RAIRO-Theor. Inf. Appl. **42** (2008) 525–537
DOI: 10.1051/ita:2008011

Available online at:
www.rairo-ita.org

PARIKH TEST SETS FOR COMMUTATIVE LANGUAGES *

Štěpán Holub[1]

Abstract. A set $T \subseteq L$ is a Parikh test set of L if $c(T)$ is a test set of $c(L)$. We give a characterization of Parikh test sets for arbitrary language in terms of its Parikh basis, and the coincidence graph of letters.

Mathematics Subject Classification. 68R15.

1. Introduction

Commutative languages represent a class of languages with the rare property of having known upper bound for the cardinality of their test sets (see [1,2]). Namely, each commutative language has a test set with cardinality at most $3n^2$, where n is the number of letters. Moreover, the bound is optimal up to a constant, since for each n there is a language the smallest test set of which has cardinality $\frac{1}{9}n^2$. On the other hand, there is a large class of commutative languages having a linear test set. The motivation for this paper was to find out which commutative languages have only test sets of quadratic size.

Commutative languages are fully described by the set of their Parikh vectors; the number of different Parikh vectors present in the language is thus the right measure for the size of test sets. It turns out that even in this measure test sets of a commutative language can be of size $\Omega(n^2)$.

Parikh basis of a language and the information about joint occurrence of letters in a word is basic finite information about a language over finite alphabet. In this paper we show that this information is sufficient to characterize Parikh test set of arbitrary commutative language.

Keywords and phrases. Combinatorics on words, test sets, commutative languages.

* *The author was supported by the research project MSM 0021620839 financed by MSMT and the Grant Agency of Czech Republic, Grant 201/03/D117.*

[1] Charles University, Faculty of Mathematics and Physics, Department of Algebra, Sokolovská 83, 175 86 Praha, Czech Republic; holub@karlin.mff.cuni.cz

Article published by EDP Sciences

For an introduction into the importance of commutative languages in general see for example [3].

2. Definitions and known facts

The reader is supposed to be familiar with common notation and basic facts of combinatorics on words.

Commutative closure of a word $u = \ell_1 \dots \ell_n$, where ℓ_i are letters, is the language

$$c(u) = \{\ell_{\sigma(1)} \dots \ell_{\sigma(n)} \mid \sigma \in S_n\}.$$

In other words, the commutative closure of u contains all words, which arise from u by permutation of its letters. *Commutative closure of a language* L is defined by

$$c(L) = \bigcup_{u \in L} c(u).$$

Language L is said to be *commutative* if $L = c(L)$. We say that morphisms g and h *agree on a language* L if $g(u) = h(u)$ for each $u \in L$, write $g \equiv_L h$. A subset T of a language L is called its *test set* if for any two morphisms g, h

$$g \equiv_L h \quad \Leftrightarrow \quad g \equiv_T h.$$

For sake of clarity we shall sometimes call those sets *classical* test sets.

The set of letters occurring in a word w is denoted by $\operatorname{alph}(w)$.

Parikh vector of a word $u = a_1^{k_1} \dots a_n^{k_n}$ over the alphabet $A = \{a_1, \dots, a_n\}$, where k_i are nonnegative integers denoted also by $|u|_{a_i}$, is defined by

$$\Psi(u) = (k_1, \dots, k_n).$$

Note that a commutative language L over A is given uniquely by the set $\Psi[L] \subseteq \mathbb{N}_0^n$ of its Parikh vectors.

Parikh basis of a commutative language L is a set $B \subseteq L$ such that $\Psi[B]$ is basis of the vector space over $\mathbb{Q}$ generated by $\Psi[L]$.

Two morphisms g and h are said to *agree lengthwise* on L if $|g(u)| = |h(u)|$ for all $u \in L$.

It is easy to see that the following lemma holds.

