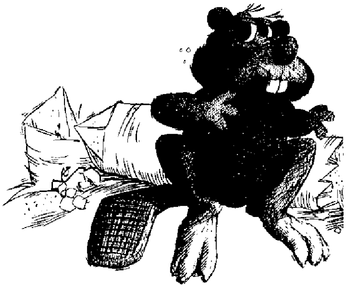
UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT



Checking Sets, Test Sets, Rich Languages and Commutatively-Closed Languages

Jürgen Albert
Derick Wood

CS-82-28

September, 1982

CHECKING SETS, TEST SETS, RICH LANGUAGES AND COMMUTATIVELY-CLOSED LANGUAGES[1)]

by

Jürgen Albert[2)] and Derick Wood[3)]

[1)] This work was partially supported by grants from the Deutsche Forschungsgemeinschaft (DFG) and from the Natural Sciences and Engineering Research Council of Canada, Grant No. A-7700.

[2)] Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe, D-7500 Karlsruhe, West Germany.

[3)] Department of Computer Science, University of Waterloo, Ontario N2L 3G1, Canada.

Abstract

The problem of homomorphism equivalence is to decide for some language L over some finite alphabet Σ and two homomorphisms f and g whether or not $f(x) = g(x)$ for all x in L. It has been conjectured that each L can be represented by some finite subset F such that for all pairs of homomorphisms f and g:

$$f(x) = g(x) \text{ for all } x \text{ in } F \text{ implies } f(x) = g(x) \text{ for all } x \text{ in } L.$$

We prove this conjecture for the families of rich and commutatively-closed languages. We derive lower and upper bounds for the sizes of these finite subsets and give examples of language families for which there are effective constructions of these subsets.

1. Introduction

Although a homomorphism can be considered to be the simplest function which translates a word (letter by letter) into some other word, many basic problems concerning pairs of homomorphisms are still open.

For example, given two homomorphisms f and g over some alphabet Σ, does there exist some x in Σ^+ such that $f(x) = g(x)$? This is a re-formulation of the well known Post Correspondence Problem, however there are some recent results about the decidability of restricted versions of it. Despite this the minimal size of Σ such that the Post Correspondence Problem over Σ becomes undecidable remains open.

We can use two homomorphisms in this way to define a language, their so called equality set. For two homomorphisms f and g their equality set $E(f,g)$ is defined as $\{x \text{ in } \Sigma^* | f(x) = g(x)\}$. The power of this mechanism has been demonstrated by yielding elegant characterizations for the recursively enumerable sets and very general families of complexity classes including NP [4].

The study of test sets can be considered as dual to this approach. Whereas for equality sets we fix a pair of homomorphisms and generate a language from them, we now fix a language and all those pairs of homomorphisms are considered which have the same images for all words in the language, i.e. all homomorphisms f, g for which $E(f,g)$ is a superset of the given language. To test this property for some f and g it would suffice for our language L to always be effectively represented by some finite subset $F \subseteq L$. Such an F is then called a test set for L.

Test sets can be constructed effectively for regular and context-free languages and are known to exist for all languages over a binary alphabet [1], and [4].

As applications of these results we obtain algorithms to decide the equivalence of deterministic gsm-mappings on context-free languages and theorems about the reducibility of systems of string equations of certain types [1], [6], and [13].

In this paper we treat the problem of test set existence for the families of rich [7] and commutatively-closed [10] languages. It turns out that in both cases the characterization of the languages by their sets of Parikh vectors is sufficient in a sense to be made precise later. Before proving our results we need some notation.

Let L be a language over some alphabet Σ and g and h be two homomorphisms defined on Σ^*. We say g and h agree on L, denoted by $g \underset{L}{\equiv} h$, if for all words x in L, $g(x) = h(x)$. We say that a finite set $F \subseteq \Sigma^*$ is a checking set for L iff for all pairs of homomorphisms g and h defined on Σ^*, g and h agree on F iff g and h agree on L. If further $F \subseteq L$ then we say that F is a test set for L. These notions were first introduced by Culik II and Salomaa [8] and have been subsequently studied in [1], [6], [7] and [9]. [4] contains a survey of recent results.

