Test Sets for Context Free Languages and Algebraic Systems of Equations over a Free Monoid†

J. Albert

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

K. Culik II

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

and

J. Karhumäki
Department of Mathematics
University of Turku
Turku, Finland

† This research was supported by the Natural Sciences and Engineering Council Canada, under Grant A7403.

No. of pages: 28

No. of tables: 0

No. of figures: 1

Proposed running head: "Test sets for context free languages" or "Test sets for CFL".

List of symbols used:

Symbol	Description
Σ	capital Greek "sigma"
Σ^*	capital Greek "sigma" with an asterisk at upper right corner
Σ^+	capital Greek "sigma" with a cross at upper right corner
Δ^*	capital Greek "delta" with an asterisk at upper right corner
L	upper case script "el"
$\in$	is a member of
$\square$	square = end of a proof
$\rightarrow$	symbol for a mapping
$\geq$	greater or equal than
$\leq$	smaller or equal than
$\neq$	not equal
$<$	smaller than
$>$	greater than
$\subseteq$	set inclusion
$\cap$	symbol for intersection
$\cup$	symbol for union
$\tilde{v}$, $\tilde{x}$	lower case Latin letters with tilde above
$\overline{X}$	upper case Latin letter with bar above
$\overline{a}$	lower case Latin letter with bar above

$\hat{v}$, $\hat{x}$	lower case Latin letters with hood above
xy	a pair of lower case Latin letters with hood above
α	lower case Greek "alpha"
β	lower case Greek "beta"
γ	lower case Greek "gamma"
η	lower case Greek "eta"
ρ	lower case Greek "rho"
σ	lower case Greek "sigma"
μ	lower case Greek "mu"
ζ	lower case Greek "theta"
φ	lower case Greek "phi"
ψ	lower case Greek "psi"
τ	lower case Greek "tau"
δ	lower case Greek "delta"
ξ	lower case Greek "Xi"
$\overline{\alpha}$, $\overline{\beta}$, $\overline{\gamma}$	lower case Greek letters with bars above
$\hat{\overline{\alpha}}$, $\hat{\overline{\beta}}$, $\hat{\overline{\gamma}}$	lower case Greek letters with bars and hoods above

Test Sets for Context Free Languages and Algebraic Systems of Equations over a Free Monoid

J. Albert
Universität Karlsruhe, Karlsruhe, W. Germany
K. Culik II
University of Waterloo
Waterloo, Ontario, Canada
and
J. Karhumäki
University of Turku, Turku, Finland

Department of Computer Science
Research Report CS-81-16
April 1981

Faculty of Mathematics

University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

Test Sets for Context Free Languages and Algebraic Systems of Equations over a Free Monoid†

J. Albert

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

K. Culik II

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

and

J. Karhumäki
Department of Mathematics
University of Turku
Turku, Finland

† This research was supported by the Natural Sciences and Engineering Council Canada, under Grant A7403.

No. of pages: 28

No. of tables: 0

No. of figures: 1

Proposed running head: "Test sets for context free languages" or "Test sets for CFL".

List of symbols used:

Symbol	Description
Σ	capital Greek "sigma"
Σ^*	capital Greek "sigma" with an asterisk at upper right corner
Σ^+	capital Greek "sigma" with a cross at upper right corner
Δ^*	capital Greek "delta" with an asterisk at upper right corner
L	upper case script "el"
$\in$	is a member of
$\square$	square = end of a proof
$\rightarrow$	symbol for a mapping
$\geq$	greater or equal than
$\leq$	smaller or equal than
$\neq$	not equal
$<$	smaller than
$>$	greater than
$\subseteq$	set inclusion
$\cap$	symbol for intersection
$\cup$	symbol for union
$\tilde{v}$, $\tilde{x}$	lower case Latin letters with tilde above
$\overline{X}$	upper case Latin letter with bar above
$\overline{a}$	lower case Latin letter with bar above

$\hat{v}$, $\hat{x}$	lower case Latin letters with hood above
xy	a pair of lower case Latin letters with hood above
α	lower case Greek "alpha"
β	lower case Greek "beta"
γ	lower case Greek "gamma"
η	lower case Greek "eta"
ρ	lower case Greek "rho"
σ	lower case Greek "sigma"
μ	lower case Greek "mu"
ζ	lower case Greek "theta"
φ	lower case Greek "phi"
ψ	lower case Greek "psi"
τ	lower case Greek "tau"
δ	lower case Greek "delta"
ξ	lower case Greek "Xi"
$\overline{\alpha}$, $\overline{\beta}$, $\overline{\gamma}$	lower case Greek letters with bars above
$\hat{\overline{\alpha}}$, $\hat{\overline{\beta}}$, $\hat{\overline{\gamma}}$	lower case Greek letters with bars and hoods above

Abstract. It is shown that for every context free language L there effectively exists a test set F, that is a finite subset F of L such that, for any pair (g, h) of morphisms, $g(x) = h(x)$ for all x in F implies $g(x) = h(x)$ for all x in L. This result was claimed earlier but a detailed correct proof is given here. Together with very recent results on systems of equations over a free monoid this result implies that every algebraic system of equations is equivalent to a finite subsystem.

1. Introduction

A number of results concerning the decidability of problems about morphisms have been obtained recently, for a survey of them see Culik (1980). Already in an early stage of this work A. Ehrenfeucht made the following intriguing conjecture: Every language L has a finite subset F, such that for any pair (g, h) of morphisms, $g(x) = h(x)$ for all x in L iff $g(x) = h(x)$ for all x in F. Such a finite set F has been called a test set in Culik and Salomaa (1980) where it has been shown that the conjecture holds true for languages over a two-letter alphabet. It is also clear from arguments in Culik and Salomaa (1978) that Ehrenfeucht's conjecture holds for regular sets over any alphabet, in which case a (finite) test set can even be effectively constructed. Effective existence of a test set for each language of family L clearly implies that morphism equivalence is decidable for family L, i.e. given a language L in L and two morphisms g, h it is decidable whether or not $g(x) = h(x)$ for each $x \in L$. Therefore test sets cannot effectively exist for context sensitive languages since morphism equivalence for them has been shown undecidable in Culik and Salomaa (1978).

