TEST SETS FOR HOMOMORPHISM EQUIVALENCE ON CONTEXT FREE LANGUAGES*

by

J. Albert[+§]

and

K. Culik II[†]

Research Report CS-79-39

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

* This research was supported by the National Sciences and Engineering Council of Canada, under Grant No. A7403.

+ Institut für Angewandte Informatik und Formale Beschreibungsverfahren
Universität Karlsruhe
Karlsruhe, West Germany

§ This paper was written during the first author's visit at the University of Waterloo.

† Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

A B S T R A C T

We show that for every context free language L over some alphabet Σ there effectively exists a test set F, that is a finite subset of L such that, for any pair (g,h) of homomorphisms on Σ^*, $g(x) = h(x)$ for each x in F implies $g(x) = h(x)$ for all x in L.

This result is then extended from homomorphisms to generalized sequential machine mappings defined by machines with uniformly bounded number of states.

1. Introduction

Problems concerning homomorphism equivalence have been intensively studied recently. Specifically, the homomorphic equivalence problem for a language family L is the following: Given a language L in L and two homomorphisms g and h determine whether g and h are equivalent on L, i.e. whether or not $g(w) = h(w)$ holds for all words w in L. It has been shown in Culik and Salomaa (1978) that there exists a uniform algorithm answering this question for any context free language L. In Culik and Richier (1979) the problem has been shown decidable also for ETOL languages over two-letter alphabets. It was conjectured in Culik and Salomaa (1978) that the problem is decidable for indexed languages, however at the present time it is open even for DOL languages (over at least three-letter alphabets). Actually, the homomorphic equivalence problem for DOL languages can be easily shown to be equivalent to the HDOL sequence equivalence problem, a well-known open problem. The homomorphic equivalence problem for (deterministic) context sensitive languages has been shown undecidable in Culik and Salomaa (1978). The decidability of homomorphic equivalence has many applications, the most important is probably in the proof of the DOL equivalence problem, Culik (1977), Culik and Fris (1977), for others see Culik (1979).

Older than the above results but closely related is the following "Ehrenfeucht's conjecture": Every language L has a finite

subset F such that , for any pair of homomorphisms (g,h) , g and h are equivalent on L iff they are equivalent on F . Such a finite set was called test set in Culik and Salomaa (1979) where it has been shown that the conjecture holds true for languages over a two-letter alphabet. It is also clear from the arguments in Culik and Salomaa (1978) that the conjecture holds for regular sets over any alphabet, and that in this case a finite test set can be effectively constructed. On the other hand it follows from the undecidability result mentioned above that for context sensitive languages finite test sets cannot exist effectively since that would, clearly, imply the decidability of homomorphic equivalence for this family.

Our main result (Theorem 1) is that a finite test set exists, and effectively so, for any context free language (given by a context free grammar). This result clearly implies the main result of Culik and Salomaa (1978), Theorem 4.1, namely the decidability of homomorphic equivalence for context free languages. Our stronger result does not follow from the proof of Theorem 4.1 in Culik and Salomaa (1978), nevertheless we use a similar basic technique ("generalized pumping").

We actually prove a somewhat stronger result, namely, that given a context free grammar G with n nonterminals and maximum m letters at the right side of productions, the set of all words of L of the length at most m^{3n+1} form a test set which does not otherwise depend on G .

We conjecture that finite test sets effectively exist even for all indexed languages, however it follows from the above discussion that to show this even for DOL languages - a very special case of indexed languages - seems to be very hard.

In the last section we extend our results from homomorphisms to deterministic generalized sequential machines (with accepting states) with uniformly bounded number of states.

2. Preliminaries

We study homomorphisms over free monoid Σ^* generated by finite set (alphabet) Σ. The unit of Σ^* (the empty word) is denoted by ε. The length of w in Σ^* is denoted by $|w|$, the cardinality of a set S by card S. For the other elementary notions of formal language theory we refer the reader to Harrison (1978), Hopcroft and Ullman (1969) or Salomaa (1973).

3. Finite Test Sets For Context Free Languages

We will show that if arbitrary two homomorphisms agree on all "short" strings of a context free language (CFL) they must agree on the whole language. The size of the strings which have to be considered will be shown to be independent on the homomorphisms. The proof will be based on "generalized pumping". It is of interest that it is not sufficient to consider all strings derived with "one loop" as it is shown by the following example.