Associated with the notion of a test set are three fundamental problems. These are:

The Test Set Existence Problem: Does every language have a test set? Ehrenfeucht has conjectured that this is indeed the case.

The Test Set Construction Problem: Given an arbitrary language can its test set be constructed effectively, if it has one? In general this is surely not the case, however [1] demonstrates effectiveness for context-free languages.

The Test Set Decision Problem: Given an arbitrary language $L \subseteq \Sigma^*$ and an arbitrary finite set $F \subseteq \Sigma^*$ then is F a test set for L?

In this paper we consider these three questions for the families of rich languages (Section 2) and commutatively-closed languages (Section 3). A language $L \subseteq \Sigma^*$ is rich iff for all homomorphisms g and h such that $g(x) = h(x)$ for all x in L then g and h are identical. We say that L is commutatively-closed if for every word x in L, L contains every word y in Σ^* which has the same Parikh vector.

We are able to solve the existence problem affirmatively for these languages and also give partial results for the two other problems.

Our approach also leads to another interesting result, namely a necessary condition for a set to be a checking set and hence a test set for an arbitrary language. This also gives a lower bound on the size of a test set. We are also able to show that every language which has a checking set has a checking set of size given by this lower bound result. Moreover Ehrenfeucht has conjectured that an upper bound on the size of test sets is 2^n where n is the cardinality of the alphabet. In the case of commutatively-closed languages we are able to derive an upper bound of $2^n(n! + n) + 5n^2$.

2. Checking and Test Sets For Languages and Rich Languages

The notions and definitions introduced here are mostly standard and can be found in most textbooks on formal language theory, e.g. [11], [12], and [14]. Let $\#_a(w)$ denote the number of occurrences of the letter a in a word w, and alph(w) the set of letters occurring in w.

If $\Sigma = \{a_1, \cdots, a_n\}$ is an alphabet and w is in Σ^*, the n-dimensional vector $p(w) = (\#_{a_1}(w), \#_{a_2}(w), \cdots, \#_{a_n}(w))$ is called the Parikh vector of w.

In the following we will use in the proofs some basic facts about combinatorial properties of words. Readers unfamiliar with these results are referred to [11, Section 11.3] for background material.

We begin with a general lemma giving a necessary condition for a finite set to be a checking set for an arbitrary language.

<u>Lemma 1</u>

Let $\Sigma = \{a_1, \cdots, a_n\}$ be an alphabet and $L \subseteq \Sigma^*$. If F is a checking set for L then there are $y_1, \cdots, y_m$ in F, $m < n$, such that for each x in L there exist rational numbers $\alpha_1, \cdots, \alpha_m$ with $p(x) = \sum_{i=1}^{m} \alpha_i \cdot p(y_i)$.

<u>Proof</u>: Let $y_1, \cdots, y_m$ be in F such that $p(y_1), \cdots, p(y_m)$ constitutes a basis for p(F). We now show that there is a pair of homomorphisms $h_1, h_2 : \Sigma^* \to \{a\}^*$ such that $h_1 \underset{F}{\equiv} h_2$ and $h_1(x) \neq h_2(x)$ if $p(y_1), \cdots, p(y_m), p(x)$ are linearly independent.

Let $p(x) = (x_1, \cdots, x_n)$ and for $i = 1, \cdots, m$ let $p(y_i) =$

$(y_{i1}, \cdots y_{in})$.

If $p(y_1), \cdots, p(y_m), p(x)$ are linearly independent the system of equations:

$$\begin{aligned} &y_{11} \cdot \Delta_1 + y_{12} \cdot \Delta_2 + \cdots + y_{1n} \cdot \Delta_n = 0 \\ &\vdots \\ &y_{m1} \cdot \Delta_1 + y_{m2} \cdot \Delta_2 + \cdots + y_{mn} \cdot \Delta_n = 0 \\ &x_1 \cdot \Delta_1 + x_2 \cdot \Delta_2 + \cdots + x_n \cdot \Delta_n = 1 \end{aligned}$$

has at least one solution $(\Delta_1, \cdots, \Delta_n)$ of rational numbers Δ_i by the basic theorems on the rank of such systems.