Lemma 2.1. *Morphisms g and h agree lengthwise on L if and only if they agree lengthwise on a Parikh basis of L.*

The following facts about test sets of commutative languages are known:

Theorem 2.2 (see [1])**.** *Any commutative language over n letters has a test set of cardinality at most $3n^2$.*

There is a commutative language over $3n$ letters the smallest test set of which has cardinality at least n^2.

Theorem 2.3 (see [2]). *The commutative language $c(a_1 \dots a_n)$ has a test set of cardinality at most $5n$.*

Each test set of the commutative language $c(a_1 \cdots a_n)$ has cardinality at least $n-1$.

The commutative language $c(a_1^{k_1} \cdots a_n^{k_n})$ has a test set of cardinality at most $10n$.

We shall call the cardinality of the set $\Psi[L]$, that is the number of Parikh vectors contained in the language, the *Parikh size* of L. Since a commutative language is fully described by its Parikh image, it suggests itself to measure the size of its test set by its Parikh size. We therefore introduce the following notion. A set $T \subseteq L$ is called a *Parikh test set* of L if $c(T)$ is a test set of $c(L)$. Note that, given a commutative language, the minimal Parikh size of its test set corresponds exactly to the minimal size of its Parikh test set.

3. Classical *vs.* Parikh test sets

In this section we give several examples, which illustrate the theory of test sets for commutative languages, and the relationship between classical and Parikh test sets.

Example 3.1. This is the example given in [1] of a language that has only test sets of size $\Omega(n^2)$. The alphabet is

$$X_1 = \{a_1, \dots, a_n, b_1, \dots, b_n, c_1, \dots, c_n\}$$

with cardinality $3n$, and the language is

$$L_1 = \{a_i b_j c_i \mid i, j = 1, \dots, n\}.$$

Both the cardinality and the Parikh size of L_1 is equal to n^2.

It turns out that no proper subset of L_1 is its test set. To prove this, suppose that a set $T \subset L_1$ does not contain $a_n b_n c_n$. The following morphisms testify that T is not a test set of L_1:

$$\begin{aligned} g(a_n) &= a^2, & g(b_n) &= b, & g(c_n) &= a, \\ h(a_n) &= a, & h(b_n) &= b, & h(c_n) &= a^2, \end{aligned}$$

and

$$\begin{aligned} g(a_i) &= a, & g(b_i) &= a, & g(c_i) &= a, \\ h(a_i) &= a, & h(b_i) &= a, & h(c_i) &= a \end{aligned}$$

for $i = 1, \dots, n-1$. The morphisms agree on all words from L_1 with the only exception of $a_n b_n c_n$.

The same reasoning and the same morphisms show that also any Parikh test set has the Parikh size of full n^2. Note that g and h in fact agree on whole $c(L_1) \setminus \{a_n b_n c_n\}$.

Example 3.2. The language L_1 should be compared with the language

$$L_2 = \{a_i b_j \mid i, j = 1, \ldots, n\},$$

which also has the Parikh size n^2. In this case, however, the set

$$T_2 = \{a_1 b_i \mid i = 1, \ldots, n\} \cup \{a_i b_1 \mid i = 1, \ldots, n\}$$

is a Parikh test set of L_2, with the Parikh size $2n - 1$.

Example 3.3. The language

$$L_3 = c(\{a_1 \ldots a_n b_i \mid i = 1, \ldots, n\})$$

contains only n Parikh vectors, and the same holds for any of its Parikh test sets. On the other hand it can be shown that any classical test set has at least $\frac{n(n-4)}{4}$, because the classical test set has to contain many words with the same Parikh vector. Exactly this source of the size of the classical test set is ignored by Parikh test sets.