Consider the context free grammar (CFG) given by productions $S \to aSb \mid c$, i.e. $L(G) = \{a^n cb^n \mid n \geq 0\}$, and homomorphisms g, h given by

$$g(a) = 0 \qquad h(a) = 01$$

$$g(b) = 100 \qquad h(b) = 00$$

$$g(c) = \varepsilon \qquad h(c) = \varepsilon$$

Here, we have $g(c) = h(c) = \varepsilon$, $g(acb) = h(acb) = 0100$, however $g(a^2cb^2) \neq h(a^2cb^2)$.

We start with a simple lemma which modifies a well-known result, see e.g. Harrison (1978), Theorem 1.3.2.

Lemma 1: For some alphabet Σ let $u \in \Sigma^+$, $v,w,x \in \Sigma^*$ such that $uvw = vx$. Then there exist $p \in \Sigma^*$, $p' \in \Sigma^+$, $i \geq 1$, $j \geq 0$ such that $u = (pp')^i$ and $v = (pp')^j p$. Furthermore, p, p', i, j can be uniquely determined by choosing $|pp'|$ minimal.

Proof: The equation $uvw = vx$ implies that there exist $y,z \in \Sigma^*$ such that $|y| = |u| \neq 0$ and $x = yz$. It follows $uv = vy$ and by Harrison (1978), Theorem 1.3.2 there exist $q, q' \in \Sigma^*$, $k \geq 0$ such that $u = qq'$ and $v = (qq')^k q$.

We can always assume $q' \neq \varepsilon$, because in the case $q' = \varepsilon$ we have $u = q \neq \varepsilon$, $v = q^{k+1}$ and by defining $r = \varepsilon$, $r' = q$ we get the desired representation;

$$u = rr' \quad , \quad v = (rr')^{k+1} r \ .$$

From $u = qq'$, $v = (qq')^k q$, $q' \neq \varepsilon$, $k \geq 0$ it is clear now that we can uniquely determine $p \in \Sigma^*$, $p' \in \Sigma^+$, $i \geq 1$, $j \geq 0$ such that $|pp'|$ is minimal and $u = (pp')^i$, $v = (pp')^j p$. □

The next lemma is crucial for the proof of our main theorem and might also have applications in the study of systems of equations over free monoids.

Definition: Let Σ be an alphabet and $\alpha, \beta, \gamma, \bar{\alpha}, \bar{\beta}, \bar{\gamma} \in \Sigma^*$. The set of pairs $M = \{(\varepsilon,\varepsilon), (\alpha,\bar{\alpha}), (\beta,\bar{\beta}), (\gamma,\bar{\gamma}), (\alpha\beta,\bar{\beta}\bar{\alpha}), (\alpha\gamma,\bar{\gamma}\bar{\alpha}), (\beta\gamma,\bar{\gamma}\bar{\beta})\}$ is then called an initial loop set.

Lemma 2: Let M be an initial loop set as above and $u,w,y \in \Sigma^*$. If for any two homomorphisms $g,h : \Sigma^* \to \Delta^*$

(1) $$g(uvwxy) = h(uvwxy)$$

holds for all $(v,x) \in M$ then (1) also holds for $(v,x) = (\alpha\beta\gamma,\bar{\gamma}\bar{\beta}\bar{\alpha})$, i.e. $g(u\alpha\beta\gamma w\bar{\gamma}\bar{\beta}\bar{\alpha}y) = h(u\alpha\beta\gamma w\bar{\gamma}\bar{\beta}\bar{\alpha}y)$.

Proof: For notational convenience let $\eta_1 := g(\eta)$, $\eta_2 := h(\eta)$ for all $\eta \in \Sigma^*$. Thus we have

(2) $$u_1v_1w_1x_1y_1 = u_2v_2w_2x_2y_2 \quad \text{for all} \quad (v,x) \in M .$$

Since u_1 must be a prefix of u_2 or vice versa, and y_1 must be a postfix of y_2 or vice versa, there are $\rho,\sigma \in \Delta^*$ such that one of the following four cases occurs:

Case 1. $u_1 = u_2\rho$, $y_1 = \sigma y_2$; Case 2. $u_1 = u_2\rho$, $y_2 = \sigma y_1$

Case 3. $u_2 = u_1\rho$, $y_2 = \sigma y_1$; Case 4. $u_2 = u_1\rho$, $y_1 = \sigma y_2$

We will only consider cases 1. and 2.. Obviously, 1. and 3., 2. and 4. are symmetrical.