The main purpose of this paper is to prove that a test set effectively exists for each context free language. This result was already claimed in Albert and Culik (1980), however R. Parchmann discovered an error in the proof of Lemma 2 of this paper and gave a counterexample (shown in Section 3) to this lemma.

In Section 3 we prove a weaker but still sufficient version of Lemma 2. The proof is quite lengthy despite an effort to make it as

succinct as possible. The next section gives the main result and its application to the testing of morphism equivalence. The new applications to systems of algebraic equations over a free monoid are discussed in Section 5.

2. Preliminaries

We need only very basic notions of formal language theory. To fix notation we specify the following, otherwise we refer the reader to Harrison (1978), Hopcroft and Ullman (1979) or Salomaa (1973).

We study morphisms of a free monoid Σ^* generated by a finite alphabet Σ. The unit of Σ^* is denoted by λ and $\Sigma^+ = \Sigma^* - \{\lambda\}$. The length of a word x in Σ^* is denoted by $|x|$. For two words x and y, xy^{-1} (resp. $y^{-1}x$) denotes the right (resp. left) difference of x by y. The notation $\mathrm{pref}_n(x)$ is used to denote the prefix of x of length n. By definition, if $|x| < n$ then $\mathrm{pref}_n(x) = x$. By pref(L) we mean the set of all prefixes of words in a language L. The corresponding notions for suffixes are obtained by replacing pref by suf.

We are almost all the time working with equations in a free monoid. The following basic facts are well known and used without any explicit reference, the reader may consult e.g. Harrison (1978). For each word x in Σ^+ there exists a unique word $\rho(x)$ such that $x \in \rho(x)^*$ and $\rho(x)$ cannot be written in the form $\rho(x) = y^n$ with $n \geq 2$. The word $\rho(x)$ is called the primitive root of x, and a word x is called primitive if $\rho(x) = x$. For two nonempty words x and y, $\rho(x) = \rho(y)$ if and only if $xy = yx$. For arbitrary x and y the identity $xy = yx$ is equivalent to the existence of a word p such that $x, y \in p^*$. We also recall the fact that if two words x^n and y^m have a common prefix of length $|x| + |y|$, then $\rho(x) = \rho(y)$. Finally, we state a simple lemma, the proof of which is straightforward.

Lemma 1. Let Σ be a finite alphabet and $u \in \Sigma^+$, $v,w,x \in \Sigma^*$. If $uvw = vx$, then there exist words $p \in \Sigma^*$, $p' \in \Sigma^+$ and integers $i \geq 1$, $j \geq 0$ such that $u = (pp')^i$ and $v = (pp')^j p$. Moreover, if pp' is chosen to be primitive, then p, p', i and j are unique. □

Next we state our crucial definition.

Definition. Let $L \subseteq \Sigma^*$. We say that a finite set F is a test set for L if $F \subseteq L$ and for any two morphisms $h,g : \Sigma^* \to \Delta^*$

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

Intuitively, the above means that to test whether two morphisms agree word by word on a language L it is enough to check whether they agree on a finite subset F of L.

Finally, we define the notion of the balance of a word with respect to two morphisms, cf. Culik (1980). Let h and $g : \Sigma^* \to \Delta^*$ be two morphisms and $w \in \Sigma^*$. The balance of w with respect to the pair (h, g), denoted by $\beta_{h,g}(w)$, is defined by

$$\beta_{h,g}(w) = |h(w)| - |g(w)| .$$

We write simply $\beta(w)$ if morphisms h,g are understood.

3. Pumping and Test Sets

In this section we show how certain pumping properties of languages are related to the existence of a test set. This is done by considering certain types of equations in a free monoid. We start with a simple example.

Example 1. Let Σ be a finite alphabet. For any words x_1, x_2, y_1, y_2, u_1, u_2, v_1 and v_2 the following holds true

$$\left\{ \begin{array}{l} x_1y_1 = x_2y_2 \\ x_1u_1y_1 = x_2u_2y_2 \\ x_1v_1y_1 = x_2v_2y_2 \end{array} \right\} \quad \text{implies} \quad x_1u_1v_1y_1 = x_2u_2v_2y_2 \ .$$

To see this, assume $x_1 = x_2w$ for some $w \in \Sigma^*$. Then $y_2 = wy_1$ and consequently the second equation yields $x_2wu_1y_1 = x_2u_2wy_1$, i.e. $wu_1 = u_2w$. Similarly we obtain $wv_1 = v_2w$. Hence, we conclude:

$$x_1u_1v_1y_1 = x_2wu_1v_1y_1 = x_2u_2wv_1y_1 = x_2u_2v_2wy_1 = x_2u_2v_2y_2 \ .$$

The above implies that any regular language, that is a language satisfying the "one place pumping property" has a test set, cf. Culik and Salomaa (1978). Indeed, if a language is given by a finite automaton, then its test set is obtained by taking all words yielding a computation where each state of the automaton is passed at most twice, i.e. by taking words with loop-free computations as well as words with at most single (but possibly nested) loop computations.

If the pumping occurs in two places, as is typical for context-free languages, then the situation is essentially more complicated. It is not only true that single loops are not enough but also that double loops

are not enough either. This is seen from the following example essentially due to R. Parchmann.

Example 2. Let $\Sigma = \{ a, b, d, \bar{a}, \bar{b} \}$ and h, g two morphisms defined by the table:

	a	b	d	$\bar{a}$	$\bar{b}$
h	bcbb	bcb	b	λ	b
g	λ	bcb	b	cbbb	b

It is striaghtforward to see that $h(xdy) = g(xdy)$ for all $(x, y) \in \{(\lambda, \lambda), (a, \bar{a}), (b, \bar{b}), (aa, \bar{a}\bar{a}), (bb, \bar{b}\bar{b}), (ab, \bar{b}\bar{a}), (ba, \bar{a}\bar{b})\}$, actually even for all $(x, y) \in \{(a^n, \bar{a}^n), (b^n, \bar{b}^n) \mid n \geq 1\}$. However, $h(abbd\bar{b}\bar{b}\bar{a}) \neq g(abbd\bar{b}\bar{b}\bar{a})$.