Let $\Delta_i = \frac{\alpha_i}{\beta_i}$, where α_i is in $\mathbb{Z}$, and β_i is in $\mathbb{N}$, for $i = 1, \cdots, n$.

Now we complete the proof by exhibiting homomorphisms h_1 and h_2 such that $h_1 \underset{F}{\equiv} h_2$ and $|h_1(x)| - |h_2(x)| \neq 0$. To this end let $k = \beta_1 \cdot \beta_2 \cdot \cdots \cdot \beta_n$, choose natural numbers σ_i, τ_i such that $\sigma_i - \tau_i = k \cdot \Delta_i$ and define $h_1(a_i) = a^{\sigma_i}, h_2(a_i) = a^{\tau_i}$ for $i = 1, \cdots, n$. Clearly, by the above system of equations

$$|h_1(y_i)| - |h_2(y_i)| = y_{i1} \cdot (\sigma_1 - \tau_1) + \cdots + y_{in} \cdot (\sigma_n - \tau_n) = 0$$

for $i = 1, \cdots, m$ and

$$|h_1(x)| - |h_2(x)| = k \neq 0 \qquad \blacksquare$$

An immediate corollary to this lemma gives a lower bound on the size of checking sets.

Corollary 2

Let L be an arbitrary language with a checking set F and let a basis for p(L) have size m. Then #F > m.

We now state a trivial result which has far reaching consequences.

Lemma 3

Let $L \subseteq \Sigma^*$ be an arbitrary languge and $F \subseteq \Sigma^*$ be an arbitrary finite language.

Then F is a checking set for L iff F is a checking set for L^*. Moreover if F is a test set for L then it is a test set for L^*.

Proof: Trivial, since (i) for homomorphisms h_1, h_2 and words $x_1, \cdots, x_k$ in L, $h_1(x_i) = h_2(x_i)$ implies $h_1(x_1 \cdots x_k) = h_2(x_1 \cdots x_k)$ and (ii) $L \subseteq L^*$. ∎

Definition

Let $L \subseteq \Sigma^*$ be an arbitrary language and $F \subseteq \Sigma^*$ be a checking set for L. We say F is minimal if #F is the size of a basis for p(L). We define minimal test sets analogously.

Theorem 4

Let $L \subseteq \Sigma^*$ be an arbitrary language having a checking set. Then L has a minimal checking set. If L has a test set then L^* has a minimal test set.

Proof: We will only prove the first statement. The second follows from it together with Lemma 3.

Let $F \subseteq \Sigma^*$ be a checking set for L and let $F = \{x_1, \cdots, x_m, w_1, \cdots, w_n\}$ where $\{p(x_i) : 1 < i < m\}$ forms a basis for p(L). If $n = 0$ then F is already minimal, therefore consider $n > 0$.

Now $p(w_n) = \sum_{i=1}^{m} \alpha_i \cdot p(x_i)$ for rational α_i, not all of which are negative. Hence without loss of generality assume $\alpha_1 > 0$.

Let $F' = \{x_1 w_n, x_2, \cdots, x_m, w_1, \cdots, w_{n-1}\}$. We will show that F' is a checking set for F and hence for L. It is well known that $\{p(x_1 w_n), p(x_2), \cdots, p(x_m)\}$ is also a basis for p(L). Consider h_1 and h_2 with $h_1 \underset{F'}{\equiv} h_2$. We will show that $h_1 \underset{F}{\equiv} h_2$. In other words we will show that $h_1(x_1 w_n) = h_2(x_1 w_n)$ implies $h_1(x_1) = h_2(x_1)$ and hence $h_1(w_n) = h_2(w_n)$.