Case 1. Since $u_1w_1y_1 = u_2w_2y_2$ and $u_1 = u_2\rho$, $y_1 = \sigma y_2$ we get $\rho w_1\sigma = w_2$ and we can write (2) as $u_2\rho v_1w_1x_1\sigma y_2 = u_2v_2\rho w_1\sigma x_2y_2$ and thus we have $\rho v_1w_1x_1\sigma = v_2\rho w_1\sigma x_2$ for all $(v,x) \in M$. If $\rho \neq \varepsilon$, then by Lemma 1, there exist $p \in \Sigma^*$, $p' \in \Sigma^+$ such that $\rho = (pp')^{i_1}p$ and for each $(v,x) \in M$ there is a number $i(v) \geq 0$ such that $v_2 = (pp')^{i(v)}$.

By symmetric application of Lemma 1 to the postfix of $\rho v_1w_1x_1\sigma = v_2\rho w_1\sigma x_2$, $\sigma \neq \varepsilon$ implies $\sigma = (qq')^{j_1}q$ for some $q \in \Sigma^*$, $q' \in \Sigma^+$ and for each $(v,x) \in M$ there is a $j(x) \geq 0$ such that $x_2 = (q'q)^{j(x)}$.

For each $(v,x) \in M$ we define

$$\tilde{v}_2 := \begin{cases} v_2 & \text{if } \rho = \varepsilon \\ (p'p)^{i(v)} & \text{if } \rho \neq \varepsilon \end{cases} \quad \text{and}$$

$$\tilde{x}_2 := \begin{cases} x_2 & \text{if } \sigma = \varepsilon \\ (qq')^{j(x)} & \text{if } \sigma \neq \varepsilon \end{cases} .$$

We can now reformulate the goal of this first section as follows: If $v_1 w_1 x_1 = \tilde{v}_2 w_1 \tilde{x}_2$ for all $(v,x) \in M$, then also $\alpha_1\beta_1\gamma_1 w_1 \bar{\gamma}_1 \bar{\beta}_1 \bar{\alpha}_1 = \tilde{\alpha}_2\tilde{\beta}_2\tilde{\gamma}_2 w_1 \tilde{\bar{\gamma}}_2 \tilde{\bar{\beta}}_2 \tilde{\bar{\alpha}}_2$.

Subcase 1.1. Let $|\alpha_1| = |\alpha_2|$. Then clearly $|\alpha_1| = |\tilde{\alpha}_2|$ and because of $\alpha_1 w_1 \bar{\alpha}_1 = \tilde{\alpha}_2 w_1 \tilde{\bar{\alpha}}_2$ we have

$$\alpha_1 = \tilde{\alpha}_2 \quad , \quad \bar{\alpha}_1 = \tilde{\bar{\alpha}}_2 \quad . \tag{3}$$

Furthermore,

$$\beta_1\gamma_1 w_1 \bar{\gamma}_1 \bar{\beta}_1 = \tilde{\beta}_2\tilde{\gamma}_2 w_1 \tilde{\bar{\gamma}}_2 \tilde{\bar{\beta}}_2 \tag{4}$$

by assumption. Combining (3) and (4) we get the desired equation

$$\alpha_1\beta_1\gamma_1 w_1 \bar{\gamma}_1 \bar{\beta}_1 \bar{\alpha}_1 = \tilde{\alpha}_2\tilde{\beta}_2\tilde{\gamma}_2 w_1 \tilde{\bar{\gamma}}_2 \tilde{\bar{\beta}}_2 \tilde{\bar{\alpha}}_2 .$$

Subcase 1.2. Let $|\alpha_1| > |\alpha_2|$. Then there is a $\mu \in \Delta^+$ such that $\alpha_1 = \tilde{\alpha}_2\mu$ and from (2) we have $\alpha_1 v_1 w_1 x_1 \bar{\alpha}_1 = \tilde{\alpha}_2 \tilde{v}_2 w_1 \tilde{x}_2 \tilde{\bar{\alpha}}_2$ and $v_1 w_1 x_1 = \tilde{v}_2 w_1 \tilde{x}_2$ for $(v,x) \in \{(\varepsilon,\varepsilon) , (\beta,\bar{\beta}), (\gamma,\bar{\gamma})\}$.

Thus,

$$(5) \qquad \mu v_1 w_1 x_1 \bar{\alpha}_1 = \tilde{v}_2 w_1 \tilde{x}_2 \tilde{\bar{\alpha}}_2 = v_1 w_1 x_1 \tilde{\bar{\alpha}}_2$$

for $(v,x) \in \{(\varepsilon,\varepsilon)\ ,\ (\beta,\bar{\beta})\ ,\ (\gamma,\bar{\gamma})\}$. In more detail we will consider now

$$(6) \qquad \mu\beta_1 w_1 \bar{\beta}_1 \bar{\alpha}_1 = \beta_1 w_1 \bar{\beta}_1 \tilde{\bar{\alpha}}_2 \quad .$$

