INFORMATIQUE THÉORIQUE ET APPLICATIONS

ISMO HAKALA
JUHA KORTELAINEN

Polynomial size test sets for commutative languages

Informatique théorique et applications, tome 31, nº 3 (1997), p. 291-304
<http://www.numdam.org/item?id=ITA_1997__31_3_291_0>

© AFCET, 1997, tous droits réservés.

L'accès aux archives de la revue « Informatique théorique et applications » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 31, n° 3, 1997, pp. 291-304)

POLYNOMIAL SIZE TEST SETS FOR COMMUTATIVE LANGUAGES (*)

by Ismo HAKALA and Juha KORTELAINEN [1]

Communicated by J. BERSTEL

Abstract. – *It is proved that any commutative language over an alphabet of* n *symbols possesses a test set of size* $O(n^2)$. *If the Parikh-map of the language is a linear set, then the minimum size of the test set is* $O(n \log n)$. *A finite commutative language over an alphabet of* n *symbols such that the smallest test set for the language is of size* $\Omega(n^2)$ *is shown to exist.*

Résumé. – *On prouve que tout langage commutatif sur un alphabet à* n *lettre possède un ensemble test de taille* $O(n^2)$. *Si l'image de Parikh du langage est un ensemble linéaire, la taille minimale de l'ensemble test est* $O(n \log n)$. *On prouve l'existence d'un langage commutatif fini sur un alphabet à* n *lettres pour lequel la taille du plus petit ensemble test est* $\Omega(n^2)$.

0. INTRODUCTION

A subset T of a language L is defined to be a test set of L if for each pair of morphisms h and g the following hold:

$$\forall x \in T : h(x) = g(x) \Rightarrow \forall x \in L : h(x) = g(x).$$

The famous Ehrenfeucht Conjecture states that each language L has a finite test set. The conjecture was proved in [3]. Since then the effectiveness and sizes of the test sets of languages belonging to certain language families have been an important subject of consideration.

Test sets for context-free languages are studied in [1], [2], [8], [9] and [10]. The research culminates in [9] where, among other things, it is proved that (i) any context-free language L over an alphabet of n symbols possesses a test set of size $O(n^6)$; and (ii) there exist a finite context-free language over n letter alphabet such that its smallest test set is of size $\Omega(n^3)$. Test sets for context-sensitive languages with a strong pumping property are studied in [5] and [6].

(*) Received February 1997, accepted June 1997.

([1]) Department of Mathematical Sciences, University of Oulu, FIN-90570 Oulu, Finland.

0988-3754/97/03/$ 7.00/© AFCET-Gauthier-Villars

In [4] it is proved that each commutative language over an alphabet of n letters possesses a test set the size of which is at most $2^n(n! + n) + 5n^2$. This upper bound is improved to $O(n^2)$ and this order of magnitude is shown to be the best possible. At last it is proved that for each commutative language with a linear Parikh-map a test set of size $O(n \log n)$ can be effectively found.

This paper is organized as follows. In the first section some prerequisites in the theory of formal languages and combinatorics on words are given.

In section 2, after some simple results on systems of word equations, it is verified that each commutative language over an alphabet of n symbols possesses a test set of size at most $3n^2 - 2n$.

In the third section we introduce a finite language F over $3n$ letter alphabet such that each test set of F is at least of the size n^2.

In section 4 we prove that each commutative language L over an n letter alphabet such that the Parikh-map of L is a linear set has a test set of size at most $2n\lceil\log(n-1)\rceil + 9n$. The procedure to construct the test set is effective.

1. PRELIMINARIES

We assume that the reader is familiar with the basic notions of formal language theory and combinatorics on words as presented in [7] and [11].

Let Z be any (finite) alphabet. As usual, Z^* (Z^+, resp.) denotes the free monoid (free semigroup, resp.) generated by Z. Let $w \in Z^*$. Then $|w|$ denotes the length of the word w and, for each $a \in Z$, $|w|_a$ is the number of occurrences of the symbol a in w. Let $\mathrm{alph}(w) = \{a \in Z \mid |w|_a > 0\}$ and $c(w) = \{u \in Z^* \mid |u|_a = |w|_a \text{ for each } a \in Z\}$. The empty word (*i.e.* the word with length zero) is denoted by ε. The word w is *primitive* if it is nonempty and for each $u \in Z^*$ and $n \in \mathbb{N}$ the equality $w = u^n$ implies $w = u$ (and, of course, $n = 1$). The words w and u are *conjugate* (*words of each other*) if there exist words w_1 and w_2 such that $w = w_1 w_2$ and $u = w_2 w_1$. For each nonempty word $u \in Z^*$ there exist a unique primitive word $t \in Z^*$ (*the primitive root of* u) such that $u \in t^+$. The morphisms h and g on Z^* are *length equivalent on* w if $|h(w)| = |g(w)|$.

For each language $L \subseteq Z^*$, let $\mathrm{alph}(L) = \bigcup_{w \in L} \mathrm{alph}(w)$. The *commutative closure* of the language $L \subseteq Z^*$ is the set $c(L) = \bigcup_{w \in L} c(w)$. We say that L is *commutative* if $L = c(L)$. The morphisms h and g on Z^* are *length equivalent on a language* L if they are length equivalent on each word of L.

Let $\mathbb{N}$ be the set of all natural numbers and $\mathbb{N}_+ = \mathbb{N}\backslash\{0\}$. For each $n \in \mathbb{N}_+$, let $a_1, a_2, ..., a_n$ be distinct symbols. The traditional Parikh-map Ψ_n (Ψ, when n is understood) from $\{a_1, a_2, ..., a_n\}^*$ onto $\mathbb{N}^n$ is defined by $\Psi_n(w) = (|w|_{a_1}, |w|_{a_2}, ..., |w|_{a_n})$.