$$
\begin{aligned}
\text{Now}\quad & p(h_1(x_1 w_n)) - p(h_2(x_1 w_n)) \\
&= p(h_1(x_1)) + p(h_1(w_n)) - p(h_2(x_1)) - p(h_2(w_n)) \\
&= (1 + \alpha_1)(p(h_1)x_1)) - p(h_2(x_1))) \\
&= 0
\end{aligned}
$$

by the representation of $p(w_n)$ given above and because $x_1 w_n, x_2, x_3, \cdots, x_m$ in F' implies $p(h_1(x_1 w_n)) = p(h_2(x_1 w_n))$ as well as $p(h_1(x_i)) = p(h_2(x_i))$ for $i = 2, 3, \cdots m$. Now because $\alpha_1 > 0$ we have $p(h_1(x_1)) = p(h_2(x_1))$ and hence $|h_1(x_1)| = |h_2(x_1)|$. But this means that in the equation

$$h_1(x_1 w_n) = h_2(x_1 w_n)$$

which can be written as

$$h_1(x_1)h_1(w_n) = h_2(x_1)h_2(w_n)$$

we have $h_1(x_1) = h_2(x_1)$. Therefore $h_1(w_n) = h_2(w_n)$ also.

We have replaced F by a checking set F' satisfying $\#F' < \#F$. Clearly this procedure can be iterated to obtain a minimal checking set. ∎

<u>Remark</u>. In [7] it is shown that the language $L = \{a^n b^n | n > 1\}$ cannot have a test set consisting of only one word, i.e. there exist languages which do not possess minimal test sets.

As an application of these results we consider the collection of rich languages. A language $L \subseteq \Sigma^*$ is <u>rich</u> if for all homomorphisms g and h satisfying $g \underset{L}{\equiv} h$, we have $g \underset{\Sigma^*}{\equiv} h$.

<u>Theorem 5</u>

Let $\Sigma = \{a_1, \cdots, a_m\}$ and $L \subseteq \Sigma^*$.
Then L is rich iff there exist $x_1, \cdots, x_m$ in L such that $p(x_1), \cdots, p(x_m)$ are linearly independent.

<u>Proof</u>: (→) Since L is rich it must contain, by the arguments in Lemma 1, words $x_1, \cdots, x_m$ with $p(x) = \sum_{i=1}^{m} \alpha_i \cdot p(x_i)$, for some α_i, for all x in Σ^*. (Essentially L is a possibly infinite checking set for Σ^*.)
(←) Each $p(a_i)$ can be expressed as $\alpha_{i1} \cdot p(x_1) + \cdots + \alpha_{im} \cdot p(x_m)$. Whenever $h_1 \underset{L}{\equiv} h_2$ for two homomorphisms h_1, h_2, it follows that $|h_1(a_i)| = |h_2(a_i)|$ for all i in $\{1, \cdots, m\}$. This immediately implies $h_1(a_i) = h_2(a_i)$

for all i and thus $h_1 \underset{\Sigma^*}{\equiv} h_2$.

Corollary 6 -- Test Set Existence

Every rich language has a test set and moreover it has a minimal test set of the same size as its alphabet.

Corollary 7 -- Test Set Construction

Given an arbitrary rich context-free language L a test set for L can be found effectively.

This follows from the fact that p(L) is semi-linear when L is context-free and hence a basis for p(L) can be found effectively. This result can be strengthened by observing that richness is decidable for context-free languages.

Corollary 8 -- Test and Checking Set Decision

Given an arbitrary rich context-sensitive language $L \subseteq \Sigma^*$ and an arbitrary finite set $F \subseteq \Sigma^*$ it is decidable whether or not F is a test set for L.

Given an arbitrary rich language $L \subseteq \Sigma^*$ and an arbitrary finite set $F \subseteq \Sigma^*$ it is decidable whether or not F is a checking set for L.

The Corollary follows by observing that a finite set F is a checking set for an arbitrary rich language iff F is rich itself. Thus, testing whether F is a test set for L involves checking whether or not p(F) contains $\#\Sigma$ linearly independent vectors and testing if $F \subseteq L$. The latter test is effective for context-sensitive languages.