Again, by Lemma 1 we conclude, that there are $r \in \Delta^*$, $r' \in \Delta^+$, $k_1 \geq 1$, $i(\beta) \geq 0$ such that $\mu = (rr')^{k_1}$, $\beta_1 = (rr')^{i(\beta)} r$ where r,r' , k_1 , $i(\beta)$ can be uniquely determined. Hence, the equation (6) is reduced to

$$(r'r)^{k_1} w_1 \bar{\beta}_1 \bar{\alpha}_1 = w_1 \bar{\beta}_1 \tilde{\bar{\alpha}}_2 \quad .$$

Thus, for w_1 there exist $s \in \Delta^*$, $s' \in \Delta^+$, $i(w) \geq 0$ such that $w_1 = (ss')^{i(w)} s$ and by choosing $|ss'|$ minimal $ss' = r'r$.

Repeating the same conclusion for $\bar{\beta}_1$, we get $\bar{\beta}_1 = (tt')^{i(\bar{\beta})} t$ and $tt' = s's$ if $|tt'|$ was chosen minimal.

Finally, (6) is reduced to $(t't)^{k_1} \bar{\alpha}_1 = \tilde{\bar{\alpha}}_2$.

In $\mu w_1 \bar{\alpha}_1 = w_1 \tilde{\bar{\alpha}}_2$, which is equation (5) for $(\varepsilon,\varepsilon)$, we insert now the representations of μ, w_1 :

$$(rr')^{k_1} (ss')^{i(w)} s \bar{\alpha}_1 = (ss')^{i(w)} s \tilde{\bar{\alpha}}_2 \quad .$$

From the uniqueness of r,r' , s,s' we derive $rr' = ss' = r'r$ and $(s's)^{k_1} \bar{\alpha}_1 = \tilde{\bar{\alpha}}_2 = (t't)^{k_1} \bar{\alpha}_1$. Thus, $t't = s's = tt'$.

From $rr' = r'r$, $tt' = t't$, $|rr'|$ and $|tt'|$ being minimal, we conclude that there are $\vartheta,\eta \in \Delta^+$ such that $r = \vartheta^d$, $r' = \vartheta^{d'}$, $d + d' = 1$, $t = \eta^e$, $t' = \eta^{e'}$, $e + e' = 1$, c.f. (Harrison 1978, Corollary on pg. 9). Furthermore, $ss' = \vartheta$ and $s's = \eta$, $\mu = \vartheta^{k_1}$, $\beta_1 = \vartheta^{i(\beta)+d}$, $w_1 = \vartheta^{i(w)}s = s\eta^{i(w)}$, $\bar{\beta}_1 = \eta^{i(\bar{\beta})+e}$. In analogy to this, we derive from $\mu\gamma_1 w_1 \bar{\gamma}_1 \bar{\alpha}_1 = \gamma_1 w_1 \bar{\gamma}_1 \tilde{\bar{\alpha}}_2$, $\mu = \varphi^{m_1}$, $\gamma_1 = \varphi^{m(\gamma)+\ell}$, $w_1 = \varphi^{m(w)}\bar{s} = \bar{s}\psi^{m(w)}$, $\bar{\gamma}_1 = \psi^{m(\bar{\gamma})+n}$. We can assume minimality for $|\varphi|$, $|\psi|$ and conclude $\varphi = \vartheta$, $m_1 = k_1$, $m(w) = i(w)$, $\bar{s} = s$, $\psi = \eta$.

The equation $\mu\beta_1\gamma_1 w_1 \bar{\gamma}_1 \bar{\beta}_1 \bar{\alpha}_1 = \beta_1\gamma_1 w_1 \bar{\gamma}_1 \bar{\beta}_1 \tilde{\bar{\alpha}}_2$ can now be proven easily just by inserting all the representations computed above and using $\vartheta = ss'$, $\eta = s's$. This completes Subcase 1.2 and the case $|\alpha_1| < |\alpha_2|$ can obviously be treated in the very same manner.

Case 2: We have $u_1 = u_2\rho$, $y_2 = \sigma y_1$. From $u_1 w_1 y_1 = u_2 w_2 y_2$ we obtain now $\rho w_1 = w_2 \sigma$ and we will treat first the case where w_1, w_2 are not overlapping.

Subcase 2.1: Let $\tau \in \Delta^*$ such that $\rho = w_2\tau$ and $\sigma = \tau w_1$. From (2) we get

$$w_2 \tau v_1 w_1 x_1 = v_2 w_2 x_2 \tau w_1 \tag{7}$$

for all $(v,x) \in M$.