In the following we show that the above is the worst possible situation, that is when we pump in two places, then three loops are enough. To make this precise let $\Delta = \{A, B, C, D, \bar{A}, \bar{B}, \bar{C}, \bar{D}\}$ and define $Q \subseteq \Delta^*$ as $Q = \{(\lambda, \lambda), (X, X), (XY, \bar{Y}\bar{X}), (XYZ, \bar{Z}\bar{Y}\bar{X}) \mid X,Y,Z \in \{A, B, C, D\}$, $X \neq Y, X \neq Z, Y \neq Z\}$. Now, let Σ be an alphabet and $\alpha, \beta, \gamma, \delta, \bar{\alpha}, \bar{\beta}, \bar{\gamma}, \bar{\delta}$ words (not necessarily distinct) in Σ^* . Define a morphism $\mu : \Delta^* \to \Sigma^*$ by $\mu(A) = \alpha$, $\mu(B) = \beta$, $\mu(C) = \gamma$, $\mu(D) = \delta$, $\mu(\bar{A}) = \bar{\alpha}$, $\mu(\bar{B}) = \bar{\beta}$, $\mu(\bar{C}) = \bar{\gamma}$ and $\mu(\bar{D}) = \bar{\delta}$. We call the set

$$M = \mu(Q) = \{(\lambda, \lambda), (\alpha, \bar{\alpha}), \ldots, (\delta\gamma\beta, \bar{\beta}\bar{\gamma}\bar{\delta})\}$$

an initial loop set.

The reason for the rather complicated formal definition of an initial loop set is that we must include only nonrepetitive combinations from $\alpha, \beta, \gamma, \delta$, however the strings $\alpha, \beta, \gamma, \delta$ do not have to be distinct. In other words for any $(v, x) \in M$ v is obtained from $\alpha\beta\gamma\delta$ and x is obtained from $\bar{\delta}\bar{\gamma}\bar{\beta}\bar{\alpha}$ by erasing some of the pairs $(\alpha, \bar{\alpha})$, $(\beta, \bar{\beta})$, $(\gamma, \bar{\gamma})$, $(\delta, \bar{\delta})$ (and possibly by changing the order of words). This is essential, since Lemma 2 will be used to show that if two morphisms disagree on some long enough word they will disagree already on a shorter word, cf. the proof of Theorem 1.

Using this notation we state our basic lemma.

Lemma 2. Let M be an initial loop set, $u,w,y \in \Sigma^*$ and $h,g : \Sigma^* \to \Delta^*$ two morphisms. If $g(uvwxy) = h(uvwxy)$ holds true for all $(v,x) \in M$, then it also holds for $(v, x) = (\alpha\beta\gamma\delta, \bar{\delta}\bar{\gamma}\bar{\beta}\bar{\alpha})$ i.e.

$$g(u\alpha\beta\gamma\delta w\bar{\delta}\bar{\gamma}\bar{\beta}\bar{\alpha}y) = h(u\alpha\beta\gamma\delta w\bar{\delta}\bar{\gamma}\bar{\beta}\bar{\alpha}y) \quad . \tag{1}$$

The proof of the lemma is rather lengthy. So we divide it into several parts, some of them formulated as independent lemmas.

Lemma 3. Let $\omega \in \Delta^*$, $\eta_1 = g(\eta)$ and $\eta_2 = h(\eta)$ for each η in Σ^*. If $v_1\omega x_1 = v_2\omega x_2$ holds true for all $(v, x) \in M$, then also

$$\alpha_1\beta_1\gamma_1\delta_1\omega\bar{\delta}_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_1 = \alpha_2\beta_2\gamma_2\delta_2\omega\bar{\delta}_2\bar{\gamma}_2\bar{\beta}_2\bar{\alpha}_2 \quad . \tag{2}$$

Proof. We first observe that the case when any of the pairs $(\alpha_1, \bar{\alpha}_1)$, $(\beta_1, \bar{\beta}_1)$, $(\gamma_1, \bar{\gamma}_1)$ or $(\delta_1, \bar{\delta}_1)$ equals (λ, λ) is clear: then equation (2) is already among the assumptions, since for any $(v, x) \in M$ we have $|v_1 x_1| = |v_2 x_2|$.

Case I. $|\alpha_1| = |\alpha_2|$.

Our assumption $\alpha_1 \omega \bar{\alpha}_1 = \alpha_2 \omega \bar{\alpha}_2$ now implies the identities $\alpha_1 = \alpha_2$ and $\bar{\alpha}_1 = \bar{\alpha}_2$. Consequently, (2) follows from

$\beta_1 \gamma_1 \delta_1 \omega \bar{\delta}_1 \bar{\gamma}_1 \bar{\beta}_1 = \beta_2 \gamma_2 \delta_2 \omega \bar{\delta}_2 \bar{\gamma}_2 \bar{\beta}_2$.

Case II. $|\alpha_1| > |\alpha_2|$.

Let $\alpha_1 = \alpha_2 \mu$ for some $\mu \in \Delta^+$. Moreover, let

$$M_\alpha = \{(v, x) \in M \mid (\alpha v, x\bar{\alpha}) \in M\} .$$

Then from our assumption and the identity $\alpha_1 = \alpha_2 \mu$ we conclude

(3) $$\mu v_1 \omega x_1 \bar{\alpha}_1 = v_2 \omega x_2 \bar{\alpha}_2 \quad \text{for all} \quad (v, x) \in M_\alpha .$$

Now we apply Lemma 1 to the equation $\mu \omega \bar{\alpha}_1 = \omega \bar{\alpha}_2$. So there exist $p \in \Delta^*$, $p' \in \Delta^+$, $i \geq 1$ and $j \geq 0$ such that

$$\mu = (pp')^i \quad , \quad \omega = (pp')^j p \quad .$$

We assume that pp' is chosen primitive. Setting

$$\bar{\mu} = (p'p)^i$$

we see that $\mu \omega = \omega \bar{\mu}$ and $\bar{\alpha}_2 = \bar{\mu} \bar{\alpha}_1$. Remembering that $v_1 \omega x_1 = v_2 \omega x_2$ for all $(v, x) \in M_\alpha$, we now conclude from (3)

(4) $$\mu v_1 \omega x_1 = v_1 \omega x_1 \bar{\mu} \quad \text{for all} \quad (v, x) \in M_\alpha \ .$$

Claim. For each $(v, x) \in M_\alpha$ there exist $i(v) \geq 0$ and $i(x) \geq 0$ such that $v_1 = (pp')^{i(v)}$ and $x_1 = (p'p)^{i(x)}$.