Remark. Richness is undecidable for context-sensitive languages $L \subseteq \Sigma^*$, since there is no algorithm to find the minimal alphabet Σ' such that $L \subseteq (\Sigma')^*$.

3. Commutatively-Closed Languages

We define the commutative closure of a language $L \subseteq \Sigma^*$, denoted by $c(L)$, by: $\{x \text{ in } \Sigma^* : p(x) = p(y) \text{ for some } y \text{ in } L\}$. We say L is commutatively-closed if $L = c(L)$.

Since a commutatively-closed language L is in some sense representable by its set of Parikh vectors $p(L)$ one is led to think that any basis of $p(L)$ can be chosen as a test set for L. The following example demonstrates that this is, in general, not the case.

Example

Let $L = \{x \text{ in } \{a, b\}^* : \#_a(x) = \#_b(x)\}$.

Then $F = \{ab, ba\}$ is not a test set for L.

Consider h_1, h_2: $\{a, b\}^* \to \{0, 1\}^*$ defined by:

$$\begin{aligned} h_1(a) &= 010 \qquad & h_2(a) &= 0 \\ h_1(b) &= 1 \qquad & h_2(b) &= 101 \end{aligned}$$

Then
$$\begin{aligned} h_1(ab) &= 0101 = h_2(ab), \\ h_1(ba) &= 1010 = h_2(ba) \end{aligned}$$

but
$$\begin{aligned} h_1(aabb) &= 01001011 \\ h_2(aabb) &= 00101101 \end{aligned}$$

The proof of our main theorem shows that $F = \{aabb, abab, abba, baab, baba, bbaa\}$ can be chosen as a test set for L, for example.

Definition

Let $L \subseteq \Sigma^*$ be any commutatively-closed language and let $F \subseteq L$ be finite and commutatively-closed. We say that F has property c1 if:

For each z in L there exist $x_1, \cdots, x_m$ in F and rational numbers $\alpha_1, \cdots, \alpha_m$ such that $alph(x_i) = alph(z)$ for $i = 1, \cdots, m$ and

$$p(z) = \sum_{i=1}^{m} \alpha_i \cdot p(x_i),$$

and F has property c2 if:

For each z in L and each a in Σ occurring more than once in z there is a y in F such that $alph(y) = alph(z)$ and a also occurs more than once in y.

Theorem 9

Let $L \subseteq \Sigma^*$ be a commutatively-closed language and F a finite commutatively-closed subset of L with properties c1 and c2.

Then F is a test set for L.

Proof: Let z be in L - F and h_1, h_2 be two homomorphisms with $h_1 \underset{F}{\equiv} h_2$. For $\Delta = alph(z)$ we can assume that there is an α in Δ with $h_1(\alpha) \neq h_2(\alpha)$. Otherwise $h_1(z) = h_2(z)$ holds trivially.

The set Δ is now partitioned as follows.

Let Δ_1 be the set of all letters in Δ which occur exactly once in all y in F satisfying $alph(y) = \Delta$. Let $\Delta_2 = \Delta - \Delta_1$.

For our proof that $h_1(z) = h_2(z)$ we show that - except for one trivial subcase - all homomorphic images of all letters of Δ commute.

Let us consider the following two cases:

Case 1: There is an a in Δ_2 with $h_1(a) \neq h_2(a)$.

Case 2: There is a b in Δ_1 with $h_1(b) \neq h_2(b)$ and $h_1(c) = h_2(c)$ for all c in Δ_2.

Case 1: Let $w = h_1(a)$, $w' = h_2(a)$. By property c2 of F and because F is commutatively-closed there are words

aya, aay in F, where alph(ay) = Δ

and

(1) $wh_1(y)w = w'h_2(y)w'$.

(2) $wwh_1(y) = w'w'h_2(y)$.

Without loss of generality we can assume $|w| > |w'|$.

Since w, w' are prefixes (suffixes) of the same word we have $w = w'x = \bar{x}w'$ for some x and $\bar{x}$ with $|x| = |\bar{x}| > 0$.