Example 3.4. For $n > 2$ denote

$$T = \{a_i b_i \mid i = 1, \ldots, n\}, \qquad S = \{a_i b_i a_j b_j \mid i, j = 1, \ldots, n,\ i < j\},$$

and consider the (non-commutative) language $L_4 = T \cup S$. Clearly T is a test set of L_4, while any Parikh test set R of L_4 has to contain a permutation of each word from S. Indeed, suppose for instance that the intersection of R with $c(a_1 b_1 a_2 b_2)$ is empty. Then morphisms g and h defined by

$$\begin{aligned} g(a_1) &= a^2, & g(b_1) &= a, \\ h(a_1) &= a, & h(b_1) &= a^2, \\ g(a_2) &= b, & g(b_2) &= b, \\ h(a_2) &= b, & h(b_2) &= b, \end{aligned}$$

and by

$$g(a_i) = g(b_i) = h(a_i) = h(b_i) = a$$

for $2 < i \leq n$, agree on $c(R)$, but do not agree on $c(L_4)$, for instance on $a_1 a_2 b_1 b_2$.

The previous example shows that T being test set of L does not imply that T is also Parikh test of L. It also shows that a test set can have smaller cardinality than any Parikh test set. If L is commutative, however, the implication trivially

holds, and Parikh test sets give a lower bound to the cardinality of classical test sets.

Theorem 2.3 moreover implies that if T is a Parikh test set of L, then L has a classical test set of cardinality at most

$$10 \sum_{v \in T} |\mathrm{alph}(v)|.$$

To see this, it is enough to consider the set

$$T' = \bigcup_{v \in T} T_v,$$

where T_v is a test set of $c(v)$, with $|T_v| \leq 10 \cdot |\mathrm{alph}(v)|$.

4. Coincidence graph and difference supports

In this section we define two basic concepts needed for characterization of Parikh test sets, and show some of their properties.

The basic description of a language L over an alphabet X is an undirected graph called *coincidence graph*, and denoted $G(L)$. It is defined as follows. The set of vertices of the graph is the alphabet X. The set of edges $E = E(L)$ is defined by saying that (a, b) is in E if and only if there is a word w in L such that ab is a factor of $c(w)$. The graph admits loops, since we do not require $a \neq b$. Therefore, the loop (a, a) is in E if and only if L contains a word with at least two occurrences of the letter a.

The sequence of vertices $x_0, \ldots, x_m$ is called *path* if $(x_0, x_1), \ldots, (x_{m-1}, x_m) \in E$. The length of such a path is m, that is the number of edges.

Let L be a language over $X = \{a_1, \ldots, a_n\}$. Given morphisms $g, h : X \to A^*$, define

$$D(g, h) = \{a_i \mid |g(a_i)| \neq |h(a_i)|\}.$$

A subset $D \subseteq X$ is called a *difference support* of L if there are morphisms g and h that agree lengthwise on L, and $D = D(g, h)$.

Difference supports have the following equivalent characterization, which allows to find all of them by means of linear algebra.

Lemma 4.1. *Let L be a language over $X = \{a_1, \ldots, a_n\}$ and B a Parikh basis of L. Then a subset D of X is difference support of L if and only if there is a vector $d = (d_1, \ldots, d_n) \in \mathbb{Q}^n$ such that*

$$D = \{a_i \mid d_i \neq 0\},$$

and

$$b \cdot d := \sum_{i=1}^{n} b_i \cdot d_i = 0$$

for each $b = (b_1, \ldots, b_n) \in \Psi[B]$.

Proof. Given $D = D(g,h)$ for morphisms g and h that agree lengthwise on L, put $d_i = |g(a_i)| - |h(a_i)|$. The length agreement implies that for each $v \in B$

$$\sum_{i=1}^{n} |v|_{a_i} \cdot |g(a_i)| = \sum_{i=1}^{n} |v|_{a_i} \cdot |h(a_i)|,$$

and therefore

$$\sum_{i=1}^{n} b_i \cdot d_i = 0.$$

Let, on the other hand, D be the set $\{a_i \mid d_i \neq 0\}$, where the vector $d = \{d_1, \ldots, d_n\}$ solves the system of equations

$$b \cdot d = 0, \quad b \in \Psi[B]. \tag{1}$$

We can choose d such that $d \in \mathbb{Z}^n$. Then D is equal to $D(g,h)$ for any g and h defined by

$$g(a_i) = a^{k_i}, \qquad\qquad h(a_i) = a^{\ell_i},$$

where k_i and ℓ_i are nonnegative integers such that $k_i - \ell_i = d_i$ for all $i = 1, \ldots, n$. Equalities (1) imply that g and h agree lengthwise on B. Since B is a Parikh basis, they agree lengthwise on the whole L. □

