REPETITIONS IN HOMOMORPHISMS AND LANGUAGES

by

A. Ehrenfeucht
G. Rozenberg

CU-CS-222-82

A. Ehrenfeucht
Dept. of Computer Science
Univ. of Colo. at Boulder
Boulder, Colorado 80309 USA

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

All correspondence to G. Rozenberg

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 BY THE NATIONAL SCIENCE FOUNDATION UNDER GRANT NO. MCS 79-Ø3838.

REPETITIONS IN HOMOMORPHISMS AND LANGUAGES

A. Ehrenfeucht
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado, 80309
U.S.A.

and

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

Repetitions of subwords in words form the very fundamental (combinatorial) structure of formal languages. A systematic investigation of such repetitions was initiated by Thue in [T]. Since then this problem area was a subject of an active investigation in numerous areas of mathematics and in formal language theory (see, e.g., [BEM], [C], [D], [MH], [P] and [S1]). As a matter of fact, recently one notices a revival of interest in "Thue problems" among formal language theorists (see, e.g., [B], [H], [K], [S2]). In particular it was discovered that the theory of nonrepetitive sequences of Thue [T] is strongly related to the theory of (iterative) homomorphisms on free monoïds. It was pointed out in [B] that most (if not all) examples of the so called square-free sequences constructed in the literature are either DOL sequences or their codings (see, e.g. [RS]). In this way a very significant connection was established between the theory of (non)repetitive sequences and the theory of DOL systems. It seems that the benefit is two-sided: the theory of nonrepetitive sequences originates a new and very interesting research area within the theory of homomorphisms on free monoids as conceived in the theory of DOL systems while the theory of DOL systems provides a better insight into the theory of (non)repetitive sequences (see, e.g., [B] and [S2]).

Since repetitions of subwords form such a basic structure in formal languages the research concerning the general area of Thue problems forms a very fundamental part of research in formal language theory.

In this paper we investigate "the repetitive properties" of homomorphisms and languages.

1. A CHARACTERIZATION OF SQUARE-FREE HOMOMORPHISMS

Let Σ be a finite nonempty alphabet. A word $x \in \Sigma^+$ is called a *pure square* if $x = yy$ for some $y \in \Sigma^+$; x is called a *square* if x contains a subword which is a pure square, otherwise x is called *square-free* . We use $SQ(\Sigma)$ and $SF(\Sigma)$ to denote the set of all squares over Σ and the set of all square-free words over Σ respectively. For a finite nonempty alphabet Δ we use $HOM(\Sigma,\Delta)$ to denote the set of all homomorphisms from Σ^*

into Δ^* . A homomorphism $h \in HOM(\Sigma,\beta)$ is called *square-free* if $(h(x) \in SF(\Delta)$ whenever $x \in SF(\Sigma)$. Hence square-free homomorphisms are homomorphisms preserving the square-free property; they form an important subject of investigation in the theory of (non) repetitive sequences and languages (see, e.g., [B],[S]).

Let $h \in HOM(\Sigma,\Delta)$. Then

$$T_h = \{w \in SF(\Sigma) : (\exists a,b)_{\Sigma} (\exists u)_{\Sigma^*} [w = aub \text{ and } \text{either } h(u) \sqsubseteq h(a) \text{ or } h(u) \sqsubseteq h(b)]\},$$

where for words x, y we write $x \sqsubseteq y$ if x is a subword of y. Also let

$$T_0 = \{w \in SF(\Sigma) : |w| \leq 3\}.$$

We have obtained the following structural characterization of square-free homomorphisms.

Theorem 1. Let $h \in HOM(\Sigma, \Delta)$. Then h is square-free if and only if $h(T_0 \cup T_h) \subseteq SF(\Delta)$. □

A well-known result by Thue (see [T] and also [BEM]) says that a sufficient condition for a homomorphism $h \in HOM(\Sigma,\Delta)$ to be square-free is as follows:

(1). $(\forall a,b)_{\Sigma}[h(a) \sqsubseteq h(b)$ implies $a = b]$ and

(2). $h(T_0) \subseteq SF(\Delta)$.

It is easily seen that this theorem by Thue is a simple corollary of our Theorem 1.