By substituting for w with $w'x$ and $\bar{x}w'$ in the equations (1) and (2) we obtain:

(1') $xh_1(y)\bar{x} \quad = h_2(y)$ giving

(2') $xw'x\ h_1(y) = w'xh_1(y)\ \bar{x}$.

By (2') xw' and $w'x$ are prefixes of the same word. Since they are of the same length, they must be equal, that is $xw' = w'x$. Because $w = w'x = \bar{x}w'$ we also conclude $\bar{x} = x$. This gives two simpler equations:

(1") $x\ h_1(y)x = h_2(y)$,

(2") $x\ h_1(y) = h_1(y)x$.

By a basic theorem on commuting words equation (2') implies the existence of a non-empty word u and numbers $i > 1$, $j > 0$ such that $x = u^i$ and $h_1(y) = u^j$; cf. [11, pp. 9 and 12]. Choosing u to be of minimal length, determines u uniquely as the "primitive root" of x. Now if y' is a word having the same Parikh vector as y then $ay'a$, aay' are in F as well. As above we derive $xh_1(y') = h_1(y')x$. Since u has been chosen uniquely and $|h_1(y')| = |h_1(y)|$ it follows that $h_1(y') = h_1(y)$.

Thus $h_1(y)$ is not changed if the letters in y are permuted.

Now by (1") $h_2(y) = xh_1(y)x = xh_1(y')x = h_2(y') = u^{2i+j}$.

Let b be in the alph(y) and $h_1(b)$ be non-empty then $h_1(b)h_1(\bar{y}) = h_1(\bar{y})h_1(b)$ where $aab\bar{y}$ and $aa\bar{y}b$ are in F and $\text{alph}(ab\bar{y}) = \Delta$. Again by [11, pp. 9 and 12] there is a unique non-empty word v, the primitive root of $h_1(b)$, and numbers $i' > 1$, $j' > 0$ such that $h_1(b) = v^{i'}$ and $h_1(\bar{y}) = v^{j'}$. Since $h_1(b)h_1(\bar{y}) = v^{i'+j'} = u^j$ and v and u are primitive it follows that $v = u$ and $j = i' + j'$.

Similarly if $h_2(c)$ is non-empty and c is in Δ, there is a number $k > 1$

such that $h_2(c) = u^k$.

Now define $r(d)$, $s(d) > 0$ such that $h_1(d) = u^{r(d)}$ and $h_2(d) = u^{s(d)}$ for each d in Δ and for each d in $\Sigma - \Delta$ define $r(d) = s(d) = 0$. Let $\underline{\bar{r}}$ and $\underline{\bar{s}}$ denote the row vectors

$$\underline{\bar{r}} = (r(a_1), r(a_2), \cdots, r(a_n)),$$
$$\underline{\bar{s}} = (s(a_1), s(a_2), \cdots, s(a_n)),$$

where $\Sigma = \{a_1, a_2, \cdots, a_n\}$.

By property c1 of F there exist $y_1, y_2, \cdots, y_m$ in F with $\text{alph}(y_i) = \Delta$ and rational numbers $\alpha_1, \cdots, \alpha_m$ such that $p(z) = \sum_{i=1}^{m} \alpha_i \cdot p(y_i)$.

Thus $h_1(z) = u^{\sigma_1}$, $h_2(z) = u^{\sigma_2}$, where

$$\sigma_1 = p(z) \cdot \underline{\bar{r}}^T = \sum_{i=1}^{m} \alpha_i \cdot p(y_i) \cdot \underline{\bar{r}}^T,$$

and $\sigma_2 = p(z) \cdot \underline{\bar{s}}^T = \sum_{i-1}^{m} \alpha_i \cdot p(y_i) \cdot \underline{\bar{s}}^T,$

where we use T to denote transpose.

Because y_i is in F and $h_1(y_i) = h_2(y_i)$ it follows that $p(y_i) \cdot \underline{\bar{r}}^T = p(y_i) \cdot \underline{\bar{s}}^T$ for $i = 1, \cdots, m$ and thus $h_1(z) = h_2(z)$.