To prove the claim we first observe, by symmetry, that it is enough to show the existence of an $i(v)$.

Let (v, x) be a fixed element in M_α . We apply Lemma 1 to the equation $\mu v_1 \omega x_1 = v_1 \omega x_1 \bar{\mu}$. So it follows that

$$\mu = (qq')^k, \quad v_1 \omega x_1 = (qq')^{k(v,x)} q, \quad \bar{\mu} = (q'q)^k$$

for some $k \geq 1$, $k(v, x) \geq 0$ and $q \in \Delta^*$, $q' \in \Delta^+$. Choosing qq' primitive we obtain $i = k$ and hence

$$qq' = pp' = \rho(\mu)$$

and

$$q'q = p'p = \rho(\bar{\mu}) \ .$$

(Note that the primitiveness of pp' implies the primitiveness of $p'p$). Our next aim is to show that $p = q$ and, consequently, $p' = q'$. Assume that this is not the case. Then e.g. $p = q\varphi$, for some $\varphi \neq \lambda$, implying that $q' = \varphi p'$. Thus, we have $\rho(\bar{\mu}) = \varphi(p'q) = (p'q)\varphi$, and since both φ and $p'q$ are nonempty $\rho(\varphi) = \rho(p'q)$. So $\rho(\bar{\mu}) \subset (\rho(\varphi))^2(\rho(\varphi))^*$, a contradiction.

In conclusion, we have proved so far that for each $(v, x) \in M_\alpha$ there exists a $k(v, x)$ such that

$$v_1 \omega x_1 = (pp')^{k(v,x)} p$$

and moreover

$$\omega = (pp')^j p, \quad \mu = (pp')^i, \quad \bar{\mu} = (p'p)^i .$$

In particular, ω is a suffix of $v_1 \omega x_1$ and hence we may define $\hat{x}_1$ by the condition

$$\omega x_1 = \hat{x}_1 \omega .$$

By the definition of $\hat{}$, it is clear that it behaves like a morphism, i.e. $\widehat{xx'} = \hat{x}\,\hat{x}'$ (whenever everything is defined). It is also clear that $v_1 \hat{x}_1 \in (pp')^+$ for each $(v, x) \in M_\alpha$.

Next we consider β- and γ-words. By the above periodicity and by the length argument we have

$$\beta_1 \hat{\bar{\beta}}_1 \gamma_1 \hat{\bar{\gamma}}_1 = \gamma_1 \hat{\bar{\gamma}}_1 \beta_1 \hat{\bar{\beta}}_1 = \beta_1 \gamma_1 \hat{\bar{\gamma}}_1 \hat{\bar{\beta}}_1 = \gamma_1 \beta_1 \hat{\bar{\beta}}_1 \hat{\bar{\gamma}}_1 .$$

Consequently,

$$\beta_1 \gamma_1 = \gamma_1 \beta_1 , \quad \hat{\bar{\beta}}_1 \hat{\bar{\gamma}}_1 = \hat{\bar{\gamma}}_1 \hat{\bar{\beta}}_1$$

and

$$\hat{\bar{\beta}}_1 \gamma_1 \hat{\bar{\gamma}}_1 = \gamma_1 \hat{\bar{\gamma}}_1 \hat{\bar{\beta}}_1, \quad \hat{\bar{\gamma}}_1 \beta_1 \hat{\bar{\beta}}_1 = \beta_1 \hat{\bar{\beta}}_1 \hat{\bar{\gamma}}_1 .$$

So it follows that $\beta_1, \gamma_1 \in s_1^*$ and $\hat{\bar{\beta}}_1, \hat{\bar{\gamma}}_1 \in s_2^*$ for some words s_1 and s_2 . Thus, we may use our last equalities to conclude that $\beta_1, \gamma_1, \hat{\bar{\beta}}_1, \hat{\bar{\gamma}}_1 \in s^*$ for some s . Since $\beta_1 \hat{\bar{\beta}}_1$ and $\gamma_1 \hat{\bar{\gamma}}_1$ are powers of a primitive word pp' we have

$$\beta_1, \gamma_1, \hat{\bar{\beta}}_1, \hat{\bar{\gamma}}_1 \in (pp')^* .$$

By symmetry, we also obtain

$$\delta_1, \hat{\bar{\delta}}_1 \in (pp')^* \ .$$

Finally, we are ready to conclude the equation (2). Indeed, we have

$$\begin{aligned}
\alpha_1\beta_1\gamma_1\delta_1\omega\bar{\delta}_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_1 &= \alpha_2\mu\beta_1\gamma_1\delta_1\hat{\bar{\delta}}_1\hat{\bar{\gamma}}_1\hat{\bar{\beta}}_1\omega\bar{\alpha}_1 \\
&= \alpha_2\beta_1\gamma_1\delta_1\hat{\bar{\delta}}_1\hat{\bar{\gamma}}_1\hat{\bar{\beta}}_1\mu\omega\bar{\alpha}_1 \\
&= \alpha_2\beta_1\gamma_1\delta_1\hat{\bar{\delta}}_1\hat{\bar{\gamma}}_1\hat{\bar{\beta}}_1\omega\mu\bar{\alpha}_1 \\
&= \alpha_2\beta_1\gamma_1\delta_1\omega\bar{\delta}_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_2 \\
&= \alpha_2\beta_2\gamma_2\delta_2\omega\bar{\delta}_2\bar{\gamma}_2\bar{\beta}_2\bar{\alpha}_2 \ .
\end{aligned}$$

Case III. $|\alpha_2| > |\alpha_1|$.

Clearly, this is symmetric to Case II.

So our proof for Lemma 3 is complete. □

Lemma 4. Let $\tau \in \Delta^*$, $\eta_1 = g(\eta)$ and $\eta_2 = h(\eta)$ for each η in Σ^* . If $\tau v_1 x_1 = x_2 v_2 \tau$ holds true for all $(v, x) \in M$, then also

(5) $$\tau\alpha_1\beta_1\gamma_1\delta_1\bar{\delta}_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_1 = \alpha_2\beta_2\gamma_2\delta_2\bar{\delta}_2\bar{\gamma}_2\bar{\beta}_2\bar{\alpha}_2\tau \ .$$

Proof. As in Lemma 3, we may assume that the words $\alpha_1\bar{\alpha}_1$, $\beta_1\bar{\beta}_1$, $\gamma_1\bar{\gamma}_1$ and $\delta_1\bar{\delta}_1$ are nonempty. We have two cases.