Now, for a homomorphism $h \in HOM(\Sigma,\Delta)$ let $maxr(h) = \max\{|h(a)| : a \in \Sigma\}$ and $minr(h) = \min\{|h(a)| : a \in \Sigma\}$, where for a word x, $|x|$ denotes its length. In [B] Berstel proves the following result:

a homomorphism $h \in HOM(\Sigma,\Sigma)$ is square-free if and only if $h(x) \in SF(\Sigma)$ for each square-free word x such that $|x| \leq 2 + \left\lfloor 2 \frac{maxr(h)}{minr(h)} \right\rfloor$.

Based on our theorem 1 we can prove the following result.

Theorem 2. A homomorphism $h \in HOM(\Sigma,\Sigma)$ is square-free if and only if $h(x) \in SF(\Sigma)$ for each square-free word x such that $|x| \leq 2 + \left\lfloor \frac{maxr(h)}{minr(h)} \right\rfloor$. □

Since $\left\lfloor \frac{2maxr(h)}{minr(h)} \right\rfloor \geq \left\lfloor \frac{maxr(h)}{minr(h)} \right\rfloor + 1$ our bound is strictly better than this of the Berstel theorem mentioned above.

2. ON SQUARE-FREENESS TEST SETS

The characterization results discussed in the last section provide one with "test sets" for testing the square-freeness of a homomorphism. A homomorphism $h \in HOM(\Sigma,\Delta)$ is square-free if h(x) is square-free for all $x \in SF(\Sigma)$. Since $SF(\Sigma)$ is infinite for $\# \Sigma \geq 3$ (where for a finite set A, $\#A$ denotes its cardinality) such a definition is not effective. On the other hand the results from the last section allow one, given a homomorphism h, to construct effectively a finite set F_h (of square-free words), such that h is square-free if and only if h(x) is square-free for every $x \in F_h$. In this sense such a set F_h is called a *square-free test set*. We will look now more closely into square-freeness test sets referred in this paper simply as *test sets*.

Thus given a homomorphism $h \in HOM(\Sigma,\Delta)$ we say that a set $X \subseteq \Sigma^{+}$ *tests* h if and only if $(h(X) \subseteq SF(\Delta))$ implies $(h(SF(\Sigma)) \subseteq SF(\Delta))$. Consequently Theorem 2 can be restated as follows.

Theorem 2'. Let $h \in HOM(\Sigma,\Delta)$. Then $\{w \in SF(\Sigma) : |w| \leq 2 + \left\lfloor \frac{maxr(h)}{minr(h)} \right\rfloor \}$. □

In order to make the test set smaller one would like to replace the "≤" sign from the above result by the "=" sign. Indeed this can be done under an additional assumption (the reader should be warned that the construction is not trivial!). In what follows, for a finite set A, #A denotes the cardinality of A.

Theorem 3. Let $h \in HOM(\Sigma,\Delta)$, $\#\Sigma \geq 3$ and let $m \geq 2 + \left\lfloor \frac{maxr(h)}{minr(h)} \right\rfloor$.
Then $\{w \in SF(\Sigma) : |w| = m\}$ tests h. □

It is easily seen that the above theorem is false if $\#\Sigma < 3$.

The tests sets we have considered above were "adjusted to h" in the sense that very specific parameters concerning h were used to define these test sets (namely *maxr*(h) and *minr*(h)). A natural next step is to consider test sets which will be universal for all homomorphisms in $HOM(\Sigma,\Delta)$ with fixed Σ and Δ. This can be done as follows.
Let $\Sigma_{\omega} = \{a_1,a_2,\ldots\}$ be a fixed infinite alphabet and then let, for each $n \geq 1$, $\Sigma_n = \{a_1,\ldots,a_n\}$. Let $n,m \geq 1$. The family $T(n,m)$ of (n,m) *test sets* is defined as follows:

$X \in T(n,m)$ if and only if
$X \subseteq SF(\Sigma_n)$ and $(\forall h)_{HOM(\Sigma_n,\Sigma_m)} [h(X) \subseteq SF(\Sigma_m)$ if and only if h is square-free].

Clearly, we are interested in the existence of *finite* (n,m) test sets.
Here we have the following result.

Theorem 4. Let $n,m \geq 1$. Then $T(n,m)$ contains a finite nonempty set if and only if either $n \leq 3$ or $m \leq 2$. □

3. REPETITIVENESS AND STRONG REPETITIVENESS

An example of repetitions (of subwords in words) in formal languages is the effect of pumping in an infinite context-free language. Then we get a word, say w, such that all its powers (repetitions) appear in words of the given language. This idea can be formalized in two different (weaker and stronger) forms. (For a language K, *sub*(K) denotes the set of its subwords).