Just as in Case 1 we can use Lemma 1 to conclude: There exist $p,q \in \Delta^*$, $p',q' \in \Delta^+$ such that, if $w_2 \neq \varepsilon$, $w_2 = (pp')^{k_2}p$,

$v_2 = (pp')^{i(v)}$ for all $(v,x) \in M$ and if $w_1 \neq \varepsilon$, $w_1 = (qq')^{k_1}q$, $x_1 = (q'q)^{j(x)}$ for all $(v,x) \in M$. For $(v,x) \in \{(\alpha,\bar{\alpha}),(\beta,\bar{\beta}),(\gamma,\bar{\gamma})\}$ we define

$$\hat{v}_2 := \begin{cases} v_2 & \text{if } w_2 = \varepsilon \\ (p'p)^{i(v)} & \text{if } w_2 \neq \varepsilon \end{cases}$$

and

$$\hat{x}_1 := \begin{cases} x_1 & \text{if } w_1 = \varepsilon \\ (qq')^{j(x)} & \text{if } w_1 \neq \varepsilon \end{cases}$$

So, we can rewrite (7) as

(8) $$\tau v_1 \hat{x}_1 = \hat{v}_2 x_2 \tau$$

for all $(v,x) \in M$.

It follows immediately that there are words r,r', $\delta(v_1)$, $\delta(\hat{x}_1)$, $\delta(\hat{v}_2)$, $\delta(x_2) \in \Delta^*$ such that for $\tau \neq \varepsilon$, $\tau = (rr')^{j_1}r$, $v_1 = (r'r)^{i(x)}\delta(v_1)$, $\hat{x}_1 = \delta(\hat{x}_1)(r'r)^{i(x)}$, where $\delta(v_1)\delta(\hat{x}_1) = r'r$ and $\hat{v}_2 = (rr')^{j(v)}\delta(\hat{v}_2)$, $x_2 = \delta(x_2)(rr')^{j(x)}$ where $\delta(\hat{v}_2)\delta(x_2) = rr'$, for all $(v,x) \in M$.

To reduce this case to Case 1 we introduce the following notation:

For $(v,x) \in \{(\alpha,\bar{\alpha}), (\beta,\bar{\beta}), (\gamma,\bar{\gamma})\}$ let

$$\overset{o}{v}_2 := \begin{cases} \hat{v}_2 & \text{if } \tau = \varepsilon \\ (r'r)^{j(v)}\xi(\hat{v}_2) & \text{if } \tau \neq \varepsilon \end{cases}$$

and

$$\overset{o}{x}_2 := \begin{cases} x_2 & \text{if } \tau = \varepsilon \\ \xi(x_2)(r'r)^{j(x)} & \text{if } \tau \neq \varepsilon \end{cases}$$

such that

$$|\xi(\hat{v}_2)| = |\delta(\hat{v}_2)| \quad , \qquad |\xi(x_2)| = |\delta(x_2)|$$

and

$$\xi(\hat{v}_2)\xi(x_2) = r'r \quad .$$

The new formulation of Subcase 2.1 is now: If $v_1\hat{x}_1 = \overset{o}{v}_2\overset{o}{x}_2$ for all $(v,x) \in M$, then also $\alpha_1\beta_1\gamma_1\bar{\bar{\gamma}}_1\bar{\bar{\beta}}_1\bar{\bar{\alpha}}_1 = \overset{o}{\alpha}_2\overset{o}{\beta}_2\overset{o}{\gamma}_2\bar{\bar{\gamma}}_2\bar{\bar{\beta}}_2\bar{\bar{\alpha}}_2$.

After appropriate renaming, this is nothing but a special subcase of Case 1, where we had proven:

$$v_1w_1x_1 = \tilde{v}_2w_1\tilde{x}_2 \quad \text{for all} \quad (v,x) \in M$$

implies $$\alpha_1\beta_1\gamma_1w_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_1 = \tilde{\alpha}_2\tilde{\beta}_2\tilde{\gamma}_2w_1\tilde{\bar{\gamma}}_2\tilde{\bar{\beta}}_2\tilde{\bar{\alpha}}_2 \ .$$

We just have to carry over the scheme of Case 1 and additionally we have $w_1 = \varepsilon$.

Subcase 2.2: In $\rho w_1 = w_2\sigma$ now let w_1, w_2 be overlapping, i.e. there is a $\tau \in \Delta^*$ such that $w_1 = \tau\sigma$, $w_2 = \rho\tau$. And so from (2) we

derive

(9) $$\rho v_1 \tau \sigma x_1 = v_2 \rho \tau x_2 \sigma$$

for all $(v,x) \in M$.