Case I. $\beta(\alpha) = 0$, i.e. the balance of the word α is zero.

Subcase (i). $N = \max\{|\beta_2\gamma_2\bar{\gamma}_2\bar{\beta}_2|, |\beta_2\delta_2\bar{\delta}_2\bar{\beta}_2|, |\gamma_2\delta_2\bar{\delta}_2\bar{\gamma}_2|\} \geq |\tau|$.

Since $\beta(\alpha) = 0$ and $\tau\alpha_1\bar{\alpha}_1 = \alpha_2\bar{\alpha}_2\tau$ we have

$$|\alpha_2^{-1}\tau\alpha_1| = |\tau| = |\bar{\alpha}_2\tau(\bar{\alpha}_1)^{-1}| \quad . \tag{6}$$

Now let e.g. $|\beta_2\delta_2\bar{\delta}_2\bar{\beta}_2| \geq |\tau|$. Then from the assumption $\tau\beta_1\delta_1\bar{\delta}_1\bar{\beta}_1 = \beta_2\delta_2\bar{\delta}_2\bar{\beta}_2\tau$ we obtain

$$\mathrm{pref}_{|\tau|}(\beta_2\delta_2\bar{\delta}_2\bar{\beta}_2) = \tau = \mathrm{suf}_{|\tau|}(\beta_1\delta_1\bar{\delta}_1\bar{\beta}_1) \quad .$$

Consequently, from the assumption $\tau\alpha_1\beta_1\delta_1\bar{\delta}_1\bar{\beta}_1\bar{\alpha}_1 = \alpha_2\beta_2\delta_2\bar{\delta}_2\bar{\beta}_2\bar{\alpha}_2\tau$ and (6) it follows that

$$\alpha_2^{-1}\tau\alpha_1 = \tau = \alpha_2\tau(\bar{\alpha}_1)^{-1} \quad .$$

So the desired equation (5) follows from $\tau\beta_1\gamma_1\delta_1\bar{\delta}_1\bar{\gamma}_1\bar{\beta}_1 = \beta_2\gamma_2\delta_2\bar{\delta}_2\bar{\gamma}_2\bar{\beta}_2\tau$.

Subcase (ii). $N < |\tau|$.

From our assumptions $\tau v_1 x_1 = v_2 x_2 \tau$ for $(v, x) \in \{(\beta, \bar{\beta}), (\gamma, \bar{\gamma}), (\beta\gamma, \bar{\gamma}\bar{\beta}), (\gamma\beta, \bar{\beta}\bar{\gamma})\}$ we conclude that

$$\beta_2\bar{\beta}_2\gamma_2\bar{\gamma}_2,\ \gamma_2\bar{\gamma}_2\beta_2\bar{\beta}_2,\ \beta_2\gamma_2\bar{\gamma}_2\bar{\beta}_2,\ \gamma_2\beta_2\bar{\beta}_2\bar{\gamma}_2 \in \mathrm{pref}(\tau) \quad .$$

Consequently,

$$\beta_2\bar{\beta}_2\gamma_2\bar{\gamma}_2 = \gamma_2\bar{\gamma}_2\beta_2\bar{\beta}_2 = \beta_2\gamma_2\bar{\gamma}_2\bar{\beta}_2 = \gamma_2\beta_2\bar{\beta}_2\bar{\gamma}_2$$

and so we derive, as in the end of the proof of case II in Lemma 3, that

$$\beta_2,\ \gamma_2,\ \bar{\beta}_2,\ \bar{\gamma}_2 \in q^*$$

for some primitive word q .

Using the symmetric reasoning for β- and δ-words we obtain that also

$$\delta_2, \overline{\delta}_2 \in q^* .$$

Now the identity $\tau\beta_1\overline{\beta}_1 = \beta_2\overline{\beta}_2\tau$ gives

(7) $$\tau \in (q'q")^*q'$$

for some words q' and $q"$ with $q = q'q"$. Consequently, $v_2x_2\tau \in (q'q")^*q'$ for all $(v, x) \in M_\alpha$, and hence our assumptions together with the primitiveness of $q'q"$ implies

(8) $$\beta_1, \gamma_1, \delta_1, \overline{\beta}_1, \overline{\gamma}_1, \overline{\delta}_1 \in (q"q')^* .$$

To complete the Case II (ii) we prove the following claim. Observe that the claim will be proved without assuming anything about the balance of the word α.

Claim. Assume that the pairs $(\beta, \overline{\beta})$, $(\gamma, \overline{\gamma})$ and $(\delta, \overline{\delta})$ satisfy

$$\beta_1, \gamma_1, \delta_1, \overline{\beta}_1, \overline{\gamma}_1, \overline{\delta}_1 \in (q"q')^*$$

and

$$\beta_2, \gamma_2, \delta_2, \overline{\beta}_2, \overline{\gamma}_2, \overline{\delta}_2 \in (q'q")^*$$

for some words q' and $q"$ with $q'q"$ primitive. Then the statement of Lemma 4 follows.