Denote by $\mathcal{B}$ a matrix the rows of which form basis of $\Psi[B]$. Denote the columns of $\mathcal{B}$ by $\mathcal{C} = \{c_1, \ldots, c_n\}$. By the previous Lemma, the difference supports correspond to subsets $\{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$ for which there are nonzero integers $d_{i_1}, \ldots, d_{i_k}$ satisfying

$$\sum_{j=1}^{k} d_{i_j} \cdot c_{i_j} = 0. \tag{2}$$

All difference supports can be found using following three lemmas. With a slight abuse of terminology we shall say, within those lemmas, that the difference supports are subsets of $\mathcal{C}$, instead of X. This should cause no confusion, since both X and $\mathcal{C}$ have cardinality n.

Lemma 4.2. *Let D be a minimal linearly dependent subset of $\mathcal{C}$. Then D is a minimal difference support.*

Proof. Since the set $D = \{c_{i_1}, \ldots, c_{i_k}\}$ is linearly dependent, there are coefficients $d_{i_1}, \ldots, d_{i_k}$ such that (2) holds. All d_{i_j} are nonzero, because all proper subsets of D are linearly independent, by hypothesis. For the same reason, no proper subset of D is a difference support. □

Lemma 4.3. *If D and D' are difference supports then also $D \cup D'$ is difference support.*

Proof. Let $d, d' \in \mathbb{Z}^n$ be vectors such that $\mathcal{B} \cdot d = \mathcal{B} \cdot d' = 0$, with nonzero entries corresponding to sets D and D', respectively. It is not difficult to see that for some integer α the vector $d'' = d + \alpha d'$ is a vector with nonzero entries corresponding to the set $D \cup D'$. □

Lemma 4.4. *Each difference support is a union of minimal difference supports.*

Proof. Let D be a difference support. We have to show that each vector $c \in D$ is an element of a minimal difference support that is a subset of D. Consider any subset S of $D \setminus \{c\}$ that is minimal with respect to the property that c is in the vector space spanned by S. It is easy to see that $S \cup \{c\}$ is a minimal difference support. □

We can summarize the previous three lemmas.

Lemma 4.5. *A subset $D = \{a_{i_1}, \ldots, a_{i_k}\}$ of X is difference support if and only if the set $C = \{c_{i_1}, \ldots, c_{i_k}\}$, which is a subset of $\mathcal{C}$, is a union of minimal linearly dependent subsets of $\mathcal{C}$.*

Given a set D of vertices of a graph G, we say that $x_0, \ldots, x_k$ is a *D-path* in G if for each $0 \le i \le k-1$ at least one of the vertices x_i, x_{i+1} is in D (*i.e.*, there is no edge in the path with both vertices out of D). If there is a D-path between a and b, we say that they are *D-connected*.

The most obvious case in which two morphisms g and h agree on the commutative closure of a word w is when $g(a) = h(a)$ for all $a \in \mathrm{alph}(w)$.

The following lemmas describe two other possibilities. Proofs may be found for example in [2].

Lemma 4.6. *Let g and h agree on $c(ab)$, and $g(a) \neq h(a)$. Suppose that $g(a)$ and $h(a)$ do not commute. Then there are unique nonempty words r, s, such that rs is primitive, and*

$$g(a) = (rs)^\ell r \qquad g(b) = (sr)^j s$$
$$h(a) = (rs)^{\ell+k} r \qquad h(b) = (sr)^{j-k} s$$

for some nonnegative integers ℓ and j, and k is a nonzero integer satisfying $-\ell \le k \le j$.