Let $n \in \mathbb{N}_+$ and P a language over the alphabet $\{a_1, a_2, ..., a_n\}$. A *basis* of P is any finite subset F of P such that (i) in the set $\{\Psi_n(v) | v \in F\}$ there are $|F|$ elements that are linearly independent (over $\mathbb{Q}$, the rationals); and (ii) for each $w \in P$, $\Psi_n(w)$ is a linear combination of some vectors in $\{\Psi_n(v) | v \in F\}$.

A set $T \subseteq \mathbb{N}^n$ is *linear* if there exist a number $m \in \mathbb{N}$ and vectors $\overline{v}, \overline{v}_1, ..., \overline{v}_m \in \mathbb{N}^n$ such that $T = \{\overline{v} + k_1\overline{v}_1 + ... + k_m\overline{v}_m | k_1, ..., k_m \in \mathbb{N}\}$. A *semilinear set* is a finite union of linear sets.

Call a commutative language with a linear (semilinear, resp.) Parikh map a CLIP-*language* (a CSLIP-*language*, resp.).

For each finite set S, let $|S|$ be the cardinality of S. For each nonnegative rational number q, let $\lceil q \rceil$ be the smallest integer $k \in \mathbb{N}$ such that $q \leq k$.

The following theorem is a reformulation of some basic results in the theory of combinatorics on words. For the proof, see for instance [11].

THEOREM 1: *Let x and y be nonempty words over the alphabet X. The following three conditions are equivalent.*

(i) *The words x and y are conjugate.*

(ii) *The words x and y are of equal length and there exist unique words $t_1 \in X^*$, $t_2 \in X^+$ such that $t = t_1t_2$ is primitive and $x \in (t_1t_2)^+$ and $y \in (t_2t_1)^+$;*

(iii) *There exists a word $z \in X^*$ such that $xz = zy$.*

Furthermore, if (ii) *holds, then for each $w \in X^*$ we have $xw = wy$ if and only if $w \in (t_1t_2)^* t_1$.*

We next prove a simple result concerning solutions of a system of two word equations with a certain commutation property. It implies three corollaries which are useful later.

THEOREM 2: *Let x and $\overline{x}$ be distinct nonempty words over the alphabet X. The following two conditions are equivalent.*

(i) *There exist words y and $\overline{y}$ in X^* such that $xy = \overline{xy}$ and $yx = \overline{yx}$.*

(ii) *There exist unique words $t_1 \in X^*$ and $t_2 \in X^+$ such that t_1t_2 is primitive and $x, \overline{x} \in (t_1t_2)^* t_1$.*

Furthermore, if (ii) *holds, then for each* w, $\overline{w} \in X^*$ *we have* $xw = \overline{x}\overline{w}$ *and* $wx = \overline{w}\overline{x}$ *if and only if* $|xw| = |\overline{x}\overline{w}|$ *and* w, $\overline{w} \in (t_2 t_1)^* t_2 \cup \{\varepsilon\}$.

Proof: Obviously (ii) implies (i).

Assume that (i) holds, and, without loss of generality, that $|x| > |\overline{x}|$. There then exists words d_1, $d_2 \in X^+$ such that $x = \overline{x} d_2 = d_1 \overline{x}$. By Theorem 1 there exist unique words $t_1 \in X^*$ and $t_2 \in X^+$ such that $d_1 \in (t_1 t_2)^+$, $d_2 \in (t_2 t_1)^+$ and $\overline{x} \in (t_1 t_2)^* t_1$. Then $x \in (t_1 t_2)^* t_1$ (in fact $x \in (t_1 t_2)^+ t_1$).

Let now w, $\overline{w} \in X^*$ be any words such that $xw = \overline{x}\overline{w}$ and $wx = \overline{w}\overline{x}$. Then certainly $\overline{w} = d_2 w = w d_1$ (since $x = \overline{x} d_2 = d_1 \overline{x}$). If $t_1 = \varepsilon$ (*i.e.* $d_1 = d_2$), the words $w, \overline{w}$ are clearly in $(t_2 t_1)^* t_2 \cup \{\varepsilon\}$. Assume that $t_1 \neq \varepsilon$. Then, again by Theorem 1, we have $w \in (t_2 t_1)^* t_2$ and also $\overline{w} = w d_1 \in (t_2 t_1)^* t_2$. $\square$

COROLLARY 3: *Let* $x, y, z, \overline{x}, \overline{y}, \overline{z}$ *be words such that* $|x| \neq |\overline{x}|$, $|y| = |z|$ *and*

$$\begin{cases} xy = \overline{x}\overline{y} & xz = \overline{x}\overline{z} \\ yx = \overline{y}\overline{x} & zx = \overline{z}\overline{x}. \end{cases}$$

Then $y = z$ *and* $\overline{y} = \overline{z}$.

Proof: If $x = \varepsilon$ or $\overline{x} = \varepsilon$, then certainly all the words x, y, z, $\overline{x}$, $\overline{y}$ and $\overline{z}$ are powers of the same (primitive) word. Since $|y| = |z|$ (and $|\overline{y}| = |\overline{z}|$), the equalities $y = z$ and $\overline{y} = \overline{z}$ hold.

Assume that $x \neq \varepsilon$ and $\overline{x} \neq \varepsilon$. By Theorem 2, there exist unique words $t_1 \in X^*$ and $t_2 \in X^+$ such that y, z, $\overline{y}$, $\overline{z} \in (t_2 t_1)^* t_2 \cup \{\varepsilon\}$. Since $|y| = |z|$ (and $|\overline{y}| = |\overline{z}|$), we have $y = z$ and $\overline{y} = \overline{z}$. $\square$

COROLLARY 4: *Let* x, y, z, $\overline{x}$, $\overline{y}$ *and* $\overline{z}$ *be words such that*

$$\begin{cases} xy = \overline{x}\overline{y} & xz = \overline{x}\overline{z} & yz = \overline{y}\overline{z} \\ yx = \overline{y}\overline{x} & zx = \overline{z}\overline{x} & zy = \overline{z}\overline{y}. \end{cases}$$

Then either $x = \overline{x}$, $y = \overline{y}$ *and* $z = \overline{z}$ *or all the words* x, y, z, $\overline{x}$, $\overline{y}$ *and* $\overline{z}$ *are powers of the same primitive words.*

Proof: Assume that $x \neq \overline{x}$ (and that $y \neq \overline{y}$ and $z \neq \overline{z}$).