To prove the claim let

$$r = \alpha_2^{-1}\tau\alpha_1 = \bar{\alpha}_2\tau(\bar{\alpha}_1)^{-1} .$$

(So we assume that $|\tau\alpha_1| \geq |\alpha_2|$; the case $|\tau\alpha_1| \leq |\alpha_2|$ is similar.) From the identities $\tau\alpha_1\bar{\alpha}_1 = \alpha_2\bar{\alpha}_2\tau$ and $\tau\alpha_1\beta_1\bar{\beta}_1\bar{\alpha}_1 = \alpha_2\beta_2\bar{\beta}_2\bar{\alpha}_2\tau$ we obtain

$$r \in \mathrm{pref}(\bar{\alpha}_2\tau) \cap \mathrm{pref}(\beta_2\bar{\beta}_2\bar{\alpha}_2\tau) .$$

Hence, by the fact $\beta_2\bar{\beta}_2 \in (q'q")^+$,

$$r \in \mathrm{pref}(q'q")^* .$$

Let $r = q^n\bar{q}$ with $|\bar{q}| < |q|$ and $n \geq 0$. Now we use the identity $\tau\alpha_1\beta_1\gamma_1\bar{\gamma}_1\bar{\beta}_1\bar{\alpha}_1 = \alpha_2\beta_2\gamma_2\bar{\gamma}_2\bar{\beta}_2\bar{\alpha}_2\tau$. This yields

$$r\beta_1\gamma_1\bar{\gamma}_1\bar{\beta}_1 = \beta_2\gamma_2\bar{\gamma}_2\bar{\beta}_2 r \in \mathrm{pref}(q'q")^* .$$

Consequently,

$$\bar{q}q"q'q"q' \in \mathrm{pref}(q'q")^*$$

which, by the primitiviness of $q'q"$, is possible only if $\bar{q} = q'$.

$$r \in (q'q")^*q'$$

and so a straightforward calculation proves (5).

Case II. $\beta(\alpha) \neq 0$.

Now the assumption $\tau v_1x_1 = v_2x_2\tau$, for each $(v, x) \in M$, implies that

(9) $$\tau(v_1x_1)^n = (v_2x_2)^n\tau \qquad \text{for all } n \geq 0 .$$

Next we make use of the identities $\tau\alpha_1\bar{\alpha}_1 = \alpha_2\bar{\alpha}_2\tau$ and $\tau\alpha_1 v_1 x_1 \bar{\alpha}_1 = \alpha_2 v_2 x_2 \bar{\alpha}_2 \tau$ for $(v, x) \in M_\alpha$. Assuming that $|\tau\alpha_1| \geq |\alpha_2|$ (the other case is quite similar) we obtain

$$\alpha_2^{-1}\tau\alpha_1 = (\alpha_2 v_2 x_2)^{-1}\tau\alpha_1 v_1 x_1 = \bar{\alpha}_2\tau(\bar{\alpha}_1)^{-1} = v_2 x_2 \bar{\alpha}_2 \, (v_1 x_1 \bar{\alpha}_1)^{-1} \ .$$

This yields

$$\tau\alpha_1(v_1x_1)^n\bar{\alpha}_1 = \alpha_2(v_2x_2)^n\bar{\alpha}_2\tau \quad \text{for all} \quad n \geq 0 \quad \text{and} \quad (v, x) \in M_\alpha \ . \tag{10}$$

From (9) and (10) it follows that

$$\tau(v_1x_1)^n, \ \alpha_2^{-1}\tau\alpha_1(v_1x_1)^n \in \mathrm{pref}(v_2x_2)^* \quad \text{for all} \quad n \geq 0 \ . \tag{11}$$

Now observe that $|\tau| \neq |\alpha_2^{-1}\tau\alpha_1|$ since $\beta(\alpha) \neq 0$. In other words, one of the words $\tau^{-1}(\alpha_2^{-1}\tau\alpha_1)$ and $(\alpha_2^{-1}\tau\alpha_1)^{-1}\tau$ is defined and nonempty. Let this word be $p(\alpha)$. Then, by (11), $(v_1x_1)^n \in \mathrm{pref}(p(\alpha))^*$ for all $n \geq 0$. Consequently, $\rho(v_1x_1) = \rho(p(\alpha))$. This means that there exists a primitive word, say p , such that

$$v_1x_1 \in p^+ \quad \text{for each} \quad (v, x) \in M_\alpha \ .$$

Observe that p is independent of (v, x) .

Now let us consider γ- and β-words. We have

$$\gamma_1\bar{\gamma}_1, \ \beta_1\bar{\beta}_1, \ \gamma_1\beta_1\bar{\beta}_1\bar{\gamma}_1 \in p^+$$

From the primitiveness of p it immediately follows that

$$\gamma_1, \ \bar{\gamma}_1, \ \beta_1, \ \bar{\beta}_1 \in p^* \ .$$

Hence, by symmetry, also

$$\delta_1, \ \bar{\delta}_1 \in p^* \ .$$

We continue as in Case I (ii) to obtain equations (7) and the analogy of (8) for the words β_2, $\bar{\beta}_2$, γ_2, $\bar{\gamma}_2$, δ_2 and $\bar{\delta}_2$. So the claim becomes applicable, which completes the proof of Lemma 4.

□

Proof of Lemma 2. Let us recall our assumption:

$$(12) \qquad u_1v_1w_1x_1y_1 = u_2v_2w_2x_2y_2 \quad \text{for} \quad (v, x) \in M .$$

Our aim is to show that this equality holds also for $(\alpha\beta\gamma\delta, \bar{\delta}\bar{\gamma}\bar{\beta}\bar{\alpha})$. We have four different cases depending on the relative lengths of u_1, u_2, y_1 and y_2 .

Case I. $u_1 = u_2\rho$, $y_1 = \sigma y_2$ for some words ρ and σ .

Case II. $u_1 = u_2\rho$, $y_2 = \sigma y_1$ for some words ρ and σ .

Case III. $u_2 = u_1\rho$, $y_2 = \sigma y_1$ for some words ρ and σ .

Case IV. $u_2 = u_1\rho$, $y_1 = \sigma y_2$ for some words ρ and σ .

Clearly, by symmetry, it is enough to prove the lemma for Cases I and II.

Case I. The identities $u_1 = u_2\rho$ and $y_1 = \sigma y_2$ applied to $u_1w_1y_1 = u_2w_2y_2$ yields $\rho w_1\sigma = w_2$ and, consequently, (12) is equivalent to

$$(13) \qquad \rho v_1w_1x_1\sigma = v_2\rho w_1\sigma x_2 \quad \text{for} \quad (v, x) \in M .$$

Clearly, for each $(v, x) \in M$, there exists $\tilde{v}_2$ and $\tilde{x}_2$ such that

$$v_2\rho = \rho\tilde{v}_2 , \quad \sigma x_2 = \tilde{x}_2\sigma .$$

By the definition of the operation $\sim$, it is clear that $\sim$, when

defined, behaves like a morphism, i.e. $(\widetilde{xx'}) = \tilde{x}\,\tilde{x}'$. Hence, with this notation , (13) is equivalent to

$$v_1 w_1 x_1 = \tilde{v}_2 w_1 \tilde{x}_2 \quad \text{for} \quad (v, x) \in M \ .$$

The proof of Case I is now completed by Lemma 3. Observe that when proving Lemma 3 we have simplified notation: w_1 is replaced by ω and the waves over the symbols are omitted.