<u>Case 2</u>: There is a b in Δ_1 with $h_1(b) \neq h_2(b)$ and for all c in Δ_2 $h_1(c) =$

$h_2(c)$.

Let us first consider the following simple subcase:

Subcase 2.1: $h_1(c) = h_2(c) = \lambda$ for all c in Δ_2.

Because of c2 there is a word y in F such that $h_1(z) = h_1(y) = h_2(y) = h_2(z)$.

Subcase 2.2: We have $h_1(b) \neq h_2(b)$ for some b in Δ_1, $h_1(c) = h_2(c)$ for all c in Δ_2 and there is an a in Δ_2 such that $h_1(a) \neq \lambda$.

Without loss of generality assume that $|h_1(b)| > |h_2(b)|$ and that there are words aaby, aayb, abay, ayab in F with alph(aby) = Δ such that h_1 and h_2 agree on these words.

For $w = h_1(a) = h_2(a)$ and $v = h_1(b)$, $v' = h_2(b)$ we have $v = v'x = \bar{x}v'$ for some x, $\bar{x} \neq \lambda$. As in Case 1 we can now derive from $h_1(t) = h_2(t)$ for t = aaby, aayb, abay, ayab that

$xw = wx$, $\bar{x}w = w\bar{x}$ and therefore $x = \bar{x}$, and $xh_1(y) = h_1(y)x$ for all y such that aaby is in F and alph(aby) = Δ. This implies again that all homomorphic images of all letters in Δ commute and $h_1(z) = h_2(z)$. ∎

It should be obvious that every commutatively-closed language L has a finite commutatively-closed subset F satisfying properties c1 and c2.

Thus, Theorem 9 implies

Corollary 10 -- Test Set Existence

Every commutatively-closed language has a test set.

As an application of Theorem 9 we obtain explicit test sets for some special commutatively-closed languages.

Corollary 11

Let $\Sigma = \{a_1, \cdots, a_m\}$ be an alphabet and $L = \{x \text{ in } \Sigma^* : \#_{a_1}(x) = \#_{a_2}(x) = \cdots = \#_{a_m}(x)\}$.

Then for each $i > 2$, $F_i = \{x \text{ in } L : \#_{a_j}(x) = i \text{ for } j = 1, \cdots m\}$ is a test set for L.

The languages F_i obviously satisfy properties c1 and c2 and therefore they are test sets for L.

Corollary 12 -- Test Set Construction

For L an arbitrary commutatively-closed context-free language a test set F for L can be effectively found.

This follows from conditions c1 and c2 for such an F. We also have:

Corollary 13 -- Test Set Decision

For L an arbitrary commutatively-closed context-free language and F an arbitrary commutatively-closed finite set it is decidable whether or not F is a test set for L.

Finally, it should be mentioned that Theorem 9 can be generalized to include all (1, 2)-complete languages. A language L Σ^* is *(1, 2)-complete* if for any pair of letters a, b in Σ, $a \neq b$, such that a occurs at least twice in a word of L and there exists an x in L with $\{a, b\} \subseteq \mathrm{alph}(x)$ then there is a y in Σ^* with aaby, aayb, abay, abya, ayab in L.

This leads to an upper bound result on the size of the test sets namely:

Corollary 14

Let $L \subseteq \Sigma^*$ be (1, 2)-complete. Then there is a test set F for L, such that $\#F < 2^n(n! + n) + 5n^2$ where $n = \#\Sigma$.