Lemma 4.7. *Let g and h agree on $c(w)$, where $|w| \ge 3$, $g(a)h(a)$ is nonempty for all $a \in \mathrm{alph}(w)$, and for at least one $a \in \mathrm{alph}(w)$ the inequality $g(a) \neq h(a)$ holds. Then all words in*

$$g[\mathrm{alph}(w)] \cup h[\mathrm{alph}(w)]$$

commute.

Fix a commutative language L over X and morphisms g and h that agree on L. In the rest of this section we shall also suppose that $g(x)h(x)$ is nonempty for each $x \in X$. In other words, we restrict the alphabet X to those letters that are not erased. Put $D = D(g, h)$.

For each primitive t define

$$P_t = \{a \mid g(a), h(a) \in t^*\},$$

and let

$$P = \bigcup_t P_t.$$

Denote

$$Z = X \setminus D = \{a \mid g(a) = h(a)\}.$$

Define also a symmetric relation $\sim$ on X by $a \sim b$ if and only if a and b satisfy conditions of Lemma 4.6. Denote

$$S = \{a \mid a \sim b \text{ for some } b\}.$$

From the definitions, and from Lemmas 4.6 and 4.7, we deduce the following claim.

Lemma 4.8. *If $(a, b) \in E$ then at least one of the following conditions is satisfied:*

(1) *$a, b \in Z$; or*
(2) *$a \sim b$; or*
(3) *there is a primitive word t such that $a, b \in P_t$.*

Moreover, $Z \subseteq P$, and S and P are disjoint.

Proof. Suppose that neither the first nor the second possibility holds. If ab is a factor of some $w \in L$, $|w| \geq 3$, then (3) follows from Lemma 4.7.

By Lemma 4.6, it remains that $g(a)$ and $h(a)$ commute, as well as $g(b)$ and $h(b)$, and $g(ab) = h(ab)$. Since $|g(a)| \neq |h(a)|$, it is easy to see that all four words commute. □

The relation $\sim$ has the following property.

Lemma 4.9.

(1) *If $x_0 \sim x_1 \sim \cdots \sim x_m$, then $x_0 \sim x_j$ for all $1 \leq j \leq m$ odd.*
(2) *If $x_0 \sim x_1 \sim \cdots \sim x_m \sim x_0$ then m is odd.*

Proof. For each i the words r_i and s_i such that $r_i s_i$ is primitive, and

$$g(x_i) = (r_i s_i)^{\ell_i} r_i, \qquad h(x_i) = (r_i s_i)^{\ell_i + k_i} r_i$$

are given uniquely by $g(x_i)^{-1} h(x_i)$ and $h(x_i) g(x_i)^{-1}$. Moreover,

$$g(x_i)^{-1} h(x_i) = g(x_{i+1}) h(x_{i+1})^{-1}.$$

This implies (1).

The second claim follows from the fact that $x_0 \sim x_0$ implies that r_0 and s_0 commute, a contradiction with the primitivity of the word $r_0 s_0$. □

We point out the following facts about paths.

Lemma 4.10.

(1) *If $x_0, \ldots, x_m$ is a D-path in $G(L)$ then either $x_i \in S$ for all $i = 0, \ldots, m$, or there is a primitive word t such that $x_i \in P_t$ for all $i = 0, \ldots, m$.*
(2) *If $x_0, \ldots, x_m, x_0$ is a path in $G(L)$ of odd length, and $x_0 \in D$, then x_0 is in P.*

Proof. (1) For each $0 \leq i < m$ either $x_i \notin Z$, or $x_{i+1} \notin Z$, since $x_0, \ldots, x_m$ is a D-path. The rest follows from Lemma 4.8.

(2) If $x_0 \in S$, then $x_0 \sim \cdots \sim x_m \sim x_0$, by Lemma 4.8, and we have a contradiction with Lemma 4.9(2). Also, x_0 is not in Z, since x_0 is an element of the difference support. □

5. Recognizing Parikh test sets

Our criterion of Parikh test sets is formulated for non-erasing pairs of morphisms. We say that the pair (g, h) is *non-erasing* if $g(x)h(x)$ is nonempty for all $x \in X$. Erased letters do not influence the agreement, and can be omitted from the alphabet. If we want to decide whether T is a Parikh test set of L for all morphisms, even erasing pairs, it is necessary to apply the criterion to all subsets of X and corresponding modifications of languages T and L obtained by erasing the missing letters.