Case II. Now identities $u_1 = u_2\rho$ and $y_2 = \sigma y_1$ applied to $u_1 w_1 y_1 = u_2 w_2 y_2$ yields $\rho w_1 = w_2 \sigma$. We have two subcases.

Subcase (i). There exists a word $\tau \in \Delta^*$ such that $\rho = w_2\tau$ and $\sigma = \tau w_1$, i.e. w_1 and w_2 are not overlapping. With these notation (12) is equivalent to

(14) $$w_2 \tau v_1 w_1 x_1 = v_2 w_2 x_2 \tau w_1 \quad \text{for} \quad (v, x) \in M \ .$$

As in Case I we define words $\hat{v}_2$ and $\hat{x}_1$ such that

$$v_2 w_2 = w_2 \hat{v}_2 \ , \quad w_1 x_1 = \hat{x}_1 w_1 \quad \text{for} \quad (v, x) \in M \ .$$

So (14) is equivalent to

$$\tau v_1 \hat{x}_1 = \hat{v}_2 x_2 \tau \quad \text{for} \quad (v, x) \in M \ ,$$

i.e. we have (after a renaming) the situation of Lemma 4.

Subcase (ii). In $\rho w_1 = w_2\sigma$ w_1 and w_2 are overlapping, i.e. there is a word $\tau \in \Delta^*$ such that $w_1 = \tau\sigma$ and $w_2 = \rho\tau$. So (12) is equivalent to

(15) $$\rho v_1 \tau\sigma x_1 = v_2 \rho\tau x_2 \sigma \quad \text{for all } (v, x) \in M .$$

Again we define words $\tilde{x}_1$ and $\tilde{v}_1$ such that

$$\sigma x_1 = \tilde{x}_1 \sigma , \quad v_2\rho = \rho\tilde{v}_2 .$$

Consequently, (15) can be rewritten as

$$v_1 \tau \tilde{x}_1 = \tilde{v}_2 \tau x_2 \quad \text{for all } (v, x) \in M ,$$

i.e. we may use Lemma 3 (after a renaming) in this case.

So our proof for Lemma 2 is complete.

□

4. A Test Set for Context Free Languages

In this section we prove our main result

Theorem 1. For every context free language $L \subseteq \Sigma^*$ (given by a λ-free context-free grammar G) there effectively exists a test set F. Moreover, F can be chosen to be $\{w \in L \mid |w| \leq m^{4n+1}\}$, where n is the cardinality of the nonterminal alphabet of G and m is the length of the longest right hand side among the productions.

Proof. Assume that L is generated by a λ-free 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 four nodes are labelled by the same nonterminal. Let L' denote the set of terminal words generated by D (the yield of D). Clearly, L' is a finite subset of L.

We claim that L' is a test set for L. To show this let h and g be arbitrary two morphisms. Assume that there exists a word z in L-L' such that $h(z) \neq g(z)$. Moreover, let z be a minimal such word (with respect to (h, g)), i.e. whenever $h(z') \neq g(z')$ with $z' \in L\text{-}L'$, then $|z'| \geq |z|$. By the definition of L', there exists a derivation tree for z of the form shown in Figure 1, for some words u, w, y and some pairs $(\alpha, \overline{\alpha})$, $(\beta, \overline{\beta})$, $(\gamma, \overline{\gamma})$ and $(\delta, \overline{\delta})$ different from (λ, λ).