Applying the same technique of splitting again, we conclude by Lemma 1:

if $\rho \neq \varepsilon$ then

$\rho = (pp')^{i_1} p$ for some $p \in \Delta^*$, $p' \in \Delta^+$ and

$v_2 = (pp')^{i(v)}$ for all $(v,x) \in M$ and if $\sigma \neq \varepsilon$

$\sigma = (qq')^{j_1} q$ for some $q \in \Delta^*$, $q' \in \Delta^+$ and

$x_1 = (q'q)^{j(x)}$ for all $(v,x) \in M$.

For all (v,x) in M we define

$$\tilde{x}_1 := \begin{cases} x_1 & \text{if } \sigma = \varepsilon \\ (qq')^{j(x)} & \text{if } \sigma \neq \varepsilon \end{cases}$$

$$\tilde{v}_2 := \begin{cases} v_2 & \text{if } \rho = \varepsilon \\ (p'p)^{i(v)} & \text{if } \rho \neq \varepsilon \end{cases}$$

Thus we get

$$v_1 \tau \tilde{x}_1 = \tilde{v}_2 \tau x_2$$

for all $(v,x) \in M$. Now $\alpha_1 \beta_1 \gamma_1 \tau \tilde{\bar{\gamma}}_1 \tilde{\bar{\beta}}_1 \tilde{\bar{\alpha}}_1 = \tilde{\alpha}_2 \tilde{\beta}_2 \tilde{\gamma}_2 \tau \bar{\gamma}_2 \bar{\beta}_2 \bar{\alpha}_2$ is to be derived.

With appropriate renaming this has been done already in Case 1. This completes the proof of Case 2 and the proof of Lemma 2. □

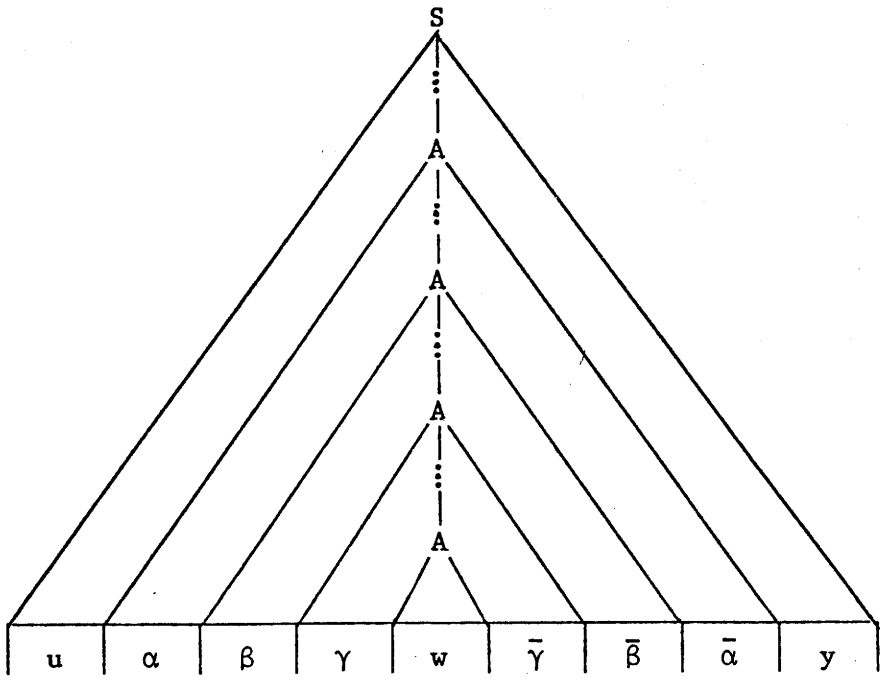
Now, we are ready to prove our main result.

Theorem 1: For every context free language $L \subseteq \Sigma^*$ (given by a CFG) there exists an effectively constructible finite subset $L' \subseteq L$, such that for any two homomorphisms g, h on Σ^*, $g(x) = h(x)$ for all $x \in L'$ implies $g(x) = h(x)$ for all $x \in L$.

Proof: Assume L is generated by some context free grammar $G = (N,\Sigma,P,S)$. Let D' be the set of all terminal derivation trees generated by G such that on each path from the root to a leaf at most three nodes are labelled by the same nonterminal from N. L' is now defined as the set of terminal words generated by D' (the yield of D'). Clearly, L' is finite and $L' \subseteq L$.

Assume that there is a string z in $L - L'$ such that $g(z) \neq h(z)$ and let z be a minimal string in the sense that for each z' in L where $|z'| < |z|$, we have $g(z') = h(z')$.