The formulation "T is Parikh test set of L for non-erasing pairs of morphisms" therefore means that for each non-erasing pair of morphisms g and h we have $g \equiv_{c(T)} h$ if and only if $g \equiv_{c(L)} h$.

Theorem 5.1. *The language T is a Parikh test set of L for non-erasing pairs of morphisms if and only if the following conditions are satisfied*

(A) *T contains a Parikh basis of L.*
(B) *If there is a difference support D such that letters a and b are D-connected in $G(L)$, then a and b are D-connected in $G(T)$ too.*
(C) *Let $a \in D$ for a difference support D. If in $G(L)$ there is a cycle of odd length containing the letter a then also in $G(T)$ there is such a cycle.*

Proof.
1. We first prove that our criterion of the test set is sufficient. Let g, h be two morphisms that agree on $c(T)$. We want to show that they agree on $c(w)$ for each $w \in L$.

Pick a word $w \in L$. Since T contains a basis of L, the morphisms agree lengthwise on $c(L)$. If $g(a) = h(a)$ for each $a \in \mathrm{alph}(w)$, then there is nothing to prove. Let therefore $d \in \mathrm{alph}(w)$ be a letter for which $g(d) \neq h(d)$; therefore d is element of the difference support $D = D(g, h)$. By a length argument, there is at least one more letter $e \neq d$ in D.

Consider an arbitrary letter $a \in \mathrm{alph}(w)$, $a \neq d$. The letters a and d are D-connected in $G(L)$, therefore, by (B), there is a D-path

$$d = d_0, d_1, \ldots, d_m = a$$

in $G(T)$.

1.1. If $|w| \geq 3$, then there is a path of odd length from d to d in $G(L)$, for example d, e, b, d, where ebd is a factor of $c(w)$. Therefore, by (C), there is a path of odd length from d to d also in $G(T)$. By Lemma 4.10(2), $d \in P_t$ for some t, and thus also $a \in P_t$, by Lemma 4.10(1).

We have shown that if $|w| \geq 3$, then $\mathrm{alph}(w) \subseteq P_t$ and $g(w) = h(w)$.

1.2. Suppose now that $|w| = 2$, which means $c(w) = c(ed)$, and $a = d$.

If $d \in P_t$ for some t, then the path $d, d_1, \ldots, d_{m-1}, a$ guarantees that $a \in P_t$ too, by Lemma 4.10(1).

Suppose $d \in S$. If m is odd, then $a \sim d$, by Lemma 4.9(1), and we are through. If m is even then $G(L)$ contains the path

$$d, d_1, \ldots, d_{m-1}, a, d$$

of odd length. By (C) and Lemma 4.10 (2), the letter d is in P, a contradiction. This completes the "only if" part of the proof.

2. Let us approach the question whether the criterion is necessary.

Clearly, T has to contain a Parikh basis of L, otherwise it is easy to define periodic morphisms, which agree on T, but do not agree lengthwise on L.

2.1. Let now T be a subset of L that contains a Parikh basis of L, and D be a difference support, for which T does not satisfy the condition (B). This means that there are letters a and b which are D-connected in $G(L)$, but not in $G(T)$. We define morphisms g and h, which agree on $c(T)$ and do not agree on $c(L)$, in the following way. Let

$$g(a) = a^{i_a}, \qquad h(a) = a^{j_a},$$

and

$$g(x) = a^{i_x}, \qquad h(x) = a^{j_x}$$

for each $a \neq x \in X$ that is D-connected to a in $G(T)$. Similarly, let

$$g(y) = b^{k_y}, \qquad h(y) = b^{\ell_y},$$

for each $a \neq y \in X$ that is *not* D-connected to a in $G(T)$. Then $X = P_a \cup P_b$. Integers i_x, j_x, k_y and ℓ_y are chosen to make sure that $D = D(g, h)$.