If any of the words x, y, z, $\overline{x}$, $\overline{y}$, $\overline{z}$ is empty we are certainly through.

Assume that all the words x, y, z, $\overline{x}$, $\overline{y}$, $\overline{z}$ are nonempty. By Theorem 2 there exist unique words $t_1 \in X^*$ and $t_2 \in X^+$ such that $t_1 t_2$ is primitive and x, $\overline{x} \in (t_1 t_2)^* t_1$ and y, $\overline{y}$, z, $\overline{z} \in (t_2 t_1)^* t_2$. Since $yz = \overline{y}\overline{z}$ and $y \neq \overline{y}$, there exist integers r_1, r_2, s_1, $s_2 \in \mathbb{N}$, $r_1 \neq r_2$, $s_1 \neq s_2$ such that $y = (t_2 t_1)^{r_1} t_2$, $\overline{y} = (t_2 t_1)^{r_2} t_2$, $z = (t_1 t_2)^{s_1} t_1$, $\overline{z} = (t_1 t_2)^{s_2} t_1$ and $(t_2 t_1)^{r_1} t_2 (t_2 t_1)^{s_1} t_2 = (t_2 t_1)^{r_2} t_2 (t_2 t_1)^{s_2} t_2$. Since $r_1 \neq r_2$, the equation

$t_1t_2 = t_2t_1$ holds. Since t_1t_2 is primitive, the word t_1 is empty. Thus $x, y, z, \overline{x}, \overline{y}, \overline{z} \in t_2^*$. □

Note: The equation $zy = \overline{zy}$ is not necessary in the previous corollary.

COROLLARY 5: *Let x, y, z, $\overline{x}$, $\overline{y}$ and $\overline{z}$ be words such that $x\overline{x} \neq \varepsilon$, $y\overline{y} \neq \varepsilon$ and $z\overline{z} \neq \varepsilon$ and*

$$\begin{cases} xyz = \overline{xyz} & zyx = \overline{zyx} \\ yzx = \overline{yzx} & yxz = \overline{yxz} \\ xzy = \overline{xzy} & zxy = \overline{zxy}. \end{cases}$$

Then either $x = \overline{x}$, $y = \overline{y}$ and $z = \overline{z}$ or all the words x, y, $z, \overline{x}$, $\overline{y}$ and $\overline{z}$ are powers of the same primitive word.

Proof: Assume that either $x \neq \overline{x}$ or $y \neq \overline{y}$ or $z \neq \overline{z}$. Suppose without loss of generality that $x \neq \overline{x}$. Then, by Corollary 3, we have $yz = zy$ and $\overline{yz} = \overline{zy}$. There thus exist primitive words t and l such that $y, z \in t^*$ and $\overline{y}, \overline{z} \in l^*$. Since $x \neq \overline{x}$, we have $yz \neq \overline{yz}$ implying that either $y \neq \overline{y}$ or $z \neq \overline{z}$. Assume without loss of generality that $y \neq \overline{y}$. Then, again by Corollary 3, the equalities $xz = zx$ and $\overline{xz} = \overline{zx}$ hold implying $x \in t^*$ and $\overline{x} \in l^*$. Since $xyz = \overline{xyz}$ and t and l are primitive, we have $t = l$. Thus $x, y, z, \overline{x}, \overline{y}, z \in t^*$ and the proof is complete. □

The last auxiliary result of this section tells that to guarantee that two morphsims h and g are length equivalent on a language L it suffices to consider the length equivalence of h and g on some basis of L.

LEMMA 6: *Let L be a language over the alphabet X, F a basis of L and h and g two morphisms on X^*. Then h and g are length equivalent on L if and only if they are length equivalent on F.*

Proof: Assume without loss of generality that $X = \{a_1, a_2, ..., a_n\}$ for some $n \in \mathbb{N}_+$. If h and g are length equivalent on L, they certainly are length equivalent on a subset F of L.

Assume that h and g are length equivalent on F. Let $r_i = |h(a_i)|$ and $s_i = |g(a_i)|$ for each $i = 1, 2, ..., n$. Let $z \in L$. Since F is a basis of L, there exist an integer $m \in \mathbb{N}_+$, (distinct) words $x_1, x_2, ..., x_m \in F$ and rational numbers $\alpha_1, \alpha_2, ..., \alpha_m$ such that

$$\Psi(z) = \alpha_1 \Psi(x_1) + \alpha_2 \Psi(x_2) + ... + \alpha_m \Psi(x_m).$$

Thus

$$|h(z)| = \Psi(z)(r_1, \ldots, r_n)^T = \Sigma_{i=1}^m \alpha_i \Psi(x_i)(r_1, \ldots, r_n)^T$$
$$= \Sigma_{i=1}^m \alpha_i \Psi(x_i)(s_1, \ldots, s_n)^T = |g(z)|.$$

where $(r_1, \ldots, r_n)^T$ $((s_1, \ldots, s_n)^T$, resp.) is the vector transpose of $(r_1, \ldots, r_n)$ $((s_1, \ldots, s_n)$, resp.) and vector multiplication is applied. Above the third equality holds since $|h(x_i)| = |g(x_i)|$ implies

$$|h(x_i)| = \Psi(x_i)(r_1, \ldots, r_n)^T = \Psi(x_i)(s_1, \ldots, s_n)^T = |g(x_i)|$$

for each $i = 1, 2, \ldots, n$. □

Note: The previous lemma implies (*see* also [4]) that if a language $L \subseteq \{a_1, a_2, \ldots, a_n\}^*$ has a basis F such that $|F| = n$, then F necessarily is a test set for L.