A language $K \subseteq \Sigma^{*}$ is called *repetitive* if and only if
$(\forall n)_{\geq 1}\ (\exists w)_{\Sigma^{+}} [w^{n} \in sub(K)]$; K is called *strongly repetitive* if
$(\exists w)_{\Sigma^{+}} (\forall n)_{\geq 1} [w^{n} \in sub(K)]$.

Clearly, every strongly repetitive language is also repetitive, but the converse does not have to be true in general. As the direct consequence of the pumping lemma we get that every infinite context-free language is strongly repetitive, and so repetitiveness implies strong repetitiveness in a trivial way.

However the situation in DOL languages is much more involved. The pumping-like properties do not hold for DOL languages and "detecting" repetitiveness in a DOL language becomes a challenging problem. We have obtained the following result.

Theorem 5. It is decidable whether or not L(G) is repetitive for an arbitrary DOL system.

This result yields also the decidability of the strong repetitiveness property because we have the following.

Theorem 6. Let K be a DOL language. Then K is repetitive if and only if K is strongly repetitive.

4. COPYING SYSTEMS

From the existing literature concerning "Thue problems" one can certainly draw the conclusion that this problem area is mathematically quite challenging. On the other hand repetitions (in languages and homomorphisms) play an important role in formal language theory and so their nature should be well understood. Thus the topic of repetitions (in languages and homomorphisms) forms an interesting and well motivated research topic.

A way to understand repetitiveness in formal languages is to consider repetitions in their "pure grammatical form", that is introduce grammatical systems that explicitly use repetitions as the way of language generation.

A *copying system* is an ordered pair G = (Σ, w) where Σ is a finite nonempty alphabet and $n \in \Sigma^*$. Then for words $u, w \in \Sigma^*$ we say that u *directly derives* w, written $u \underset{G}{\Rightarrow} w$, if u = xyz and w = xyyz for some $x, y, z \in \Sigma^*$. Then $\overset{*}{\underset{G}{\Rightarrow}}$ denotes the reflexive and the transitive closure of $\underset{G}{\Rightarrow}$. The *language of*G is defined by $L(G) = \{x \in \Sigma^* : w \overset{*}{\underset{G}{\Rightarrow}} x\}$; L(G) is referred to as a *copying language*. Analyzing copying languages turns out to be a difficult task. (The reader should consider the problem of proving or disproving whether the language of the copying system G = $(\{a,b,c\}, abc)$ is context-free).

We will provide now a result allowing one to prove that certain copying languages are not regular.

Given a copying system G = (Σ, w), the relation $\overset{*}{\underset{G}{\Rightarrow}}$ is a partial order on Σ^*. Hence for a language $K \subseteq \Sigma^*$ we can distinguish the set of minimal elements of K, *min*(K) = $\{x \in K : (\forall y)_K[\text{if } y \overset{*}{\underset{G}{\Rightarrow}} x \text{ then } x = y]\}$. Also we say that K is *upward closed* under $\overset{*}{\underset{G}{\Rightarrow}}$ if $(\forall x,y)_{\Sigma^*}[\text{if } x \in K \text{ and } x \overset{*}{\underset{G}{\Rightarrow}} y \text{ then } y \in K]$. In what follows, for a word z, *alph*(z) denotes the set of letters occurring in z.

Theorem 6. Let G = (Σ, w) be a copying system and let $K \subseteq \Sigma^*$. If

(1). K is regular,

(2). K is upwards closed under $\overset{*}{\underset{G}{\Rightarrow}}$, and

(3). $(\exists x)_K[\# alph(x) \geq 3]$

then *min*(K) is infinite

As an application of this result we can show that

Corollary. Let G = ({a,b,c}, abc). Then L(G) is not regular. □

We don't know of any other way to prove the above corollary.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of NSF grant MSC 79-03838.

REFERENCES

[B] J. Berstel, Sur les mots sans carré définis par un morphisme, 1979, *Springer Lecture Notes in Computer Science*, v71, 16-25.

[BEM] D.R. Bean, A. Ehrenfeucht and G.F. Mc Nulty, 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.

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

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

[K] J. Karhumaki, On cubic-free ω-words generated by binary morphisms, 1981 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.

[P] P.A. Pleasants, Non-repetitive sequences, 1970, *Proc. Cambridge Phil. Society*, v.68, 267-274.

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

[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, 141-166.

[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.