Let

$$a = x_0, x_1, \ldots, x_k = b$$

be a D-path in $G(L)$. Since $a \in P_a$ and $b \in P_b$, there is some $0 \leq i \leq k-1$ such that $x_i \in P_a$ and $x_{i+1} \in P_b$. Moreover, at least one of the vertices x_i, x_{i+1} is in

the difference support. Since there is an edge between x_i and x_{i+1} in $G(L)$, there is a word w in L such that x_i and x_{i+1} are elements of alph(w). It is manifest that g and h do not agree on $c(w)$.

It remains to prove in this part that g and h agree on $c(T)$. Suppose, on the contrary, that g and h do not agree on $c(u)$ for some $u \in T$. Then, obviously, there are letters d and e in alph(u) such that $d \in P_a$ and $e \in P_b$, and at least one of the letters e and d is an element of D. By definition of g and h, there is a D-path between a and d. Then the same path extended by the edge (d, e) is a D-path between a and e, a contradiction with $e \in P_b$. Therefore (B) is a necessary condition.

2.2. We show that also (C) is necessary. Let again T be a subset of L containing a Parikh basis of L; suppose that for some difference support D of L there is a path of odd length from a letter $a \in D$ to itself in the graph $G(L)$, but there is no such cycle in $G(T)$. Let g' and h' be morphisms, which are length equivalent on L, such that $D = D(g', h')$.

Denote by Y all letters connected in $G(T)$ to a, including a itself. We claim that each word in T containing a letter from Y has length 2. Suppose the contrary, and let a word w in T contain a factor cde, where c, d, e are (not necessarily distinct) letters from Y. (Clearly, if one letter in alph(w) is in Y, then all are.) Let $a, z_1, \ldots, z_\ell, c$ be a path in $G(T)$. Then

$$a, z_1, \ldots, z_\ell, c, d, e, c, z_\ell, \ldots, a$$

has odd length, a contradiction.

The morphisms g' and h' are length equivalent on T, and a is in $D(g, h)$, that is, $|g(a)| \neq |h(a)|$; therefore, by the above claim, the set Y is a subset of D.

Define morphisms g and h in the following way. Let

$$g(x) = ab \qquad\qquad h(x) = a$$

for each $x \in X$, for which there is a D-path from a to x of *even* length in $G(T)$ (in particular this definition applies a itself); let

$$g(y) = a \qquad\qquad h(y) = ba$$

for each $y \in X$, for which there is a D-path from a to y of *odd* length in $G(T)$; and, finally, let

$$g(z) = c^{|g'(z)|} \qquad\qquad h(z) = c^{|h'(z)|}$$

for each $z \in X$ that is not D-connected to a in $G(T)$.

First, we have to assure that the definition is correct, namely that for each x in Y any two paths in $G(T)$ from a to x have length of the same parity. This holds, because, by assumption, any path $a, \ldots, x, \ldots a$, linking two paths from a to x, has even length.

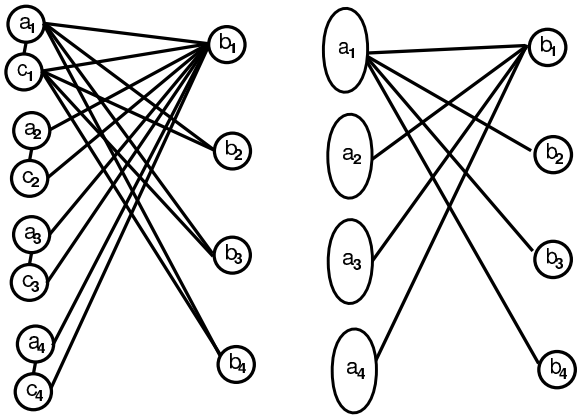

FIGURE 1. Coincidence graphs of sets T_1 and T_2, $n = 4$.