2. CONSTRUCTING TEST SETS FOR COMMUTATIVE LANGUAGES

Let L be a commutative language over the alphabet X.

For each unordered pair $\{a, b\}$ of two distinct symbols in X construct the language $L_{\{a,b\}}$ as follows.

If $L \cap ab\,X^* = \emptyset$, then $L_{\{a,b\}} = \emptyset$.

Assume that $L \cap ab\,X^* \neq \emptyset$. We have three possibilities: 1° $L \cap a^2\,b\,X^* \neq \emptyset$; 2° $L \cap a^2\,b\,X^* = \emptyset$ and $L \cap ab^2\,X^* \neq \emptyset$; 3° $L \cap a^2\,b\,X^* = L \cap ab^2\,X^* = \emptyset$.

Case 1°. Let $x \in X^*$ be a word such that $a^2\,bx \in L$. Then

$$L_{\{a,b\}} = \{ab\,(ax),\ ba\,(ax),\ a\,(ax)\,b,\ b\,(ax)\,a,\ (ax)\,ab,\ (ax)\,ba\}.$$

Case 2°. Let $y \in X^*$ be a word such that $ab^2\,y \in L$. Then

$$L_{\{a,b\}} = \{ab\,(by),\ ba\,(by),\ a\,(by)\,b,\ b\,(by)\,a,\ (by)\,ab,\ (by)\,ba\}.$$

Case 3°. Let $z \in X^*$ be a word such that $abz \in L$. Then

$$L_{\{a,b\}} = \{abz,\ baz,\ azb,\ bza,\ zab,\ zba\}.$$

Let B be a basis of L such that, for each $a \in X$, if $L \cap a^+ \neq \emptyset$, then $a^r \in B$ where r is the smallest number $m \in \mathbb{N}_+$ such that $a^m \in L$. Let

$$T_L = \bigcup_{\substack{a,\,b \in X \\ a \neq b}} L_{\{a,b\}} \cup B.$$

Obviously $|T_L| \leq 6\binom{n}{2} + n = 3\,n^2 - 2\,n$, where $n = |X|$.

We shall next prove that T_L is a test set for L.

THEOREM 7: *Let L be a commutative language over the alphabet X. Then T_L is a test set for L.*

Proof: Let h and g be morphisms on X^* such that $h(x) = g(x)$ for each $x \in T_L$. Let Y denote $\{a \in X | h(a) \neq \varepsilon \text{ or } g(a) \neq \varepsilon\}$. Let $z \in L$.

If $\text{alph}(z) \cap Y = \emptyset$, then certainly $h(z) = g(z) = \varepsilon$.

Suppose that $\text{alph}(z) \cap Y \neq \emptyset$. Consider three cases: 1° $|\text{alph}(z) \cap Y| = 1$; 2° $|\text{alph}(z) \cap Y| = 2$; and 3° $|\text{alph}(z) \cap Y| > 2$.

Case 1°. Let $a \in X$ be such that $\text{alph}(z) \cap Y = \{a\}$. There surely exists a word v such that $av \in T_L$. Then $h(av) = g(av)$ by the assumption. By Lemma 6, $|h(a^{|z|_a})| = |h(z)| = |g(z)| = |g(a^{|z|_a})|$. Thus $|h(a)| = |g(a)|$ which implies that $h(a) = g(a)$.

Case 2°. Let $a, b \in X$, $a \neq b$, be such that $\text{alph}(z) \cap Y = \{a, b\}$. If $h(a) = g(a)$ and $h(b) = g(b)$, then clearly $h(z) = g(z)$. Assume without loss of generality that $h(a) \neq g(a)$. Consider first the case that either $a^2 b X^* \cap L \neq \emptyset$ or $ab^2 X^* \cap L \neq \emptyset$. Assume without loss of generality that $a^2 b X^* \cap L \neq \emptyset$. By construction, there exists a word $u \in X^*$ such that $abau$, $baau$, $aaub$, $baua$, $auab$, $auba \in T_L$. Then

$$\begin{cases} h(a)h(b)h(au) = g(a)g(b)g(au) & h(b)h(au)h(a) = g(b)g(au)g(a) \\ h(b)h(a)h(au) = g(b)g(a)g(au) & h(au)h(a)h(b) = g(au)g(a)g(b) \\ h(a)h(au)h(b) = g(a)g(au)g(b) & h(au)h(b)h(a) = g(au)g(b)g(a). \end{cases}$$

By Corollary 5, the words $h(a)$, $h(b)$, $g(a)$ and $g(b)$ are powers of the same (primitive) word. By Lemma 6,

$$|h(z)| = |h(a^{|z|_a} b^{|z|_b})| = |g(a^{|z|_a} b^{|z|_b}) = |g(z)|.$$

Then $h(z) = g(z)$. Let us now turn to the case $a^2 b X^* \cap L = ab^2 X^* \cap L = \emptyset$. Then, by construction, there exists a word in X^* such that abw, baw, awb, bwa, wab, $wba \in T_L$. Then

$$\begin{cases} h(a)h(b)h(w) = g(a)g(b)g(w) & h(b)h(w)h(a) = g(b)g(w)g(a) \\ h(b)h(a)h(w) = g(b)g(a)g(w) & h(w)h(a)h(b) = g(w)g(a)g(b) \\ h(a)h(w)h(b) = g(a)g(w)g(b) & h(w)h(b)h(a) = g(w)g(b)g(a). \end{cases}$$

If $h(w) \neq \varepsilon$ or $g(w) \neq \varepsilon$ then, just as above, the words $h(a)$, $h(b)$, $g(a)$ and $g(b)$ are powers of the same primitive word and we are through.

Assume that $h(w) = g(w) = \varepsilon$. Then

$$\begin{cases} h(a)\,h(b) = g(a)\,g(b) \\ h(b)\,h(a) = g(b)\,g(a). \end{cases}$$

and since either $h(z) = h(ab)$ and $g(z) = g(ab)$ or $h(z) = h(ba)$ and $g(z) = g(ba)$, we must have $h(z) = g(z)$.