By the construction of L' there is a derivation tree for z of the form



for some nonterminal A, some words u, w, y and pairs of strings in Σ^* $(\alpha,\bar{\alpha})$, $(\beta,\bar{\beta})$, $(\gamma,\bar{\gamma})$ distinct from $(\varepsilon,\varepsilon)$.

Thus, by taking out any of these A-loops here, we get derivation trees generating words shorter than z. Now, clearly Lemma 2 applies and

$$g(u\alpha\beta\gamma w\bar{\gamma}\bar{\beta}\bar{\alpha}y) = h(u\alpha\beta\gamma w\bar{\gamma}\bar{\beta}\bar{\alpha}y) ,$$

completing the proof of Theorem 1. □

Definition: Let $L \subseteq \Sigma^*$. We say that F is a test set for L if $F \subseteq L$ and for any homomorphisms $g, h : \Sigma^* \to \Delta^*$, $g(x) = h(x)$ for all $x \in F$ implies $g(x) = h(x)$ for all $x \in L$.

Corollary 1: Let $G = (N,\Sigma,P,S)$ be a context free grammar with $n = \text{card } N$ and $m = \max (|X| : A \to X \in P)$. Let $F = \{w \in L(G) : |w| \leq m^{3n+1}\}$. Then F is a (finite) test set for $L(G)$.

Proof: Clear by the proof of Theorem 1.

Obviously, Corollary 1 implies the main result of (Culik and Salomaa, 1979, Theorem 4.1), namely, that given a CFL L and homomorphisms g, h, it is decidable whether $g(x) = h(x)$ for all x in L . □

4. Extension to gsm

Now, we will extend our result from homomorphisms to the mappings defined by deterministic generalized machines (gsm's) (with accepting states). We will construct a single test set for all deterministic gsm with bounded number of states.

Theorem 2: For every context free language $L \subseteq \Sigma^*$ (given by a CFG) and each natural number q there exists a finite subset $L' \subseteq L$ such that for any two functions $f_1, f_2 : \Sigma^* \to \Delta^*$ given by deterministic gsm's with at most q states, $f_1(x) = f_2(x)$ for all x in L' implies $f_1(x) = f_2(x)$ for all x in L.

Note: The above theorem clearly does not hold, if the numbers of states are arbitrary.

Proof of Theorem 2: Let $G = (N,\Sigma,P,S)$ be an ε-free CFG generating L, where $n = \text{card } N$, $d = \text{card } \Sigma$, $m = \max(|X| \mid A \to X \in P)$ and k is defined as $k = 2 \cdot q^4(n + d) + 1$. Let

$L' = \{w \in L(G) \mid |w| \leq m^{3k+1}\}$.

Consider any two deterministic gsm's $S_i = (Q_i,\Sigma,\Delta,\delta_i,q_i,F_i)$, $i = 1,2$ (c.f. Hopcroft and Ullman (1969) or Salomaa (1973)) such that $\text{card } Q_i \leq q$. Let D_i be the domain of S_i, $i = 1,2$. For $x \in \Sigma^*$ and $i = 1,2$, define

$$f_i(x) := \begin{cases} y & \text{where } x \in D_i,\ \delta_i(q_i,x) = (p_i,y) \\ & \text{for some } p_i \in F_i \\ & \\ & \text{undefined otherwise} \end{cases}$$

i.e. f_i is the mapping defined by machine S_i. Furthermore, let $M = L \cap (D_1 \cup D_2)$ and $M' = L' \cap (D_1 \cup D_2)$. Then, proving that $f_1(x) = f_2(x)$ for all $x \in M'$ implies $f_1(x) = f_2(x)$ for all $x \in M$ clearly establishes Theorem 2.

We proceed as follows: f_1, f_2 are decomposed into one injective, length preserving function g and two homomorphisms h_1, h_2, such that for all $x \in D_1 \cup D_2$: $f_1(x) = f_2(x)$ iff $h_1(g(x)) = h_2(g(x))$, and furthermore: $h_1(y) = h_2(y)$ for all $y \in g(M')$ implies $h_1(y) = h_2(y)$ for all $y \in g(M)$. This function g operates on strings $x = a_1a_2,\ldots,a_r \in D_1 \cup D_2$ as follows. For $i=1,2,\ldots,r$, the letter a_i of x is indexed by the states reached in S_1 and S_2 just after reading $a_1a_2,\ldots,a_{i-1}$; and the last letter is barred if x is accepted by exactly one of the gsm's S_1, S_2. In more detail, for $x = a_1a_2,\ldots,a_{r-1}a_r \in D_1 \cup D_2$ let $g(x) = a_1(m_1,n_1)a_2(m_2,n_2),\ldots,a_{r-1}(m_{r-1},n_{r-1})\tilde{a}_r(m_r,n_r)$ where $m_1 = q_1$, $n_1 = q_2$, $\delta_1(q_1,a_1,\ldots,a_{i-1}) = (m_i,y_1)$ for some $y_1 \in \Delta^*$, $\delta_2(q_2,a_1,\ldots,a_{i-1}) = (v_i,y_2)$ for some $y_2 \in \Delta^*$, and