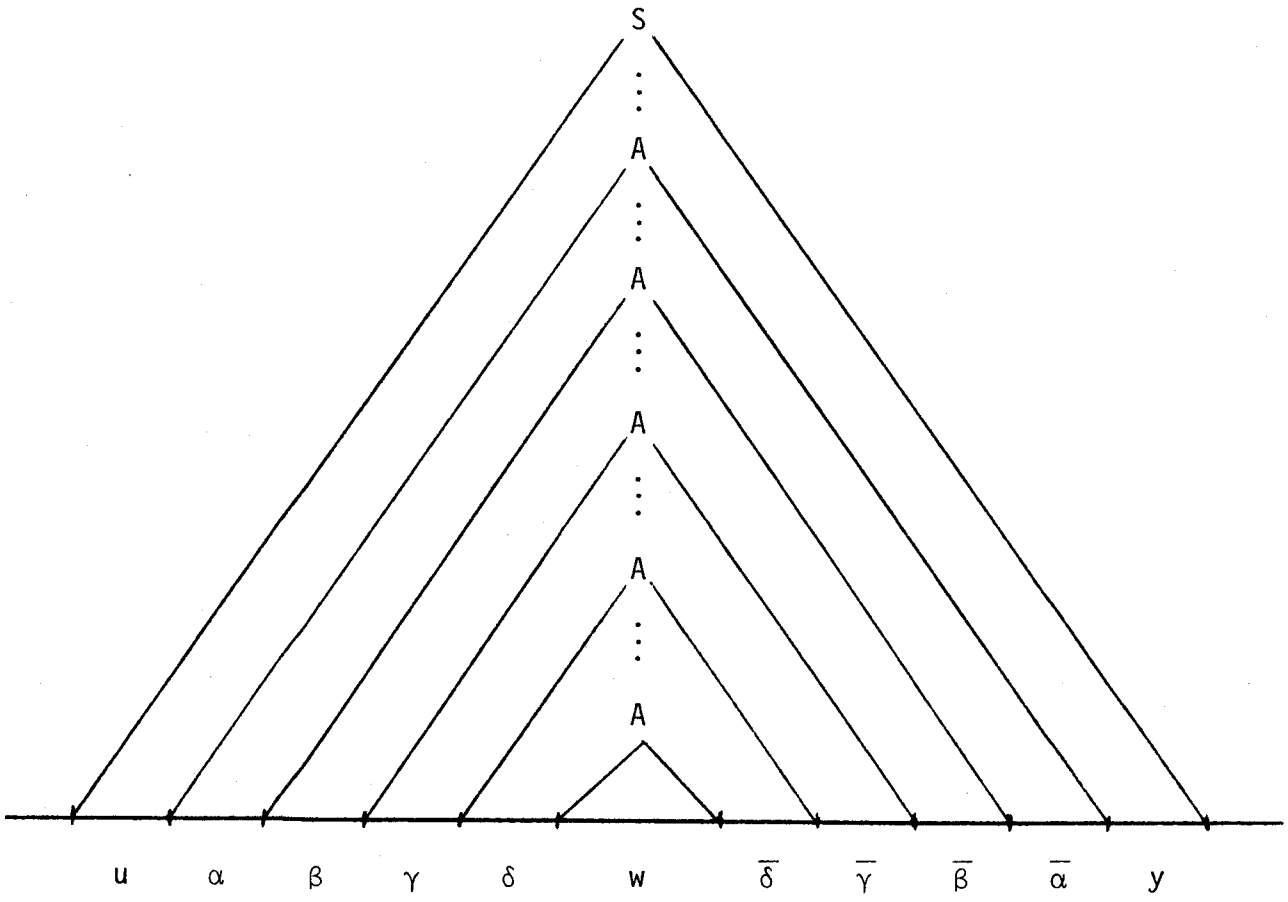

Figure 1

By Lemma 2 and by the relation $h(z) \neq g(z)$, we conclude that deleting some A-loops from the above derivation tree, i.e. erasing from z some of the pairs $(\alpha, \overline{\alpha})$, $(\beta, \overline{\beta})$, $(\gamma, \overline{\gamma})$ or $(\delta, \overline{\delta})$, we obtain a shorter word z_1 such that $h(z_1) \neq g(z_1)$. By the minimality of z, z_1 must be in L'. Consequently, L' tests whether h and g agree on L. Because this is true for any pair of morphisms, L' is really a test set for L.

The second sentence of the theorem is immediate.

□

Theorem 1 immediately implies the main result of Culik and Salomaa (1978), the decidability of morphism equivalence for the family of context free languages.

Corollary 1. Given a context free grammar $G = (N, T, P, S)$ and two morphisms $g,h : T^* \to \Delta^*$ it is decidable whether $g(w) = h(w)$ for all w in $L(G)$. □

We can easily extend the claims of both Theorem 1 and Corollary 1 to some non-context-free languages using the following

Lemma 5. If F is a test set for L and $F \subseteq L' \subseteq L$ then F is a test set for L' .

Proof. Obvious by definition of a test set.

□

Example. It has been shown in Culik and Salomaa (1978) that any two distinct strings of the language $L = \{a^n b^n \mid n \geq 1\}$ form its test set. Hence, by Lemma 5 the same statement holds true also for every subset of L of cardinality at least two.

5. Applications to Systems of Equations over a Free Monoid

In Culik and Karhumäki (1981) it has been shown that the effective existence of a test set for a language of certain type is equivalent to the effective existence of an equivalent finite subsystem for every system of equations of "the same type". Here we are concerned with context free (algebraic) languages and correspondingly with context free (algebraic) systems of equations, which we will now introduce formally.

A system of equations over Σ^* with unknowns N is a binary relation $S \subseteq (N \cup \Sigma)^* \times (N \cup \Sigma)^*$. A pair (u, v) in S represents the equation $u = v$. We say that system S is rational (regular) or algebraic (push-down) if N is finite and relation S is rational (regular) or algebraic (defined by a push-down transducer), respectively.

The following lemma follows immediately from Theorem 2.1 in Culik and Karhumäki (1981).

Lemma 6. If there effectively exists a test set for every context free language, then for every algebraic (push-down) system S of equations over Σ^* we can effectively construct a finite equivalent subsystem of S.

□

Thus our main result has an immediate corollary which extends a result obtained for rational systems in Culik and Karhumäki (1981).

Corollary 2. Given (effectively) an algebraic (push-down) system of equations S, we can effectively construct a finite equivalent subsystem of S.

Proof. By Lemma 6 and Theorem 1. □

Example. Let S be the system of equations over $\{a, b\}^*$ with unknowns $w, x, y,$ and z

$$w^n x^n = y^n z^n \quad \text{for all} \quad n \geq 1 .$$

Clearly, S is an algebraic system, and every solution of S is either of the form $w = y$, $x = z$ or of the form $w = p^r$, $x = p^q$, $y = p^s$, $z = p^t$ where $p \in \{a, b\}^*$; $r, q, s, t \geq 0$ and $r + q = s + t$. Here any two equations form an equivalent finite subsystems of S. This is just a translation into the terminology of equations of the Theorem 3.1 from Culik and Salomaa (1980) which states that every two distinct strings of the context free language $\{a^n b^n \mid n \geq 1\}$ constitute a test set.

The inclusion problem for systems of equations is to test for two given systems S_1 and S_2 whether every solution of S_1 is also a solution of S_2. Theorem 3.1 and Corollary 3.3 from Culik and Karhumäki (1981), together with Corollary 2 imply the decidability of the inclusion and the equivalence problem for algebraic systems of equations.

Corollary 3. The equivalence problem and the inclusion problem for algebraic systems of equations (with finite number of unknowns) over Σ^* are decidable.

□

Corollary 4. It is decidable whether an algebraic system of equations on a free monoid has a solution.

Proof. By Corollary 3 we can effectively construct an equivalent finite system of equations. Its solvability is decidable by Makanin (1977).

□

References

Albert, J. and Culik, K. II (1980), Tests sets for homomorphism equivalence on context free languages, *Inform. Contr.* 45, 273-284.

Culik, K. II (1980), Homomorphisms: Decidability, equality and test sets, in "Formal Language Theory, Perspectives and Open Problems", (R.V. Book, Ed.), pp. 167-194, Academic Press, New York.

Culik, K. II and Karhumäki, J. (1981), Systems of equations over a free monoid and Ehrenfeucht Conjecture, Research Report CS-81-15, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

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

Culik, K. II and Salomaa, A. (1980), Test sets and checking words for homomorphism equivalence, *J. Comput. System Sci.*, 20, 379-395.

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

Hopcroft, J. E. and Ullman, J. D. (1979), "Introduction to Automata Theory, Languages and Computation", Addison-Wesley, Reading, Mass.

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

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