Case 3°. Assume now that $|\mathrm{alph}(z) \cap Y| > 2$. If $h(a) = g(a)$ for each $a \in \mathrm{alph}(z) \cap Y$, then $h(z) = g(z)$. Let $a \in \mathrm{alph}(z) \cap Y$ be such that $h(a) \neq g(a)$. Let b and c be any two symbols in $\mathrm{alph}(z) \cap Y$ such that $b \neq a \neq c$. By construction, there exist words u_1, u_2, $u_3 \in X^*$ such that the words abu_1, bau_1, au_1b, bu_1a, u_1ab, u_1ba, acu_2, cau_2, au_2c, cu_2a, u_2ac, u_2ca, bcu_3, cbu_3, bu_3c, cu_3b, u_3bc, u_3cb are all in T_L. Thus

$$\begin{cases} h(a)\,h(b)\,h(u_1) = g(a)\,g(b)\,g(u_1) & h(c)\,h(u_2)\,h(a) = g(b)\,g(u_2)\,g(a) \\ h(b)\,h(a)\,h(u_1) = g(b)\,g(a)\,g(u_1) & h(u_2)\,h(a)\,h(c) = g(u_2)\,g(a)\,g(c) \\ h(a)\,h(u_1)\,h(b) = g(a)\,g(u_1)\,g(b) & h(u_2)\,h(c)\,h(a) = g(u_2)\,g(c)\,g(a) \\ h(b)\,h(u_1)\,h(a) = g(b)\,g(u_1)\,g(a) & h(b)\,h(c)\,h(u_3) = g(b)\,g(c)\,g(u_3) \\ h(u_1)\,h(a)\,h(b) = g(u_1)\,g(a)\,g(b) & h(c)\,h(b)\,h(u_3) = g(c)\,g(b)\,g(u_3) \\ h(u_1)\,h(b)\,h(a) = g(u_1)\,g(b)\,g(a) & h(b)\,h(u_3)\,h(c) = g(b)\,g(u_3)\,g(c) \\ h(a)\,h(c)\,h(u_2) = g(a)\,g(c)\,g(u_2) & h(c)\,h(u_3)\,h(b) = g(c)\,g(u_3)\,g(b) \\ h(c)\,h(a)\,h(u_2) = g(c)\,g(a)\,g(u_2) & h(u_3)\,h(b)\,h(c) = g(u_3)\,g(b)\,g(c) \\ h(a)\,h(u_2)\,h(c) = g(a)\,g(u_2)\,g(c) & h(u_3)\,h(c)\,h(b) = g(u_3)\,g(c)\,g(b). \end{cases}$$

We show that all the words $h(a)$, $h(b)$, $h(c)$, $g(a)$, $g(b)$ and $g(c)$ are powers of the same (primitive) word.

Assume first that $h(u_1)\,g(u_1) \neq \varepsilon$. Then, by Corollary 5, there exists a primitive word t such that $h(a)$, $h(b)$, $g(a)$, $g(b)$, $h(u_1)$, $g(u_1) \in t^*$. If either $h(u_2)\,g(u_2) \neq \varepsilon$ or $h(u_3)\,g(u_3) \neq \varepsilon$, we have (again by Corollary 5) that either $h(a)$, $h(c)$, $g(a)$, $g(c) \in t^*$ or $h(b)$, $h(c)$, $g(b)$, $g(c) \in t^*$ and we are done. Suppose that $h(u_2)\,g(u_2) = h(u_3)\,g(u_3) = \varepsilon$. Then the previous system of equations implies

$$\begin{cases} h(a)\,h(c) = g(a)\,g(c) \\ h(c)\,h(a) = g(c)\,g(a). \end{cases}$$

Since $h(a)$, $g(a) \in t^*$ and $h(a) \neq g(a)$, it is clear that $h(c)$, $g(c) \in t^*$.

Let now $h(u_1)\,g(u_1) = \varepsilon$. Then, since $h(a) \neq g(a)$, it must be $h(b) \neq g(b)$. If now either $h(u_2)\,g(u_2) \neq \varepsilon$ or $h(u_3)\,g(u_3) \neq \varepsilon$, we are through as above. Assume thus that $h(u_2)\,g(u_2) = h(u_3)\,g(u_3) = \varepsilon$. Then we have

$$\begin{cases} h(a)\,h(b) = g(a)\,g(b) & h(c)\,h(a) = g(c)\,g(a) \\ h(b)\,h(a) = g(b)\,g(a) & h(b)\,h(c) = g(b)\,g(c) \\ h(a)\,h(c) = g(a)\,g(c) & h(c)\,h(b) = g(c)\,g(b). \end{cases}$$

By Corollary 4, $h(a)$, $h(b)$, $h(c)$, $g(a)$, $g(b)$ and $g(c)$ are powers of the same primitive word. $\square$

3. A LOWER BOUND OF SIZE $\Omega(n^2)$

Let $n \in \mathbb{N}_+$ and $b_1, b_2, \ldots, b_n, c_1, c_2, \ldots, d_1, d_2, \ldots, d_n$ be distinct symbols. Let $F_1 = \{b_i\,c_j\,d_j \mid i, j = 1, 2, \ldots, n\}$ and $F = c(F_1)$. Thus F is a commutative language such that $|F| = 6\,n^2$.

Consider any subset Y of F such that $|Y| < n^2$. There then exist $i, j \in \{1, 2, \ldots, n\}$ such that $c(b_i\,c_j\,d_j) \cap Y = \emptyset$. Without loss of generality we may assume that $i = j = n$. Let a and b be distinct symbols. Define two morphisms h_1 and g_1 on $\{b_1, b_2, \ldots, b_n, c_1, c_2, \ldots, c_n, d_1, d_2, \ldots, d_n\}^*$ as follows:

$$h_1(b_i) = h_1(c_i) = h_1(d_i) = g_1(b_i) = g_1(c_1) = g_1(d_i) = a$$

for each $i \in \{1, 2, \ldots, n-1\}$, and

$$h_1(b_n) = g_1(b_n) = b \quad h_1(c_n) = g_1(d_n) = a^2 \quad h_1(d_n) = g_1(c_n) = a.$$

Then certainly $h_1(y) = g_1(y)$ for each $y \in Y$. On the other hand

$$h_1(c_n\,b_n\,d_n) = a^2\,ba \neq aba^2 = g_1(c_n\,b_n\,d_n).$$

Thus Y is not a test set for F.