Proof: It is easy to see that every (1, 2)-complete language $L \subseteq \Sigma^*$ contains a finite subset F with the following properties:

c1: For each z in L there exist $y_1, \cdots, y_m$ in F and rational numbers $\alpha_1, \cdots, \alpha_m$ such that $\mathrm{alph}(y_i) = \mathrm{alph}(z)$ for $i = 1, \cdots, m$ and $p(z) = \sum_{i=1}^{m} \alpha_i \cdot p(y_i)$.

d1: Let $z = x_0a_1x_1a_2x_2 \cdots x_{t-1}a_tx_t$ be in L, x_i in Σ^*, a_i in Σ, such that the letters a_i occur at most once in all words of L and each letter in the x_i's occurs at least twice in some word of L. Then for each such z F contains some word $y_0a_1y_1a_2y_2 \cdots y_{t-1}a_ty_t$.

d2: For each pair of letters a, b in Σ, $a \neq b$, such that a occurs at least twice in a word of L and $\{a, b\} \subseteq \mathrm{alph}(x)$ for some x in L, F contains five words aaby, aayb, abay, abya, ayab for some y in Σ^*.

To show that F is a test set for L we carry over the case-analysis in the proof of Theorem 9. Case 2 can obviously be covered by properties d1, and d2. In Case 1 we show first that $h_1(a)$, $h_2(a)$, $h_1(by)$, $h_2(by)$ commute as before. Now we represent $h_1(b)$, $h_1(y)$, $h_2(b)$, $h_2(y)$ in terms of the primitive root u of $h_1(a)$. Inserting these in $h_1(aayb) = h_2(aayb)$ we derive easily that $h_1(a)$, $h_1(b)$, and $h_2(b)$ commute.

Since there are less than 2^n non-empty sub-alphabets of Σ we need less than $2^n \cdot n$ words in F to satisfy c1. For d1 and d2 we need less than $2^n \cdot n!$ and $5n^2$ words, respectively. Thus we can find an F with $\#F < 2^n(n! + n) + 5n^2$ satisfying properties c1, d1 and d2. ∎

References

[1] Albert, J., Culik II, K., and Karhümaki, J. Test sets for context-free languages and algebraic systems of equations over a free monoid. Techn. Report Nr. 104, Inst. f. Angew. Inform. u. Form. Beschr. verf., Univ. Karlsruhe, June 1981.

[2] Book, R.V., and Brandenburg, F.-J. Equality sets and complexity classes, SIAM Journal of Computing 9 (1980), 729-743.

[3] Culik II, K. Some decidability results about regular and pushdown translations, Information Processing Letters 8 (1979), 5-8.

[4] Culik II, K. Homomorphisms: decidability, equality and test sets, In Formal Language Theory: Perspectives and Open Problems (ed., R.V. Book), Academic Press (1980), New York, 167-194.

[5] Culik II, K., and Diamond, N.D. A homomorphic characteriztion of time and space complexity classes of languages, International Journal of Computer Mathematics 8A (1980), 207-222.

[6] Culik II, K., and Karhümaki, J. (1981), Systems of equations over a free monoid and Ehrenfeucht Conjecture, Discrete Mathematics (1982), to appear.

[7] Culik II, K., and Salomaa, A. On the decidability of homomorphism equivalence for languages. Journal of Computer and System Sciences 17

(1978), 163-175.

[8] Culik II, K., and Salomaa, A. Test sets and checking words for homomorphism equivalence, Journal of Computer and System Sciences 20 (1980), 379-395.

[9] Ehrenfeucht, A., and Rozenberg, G. Elementary homomorphisms and a solution to the DOL sequence equivalence problem, Theoretical Computer Science 7 (1978), 169-183.

[10] Ginsburg, S. Algebraic and Automata-Theoretic Properties of Formal Languages. North-Holland Publishing Co., Amsterdam, 1975.

[11] Harrison, M.A. Introduction to Formal Language Theory. Addison-Wesley, Reading, Mass., 1978.

[12] Hopcroft, J.E., and Ullman, J.D. Formal Languages and their Relation to Automata, Second Edition. Addison-Wesley, Reading, Mass., 1980.

[13] Makanin, G.S. The problem of solvability of equations in a free semigroup (in Russian), Matematiceskij Sbornik 103 (145), (1977), 148-236.

[14] Salomaa, A. Formal Languages. Academic Press, New York, 1973.

Aug. 19, 1982
UCS/M/DW.4/DW9.1/dm