$$\tilde{a}_r = \begin{cases} a_r & \text{if } x \in D_1 \cap D_2 \\ \bar{a}_r & \text{if } x \in (D_1 - D_2) \cup (D_2 - D_1) \,. \end{cases}$$

Clearly, g is length-preserving and injective and can be provided effectively by a deterministic gsm .

Since the family of context free languages is effectively closed under gsm-mappings, we can construct a context free grammar $G' = (N', \Sigma', P', S')$ such that $L(G') = g(M) = g(L(G) \cap (D_1 \cup D_2))$. Since the construction of G' is just a straightforward variant of the well-known construction with new nonterminals being triples from $Q_i \times N \times Q_i$, we omit the details for P' and consider only N' . It is obvious that the choice of

$$N' := \{(p,q,X,p',q') \,,\ (p,q,\bar{X},p',q') \mid X \in N \in \Sigma \,,$$
$$p,p' \in Q_1, q,q' \in Q_2\} \cup \{S'\}$$

is sufficient for our construction.

Since, card $N' \leq 2 \cdot q^4 \cdot (n + d) + 1 = k$, by Corollary 1 it holds for any two homomorphisms h_1, h_2 on Σ'^* that $h_1(y) = h_2(y)$ for all $y \in g(M')$ implies $h_1(y) = h_2(y)$ for all $y \in g(M)$. Specifically, for any $a \in \Sigma$, $p \in Q_1$, $q \in Q_2$ let

$h_1(a(p,q)) = y_1$, where $\delta_1(p,a) = (p',y_1)$ for some $p' \in Q_1$,

$h_1(\bar{a}(p,q)) = \#_1$, for some new symbol $\#_1$. Analogously, let

$h_2(a(p,q)) = y_2$, where $\delta_2(q,a) = (q',y_2)$ for some $q' \in Q_2$,

$h_2(\bar{a}(p,q)) = \#_2$, for some new symbol $\#_2 \neq \#_1$. Since $g : M \to L(G')$

is bijective and length-preserving, we conclude: $h_1(g(x)) = h_2(g(x))$ for all $x \in M'$ implies $h_1(g(x)) = h_2(g(x))$ for all $x \in M$, which proves Theorem 2, because of $h_1(g(w)) = f_1(x)$, $h_2(g(x)) = f_2(x)$ for all $x \in L \cap D_1 \cap D_2$ and $h_1(g(x)) \in \Delta^* \cdot \{\#_1\}$, $h_2(g(x)) \in \Delta^* \cdot \{\#_2\}$ for all $x \in L \cap ((D_1 - D_2) \cup (D_2 - D_1))$. □

Finally, we note that the proof of Theorem 2 suggests that Theorem 2 might be possible to extend to a larger family of languages L, e.g. even indexed languages, if the effective existence of finite test sets were shown for L and if L has some other properties like the family of CFL .

References

Culik, K. II (1977), On the decidability of the sequence equivalence problem for DOL-systems, Theoretical Computer Sci. 3, 75-84.

Culik, K. II (1979), Some decidability results about regular and push down translations, Information Processing Letters 8, 5-8.

Culik, K. II and Fris, J. (1977), The decidability of the equivalence problem for DOL systems, Inform. Control 35, 20-39.

Culik, K. II and Richier, J.L. (1979), Homomorphism equivalence on ETOL languages, Int. J. Computer Math. Section A, 7, 43-51.

Culik, K. II and Salomaa, A. (1978), On the decidability of homomorphism equivalence for languages, J. Computer Syst. Sci. 17, 163-175.

Culik, K. II and Salomaa, A. (1979), Test sets and checking words for homomorphism equivalence, submitted to J. Computer Syst. Sci; also Research Report CS-79-04, Department of Computer Science, University of Waterloo, Waterloo.

Harrison, M.A. (1978), "Introduction to Formal Language Theory," Addison-Wesley, Reading, Massachusetts.

Hopcroft, J.E. and Ullman, J.D. (1969), "Formal Languages and Their Relation to Automata," Addison-Wesley, Reading, Massachusetts.

Salomaa, A. (1973), "Formal Languages" Academic Press, New York.