Consider the example above with erasing morphisms. Define the two morphisms h_2 and g_2 on $\{b_1, b_2, \ldots, b_n, c_1, c_2, \ldots, c_n, d_1, d_2, \ldots, d_n\}^*$ as follows. Let $h_2(b_i) = g_2(b_i) = \varepsilon$ for $i = 1, 2, \ldots, n-1$ and $h_2(b_n) = g_2(b_n) = a$. Let

$$h_2(c_j) = g_2(c_j) = h_2(d_j) = g_2(d_j) \qquad \text{for} \quad j = 1, 2, \ldots, n-1,$$

and $h_2(c_n) = (ab)^2\,a$, $g_2(c_n) = (ab)\,a$, $h_2(d_n) = (ba)\,b$, and $g_2(d_n) = (ba)^2\,b$. Then $h_2(x) = g_2(x) = a^2$ for each $x \in c(\{b_i\,c_j\,d_j\})$

where $i, j \in \{1, 2, ..., n-1\}$. For each $y \in c(\{b_j\, c_n\, d_n\})$ where $j \in \{1, 2, ..., n-1\}$ we have $h_2(y) = g_2(y) \in \{(ab)^4, (ba)^4\}$. Certainly

$$h_2(c_n\, b_n\, d_n) = (ab)^2\, a^2\, (ba)\, b \neq (ab)\, a^2\, (ba)^2\, b = g_2(c_n\, b_n\, d_n).$$

We have thus proved

THEOREM 8: *The lower bound for the size of a test set for languages from the family of all commutative languages over an alphabet of n symbols is* $\Omega(n^2)$.

Note: By construction, the previous theorem remains true if the string' commutative languages' is substituted by the word 'CSLIP-languages'.

4. TEST SETS FOR COMMUTATIVE LANGUAGES WITH A LINEAR PARIKH-MAP

In the following we shall see that each CLIP-language over an alphabet of n symbols possesses a test set of size $O(n \log n)$.

For each m and j in $\mathbb{N}$, $j \leq m$, define the function p_{mj} from $(X^*)^{2m}$ into X^* inductively as follows.

$$p_{m0}(w_1, ..., w_{2^m}) = w_1 ... w_{2^m}$$
$$p_{m1}(w_1, ..., w_{2^m}) = (w_{2^{m-1}+1} ... w_{2^m})(w_1 ... w_{2^m - 1})$$
$$p_{m+1, j+1}(w_1, ..., w_{2^{m+1}}) = p_{mj}(w_1, ..., w_{2^m})\, p_{mj}(w_{2^m+1}, ..., w_{2^{m+1}})$$

The classical result concerning the word equation $xy = yx$ can now be generalized.

THEOREM 9: *Let* $m \in \mathbb{N}_+$ *be a number and* $x_1, x_2, ..., x_{2^m}$ *words in* X^* *such that*

$$x_1 ... x_{2^m} = p_{mj}(x_1, ..., x_{2^m})$$

for $j = 1, 2, ..., m$. *Then the words* $x_1, x_2, ..., x_{2^m}$ *are powers of the same (primitive) word.*

Proof: By induction on m.

The case $m = 1$ is trivial: certainly $x_1\, x_2 = x_2\, x_1$ implies the claim. Assume that the theorem is true for $m = k$.

Consider the case $m = k + 1$. Since

$$\begin{aligned}(x_1 ... x_{2^k})(x_{2^k+1} ... x_{2^{k+1}}) &= p_{k+1,1}(x_1, ..., x_{2^{k+1}}) \\ &= (x_{2^k+1} ... x_{2^{k+1}})(x_1 ... x_{2^k}),\end{aligned}$$

we notice that there exists a (primitive) word t such that $x_1 \ldots x_{2^k}$, $x_{2^k+1} \ldots x_{2^{k+1}} \in t^*$. Also, by assumption,

$$(x_1 \ldots x_{2^k})(x_{2^k+1} \ldots x_{2^{k+1}}) = p_{kj}(x_1, \ldots, x_{2^k}) \, p_{kj}(x_{2^k+1}, \ldots, x_{2^{k+1}})$$

for each $j \in \{1, , \ldots, k\}$ implying

$$\begin{cases} x_1 \ldots x_{2^k} = p_{kj}(x_1, \ldots, x_{2^k}) \\ x_{2^k+1} \ldots x_{2^{k+1}} = p_{kj}(x_{2^k+1}, \ldots, x_{2^{k+1}}) \end{cases}$$

for each $j \in \{1, 2, \ldots, k\}$. By induction, there exist (primitive) words t_1 and t_2 such that $x_1, \ldots, x_{2^k} \in t_1^*$ and $x_{2^k+1}, \ldots, x_{2^{k+1}} \in t_2^*$. Since $x_1 \ldots x_{2^k}$, $x_{2^k+1} \ldots x_{2^{k+1}} \in t^*$, we have $t_1 = t_2 = t$. Thus $x_1, \ldots, x_{2^{k+1}} \in t^*$ and the induction is extended.