Since Y has been proved to be a subset of D, our construction implies that $D = D(g', h') = D(g, h)$.

It is straightforward to verify that g and h agree on $c(T)$. On the other hand, Lemma 4.10(2) implies that g and h do not agree on $c(L)$. □

The theorem claims that the Parikh test set has to preserve three properties of the tested language: Parikh basis, D-connectedness and cycles of odd length.

The first requirement is obvious. The second one is well illustrated by languages L_1 and L_2 from Section 3, which have Parikh test sets of significantly different size. The dividing line between them is the set of difference supports. The only nonempty difference support of L_2 is the whole X_2. Therefore, to satisfy the condition (B) it is enough to choose a test set T, for which $G(T)$ is connected. To be connected, and to be D-connected coincides. That is why the set T_2 is Parikh test set of L_2.

By contrast, a similar set

$$T_1 = \{a_1 b_i c_1 \mid i = 1, \ldots, n\} \cup \{a_i b_1 c_i \mid i = 1, \ldots, n\}$$

is not a Parikh test set of L_1. For instance, if g and h are as in Example 1, then we obtain

$$D = D(g, h) = \{a_n, c_n\},$$

and the letters a_n and b_n are D-connected in $G(L_1)$, but not in $G(T_1)$.

The importance of the third condition can be seen from the following example.

Example 5.2. Let

$$L_5 = \{a_1 a_2, a_2 a_3, a_3 a_4, a_4 a_5, a_5 a_6, a_6 a_1, a_1 a_3\}.$$

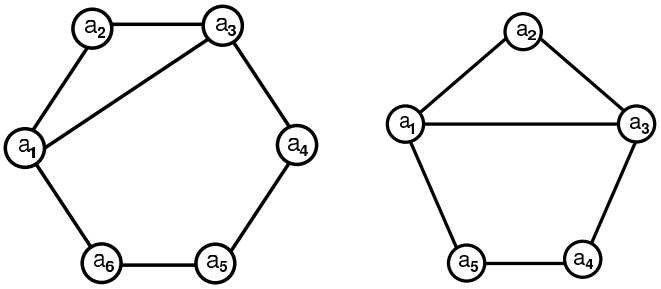

FIGURE 2. Coincidence graphs of sets L_5 and L'_5.

The language

$$T_5 = \{a_1a_2, a_2a_3, a_3a_4, a_4a_5, a_5a_6, a_6a_1\} = L_5 \setminus \{a_1a_3\}$$

is not a Parikh test set of L_5, since it misses the odd cycle a_1, a_2, a_3. Indeed, the morphisms g, h defined by

$$g(a_i) = \begin{cases} aba & i = 1,3,5 \\ b & i = 2,4,6 \end{cases}$$
$$h(a_i) = \begin{cases} a & i = 1,3,5 \\ bab & i = 2,4,6 \end{cases}$$

agree on T_5, but not on a_1a_3.

On the other hand, the language

$$L'_5 = \{a_1a_2, a_2a_3, a_3a_4, a_4a_5, a_5a_1, a_1a_3\}$$

has a Parikh test set

$$T'_5 = \{a_1a_2, a_2a_3, a_3a_4, a_4a_5, a_5a_1\},$$

since the cycle $a_1, \ldots, a_5$ has also odd length, which makes the edge a_1a_3 unnecessary.

REFERENCES

[1] Ismo Hakala and Juha Kortelainen, Polynomial size test sets for commutative languages. *RAIRO-Theor. Inf. Appl.* **31** (1997) 291–304.

[2] Štěpán Holub and Juha Kortelainen, Linear size test sets for certain commutative languages. *RAIRO-Theor. Inf. Appl.* **35** (2001) 453–475.

[3] Michel Latteux, Rational cones and commutations. In *Machines, languages, and complexity (Smolenice, 1988). Lect. Notes Comput. Sci.* **381** (1989) 37–54.