proof of Lemma 6:

By Lemma 2, there exists a word x such that $w_0\ w(m)\ x \in X$. Thus by Lemma 5 there exists a word $v \in \Sigma^*$ such that $w_0 w(m)\ v \in X$ where $|v| \leq q(X)$. Consequently there exists a word $z \in MIN(X)$ such that $z\ copy^*\ w_0\ w(m)\ v$.

Thus lemma 6 holds. ∎

Obviously, Lemma 6 and Lemma 4 imply the theorem. ∎

Corollary. The language of the copying system $G = (\Sigma, abc)$ is not regular.

Proof:

Obviously $L(G)$ satisfies the assumption of Theorem 1. However $MIN(L(G) = \{abc\}$ and so, by Theorem 1, $L(G)$ cannot be regular. ∎

DISCUSSION

In this note we have studied the regularity aspect of copying systems. While we are convinced that the study of copying systems should contribute to the understanding of repetitions in formal languages we realize that this note represents only a step in this direction. Two concrete problem areas suggest themselves.

(1) A characterization of regular copy languages; we have presented here only a sufficient condition.

(2) The relationship of copying systems to their "symmetric version" (that is introducing a square xx as well as reducing it to x are allowed) considered in [*Brz*].

Moreover it seems natural to consider various other standard language theoretic questions in the framework of coping systems.

ACKNOWLEDGEMENTS

The authors greatfully 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.

[Br] F. Brandenburg, Uniformly growing k-th power free homomorphisms, *Theoretical Computer Science*, to appear.

[Brz] J. Brzozowski, Open problems about regular languages, in R.V. Book, ed., *Formal language theory, perspectives and open problems*, 1980, Academic Press, London, New York.

[BEM] D.R. Bean, A. Ehrenfeucht and G.F. McNulty, Avoidable patterns in strings of symbols, 1979, *Pacific Journal of Mathematics*, v85, no.2, 261-293.

[C] A. Cobham, Uniform tag sequences, *Mathematical Systems Theory*, 1972, v.6, n.2, 164-191.

[Cr] M. Crochemore, Sharp characterizations of square free morphisms, 1982, *Theoretical Computer Science*, v. 18, 221-226.

[D] F.M. Dekking, Combinatorial and statistical properties of sequences generated by substitutions, 1980, Ph.D. Thesis, University of Nijmegen, Holland.

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

[K] J. Karhumaki, On cubic-free ω-words generated by binary morphisms, *Discrete Applied Mathematics*, to appear.

[MH] M. Morse and G. Hedlund, Unending chess, symbolic dynamics and a problem of semigroups, 1944, *Duke Math. Journal*, v. 11, 1-7.

[S1] A. Salomaa, Morphisms on free monoids and language theory, in R.V. Book, ed., *Formal language theory, perspectives and open problems*, 1980, Academic Press, London, New York.

[S2] A. Salomaa, *Jewels of formal language theory*, Computer Science Press, 1981.

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