We still give an example. Assume that $m = 3$. Then we have the following system of equations

$$\begin{cases} x_1x_2x_3x_4x_5x_6x_7x_8 = (x_5x_6x_7x_8)(x_1x_2x_3x_4) \\ x_1x_2x_3x_4x_5x_6x_7x_8 = (x_7x_8)(x_5x_6)(x_3x_4)(x_1x_2) \\ x_1x_2x_3x_4x_5x_6x_7x_8 = x_8x_7x_6x_5x_4x_3x_2x_1. \end{cases}$$

The last two equations imply that $x_1, x_2 \in p_1^*$, $x_3, x_4 \in p_2^*$, $x_5, x_6 \in p_3^*$ and $x_7, x_8 \in p_4^*$ where p_1, p_2, p_3 and p_4 are primitive words. From the first and the second equation we obtain that $p_1 = p_2$ and $p_3 = p_4$. Finally, the first equation gives $p_1 = p_2 = p_3 = p_4$. $\square$

Let L be a CLIP-language over the alphabet $\{a_1, a_2, \ldots, a_n\}$, where $n \geq 2$. By definition, there exist a number $p \in \mathbb{N}_+$ and words $u_0, u_1, \ldots, u_p$ such that $L = c(u_0 u_1^* \ldots u_p^*)$.

Let $u = u_0 u_1^2 \ldots u_p^2$ and $m = \lceil \log(n-1) \rceil$. Thus m is the smallest number $k \in \mathbb{N}$ such that $n - 1 \leq 2^k$. Let $a_{n+1}, \ldots, a_{2^m}$ be new symbols and $r_j = |u|_{a_j}$ for each $j \in \{1, 2, \ldots, 2^m\}$.

Note that each symbol a_i occurs exactly once (at least twice, resp.) in u if and only if it occurs exactly once (at least twice, resp.) in some word of in $c(u_0 u_1^* \ldots u_p^*)$.

Using the words in $c(u)$ we construct a test set (of size $O(n \log n)$) for the language $L = c(u_0 u_1^* \ldots u_p^*)$

For each $i \in \{1, 2, \ldots, n\}$, let the words w_{i1}, w_{i2} and w_{i3} be defined as follows.

If $r_i = 0$, let $w_{i1} = w_{i2} = w_{i3} = \varepsilon$.

If $r_i = 1$, let $w_{i_1} = a_1^{r_1} \dots a_{i-1}^{r_{i-1}}$, $w_{i2} = a_i$ and $w_{i3} = a_{i+1}^{r_{i+1}} \dots a_n^{r_n}$.

If $r_i \geq 2$, let $w_{i_1} = w_{i_2} = a_i$ and $w_{i_3} = a_1^{r_1} \dots a_{i-1}^{r_{i-1}} a_i^{r_i - 2} a_{i+1}^{r_{i+1}} \dots a_n^{r_n}$.

Let $A(u_0; u_1, \dots, u_p)$ be the set of all words $w_{i_{\sigma(1)}} w_{i_{\sigma(2)}} w_{i_{\sigma(3)}}$ where σ is any permutation of 1, 2, 3 and $i = 1, 2, \dots, n$. Clearly $|A(u_0; u_1, \dots, u_p)| \leq 6n$.

For each $i \in \{1, 2, \dots, n\}$ define the words $v_{i1}, v_{i2}, \dots, v_{i2^m}$ as follows.

$$
\begin{aligned}
v_{ij} &= a_j^{r_j} && \text{for} \quad j = 1, \dots, i-1; \qquad \text{and} \\
v_{ij} &= a_{j+1}^{r_{j+1}} && \text{for} \quad j = i, i+1, \dots, 2^m.
\end{aligned}
$$

Let

$$
\begin{aligned}
B(u_0; u_1, \dots, u_p) = \{a_i^{r_i}\, p_{mk}(v_{i1}, \dots, v_{i2^m}),\ p_{mk}(v_{i1}, \dots, v_{i2^m})\, a_i^{r_i} \mid i \\
= 1, 2, \dots, n,\ k = 0, 1, \dots, m\}.
\end{aligned}
$$

Obviously $B(u_0; u_1, \dots, u_p) \subseteq L$ and $|B(u_0; u_1, \dots, u_p)| \leq 2n(m+1)$. Let $C(u_0; u_1, \dots, u_p) \subseteq \{u, uu_1, \dots, uu_p\}$ be a base of L and

$$
\begin{aligned}
&T(u_0; u_1, \dots, u_p) = \\
&\quad A(u_0; u_1, \dots, u_p) \cup B(u_0; u_1, \dots, u_p) \cup C(u_0; u_1, \dots, u_p).
\end{aligned}
$$

Then $T(u_0; u_1, \dots, u_p) \subseteq L$ and $|T(u_0; u_1, \dots, u_p)| \leq 2nm + 9n \leq 2n(\lceil \log(n-1) \rceil + 9n$. It is a bit tedious but straightforward to prove the following.

THEOREM 10: *Let $p \in \mathbb{N}$ be a number and $u_0, u_1, \dots, u_p$ be words over the alphabet $\{a_1, a_2, \dots, a_n\}$, where $n \geq 2$. Then $T(u_0; u_1, \dots, u_p)$ is a test set for the language $c(u_0 u_1^* \dots u_p^*)$.*

Proof: We use the notation preceding the theorem. Denote $L = c(u_0 u_1^* \dots u_p^*)$ and $D = D(u_0; u_1, \dots, u_p)$ for each $D \in \{A, B, C, T\}$.

Consider two morphisms h and g defined on $\{a_1, a_2, \dots, a_n\}^*$ such that $h(x) = g(x)$ for each $x \in T$. We shall show that $h(z) = g(z)$ for each $z \in L$.

If $h(a_i) = g(a_i)$ for each $i \in \{1, 2, \dots, n\}$, there remains nothing to prove.

Assume thus that $h(a_j) \neq g(a_j)$ for some $j \in \{1, 2, \dots, n\}$. Let Y be the set of all $j \in \{1, 2, \dots, n\}$ such that $h(a_j) \neq g(a_j)$. Since T, by construction, contains a base C of L, the morphisms h and g, by Lemma 6, are length equivalent on L. This certainly implies that $|Y| \geq 2$.

Suppose, without loss of generality, that there exists $s \in Y$, $1 < s < n$ such that both $h(w_{s1})\, g(w_{s1})$ and $h(w_{s3})\, g(w_{s3})$ are nonempty. By the construction of A, we have $w_{s\sigma(1)}\, w_{s\sigma(2)}\, w_{s\sigma(3)} \in T$ for each permutation σ of 1, 2, 3. Then

$$h(w_{s\sigma(1)})\, h(w_{s\sigma(2)})\, h(w_{s\sigma(3)}) = g(w_{s\sigma(1)})\, g(w_{s\sigma(2)})\, g(w_{s\sigma(3)})$$

for each permutation σ of 1, 2, 3. By Corollary 5, there exists a primitive word t such that all the words $h(w_{s1})$, $h(w_{s2})$, $h(w_{s3})$, $g(w_{s1})$, $g(w_{s2})$ and $g(w_{s3})$ are in t^*. Since $w_{s2} = a_s$, we have $h(a_s)$, $g(a_s) \in t^*$ as well as the words $h(a_1^{r_1} \cdots a_{s-1}^{r_{s-1}})$, $h(a_{s+1}^{r_{s+1}} \cdots a_n^{r_n})$, $g(a_1^{r_1} \cdots a_{s-1}^{r_{s-1}})$ and $g(a_{s+1}^{r_{s+1}} \cdots a_n^{r_n})$ respectively. By the construction of B the words

$$a_s^{r_s}\, p_{mk}(v_{s1}, \ldots, v_{s2}), \qquad p_{mk}(v_{s1}, \ldots, v_{s2^m})\, a_s^{r_s}$$

are in T for $k = 0, 1, \ldots, m$. By assumption

$$\begin{aligned} h(a_s^{r_s})\, p_{mk}(h(v_{s1}), \ldots, h(v_{s2^m})) &= g(a_s^{r_s})\, p_{mk}(g(v_{s1}), \ldots, g(v_{s2^m})) \\ p_{mk}(h(v_{s1}), \ldots, h(v_{s2^m}))\, h(a_s^{r_s}) &= p_{mk}(g(v_{s1}), \ldots, g(v_{s2^m}))\, g(a_s^{r_s}) \end{aligned}$$

for $k = 0, 1, \ldots, m$. Since $h(a_s^{r_s}) \neq g(a_s^{r_s})$, we have, by Corollary 3, that

$$\begin{aligned} h(v_{s1}) \ldots h(v_{s2^m}) &= p_{mk}(h(v_{s1}), \ldots, h(v_{s2^m})) \\ g(v_{s1}) \ldots g(v_{s2^m}) &= p_{mk}(g(v_{s1}), \ldots, g(v_{s2^m})) \end{aligned}$$

for $k = 0, 1, \ldots, m$. By Theorem 9 there exist primitive words t_1 and t_2 such that $h(v_{s1}), \ldots, h(v_{s2^m}) \in t_1^*$ and $g(v_{s1}), \ldots, g(v_{s2^m}) \in t_2^*$. This means that the words $h(a_1), \ldots, h(a_{s-1})$, $h(a_{s+1}), \ldots, h(a_{2^n})$ are in t_1^* and $g(a_1), \ldots, g(a_{s-1})$, $g(a_{s+1}), \ldots, g(a_{2^n})$ are in t_2^*. Then $t_1 = t_2 = t$. Now all the words $h(a_1), \ldots, h(a_n)$, $g(a_1), \ldots, g(a_n)$ are powers of t. Since h and g are length equivalent on L, the set T is a test set of L. □

COROLLARY 11: *For each* CLIP-*language over an alphabet of n symbols, $n \in \mathbb{N}_+$, there exists a test set of the size $O(n \log n)$.*

The following question remains open.

OPEN PROBLEM: *Does each CLIP-language over an alphabet of n symbols possess a test set of size $O(n)$?*

We do not even know whether or not the language $c(a_1 \ldots a_n)$ has a test set of size $O(n)$.

REFERENCES

1. J. Albert and K. Culik II, *Test sets for homomorphism equivalence on context free languages,* Information and Control, 1980, *45*, pp. 273-284.
2. J. Albert, K. Culik II and J. Karhumäki, *Test sets for context free languages and algebraic systems of equations over free monoid,* Information and Control, 1982, *52*, pp. 172-186.
3. M. H. Albert and J. Lawrence, A proof of Ehrenfeucht's conjecture, *Theoret. Comput. Sci.,* 1985, *41*, pp. 121-123.
4. J. Albert and D. Wood, Checking sets, test sets rich languages and commutatively closed languages, *Journal of Computer and System Sciences*, 1983, *26*, pp. 82-91.
5. I. Hakala and J. Kortelainen, On the system of word equations $x_1^i x_2^i ... x_m^i = y_1^i y_2^i ... y_n^i$ $(i = 1, 2, ...)$ in a free monoid, *Acta Inform.*, 1997, *34*, pp. 217-230.
6. I. Hakala and J. Kortelainen, On the system of word equations $x_0 u_1^i$ $x_0 u_1^i x_1 u_2^i x_2 u_3^i x_3 = y_0 v_1^i y_1 v_2^i y_2 v_3^i y_3$ $(i = 0, 1, 2, ...)$ in a free monoid, *Theor. Comput. Sci.* (to appear).
7. M. A. Harrison, *Introduction to Formal Language Theory,* Addison-Wesley, Reading Massachusetts, 1978.
8. J. Karhumäki, W. Plandowski and W. Rytter, Polynomial-size test sets for context-free languages, *Lecture Notes in Computer Sciences*, 1992, *623*, pp. 53-64.
9. J. Karhumäki, W. Plandowski and W. Rytter, Polynomial-size test sets for context-free languages, *Journal of Computer and System Sciences,* 1995, *50*, pp. 11-19.
10. J. Karhumäki, W. Plandowski and S. Jarominek, Efficient construction of test sets for regular and context-free languages, *Theor. Comp. Sci.,* 1993, *116*, pp. 305-316.
11. M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading Massachusetts, 1983.