Theoretical Computer Science 33 (1984) 239-260
North-Holland

ON TEST SETS FOR CHECKING MORPHISM EQUIVALENCE ON LANGUAGES WITH FAIR DISTRIBUTION OF LETTERS

Yael MAON and Amiram YEHUDAI

Computer Science Department, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Communicated by A. Salomaa
Received October 1983
Revised March 1984

Abstract. A test set for a language L is a finite subset T of L with the property that each pair of morphisms that agrees on T also agrees on L. Some results concerning test sets for languages with fair distribution of letters are presented. The first result is that every D0L language with fair distribution of letters has a test set. The second result shows that every language L with fair distribution has a test set relative to morphisms g, h which have bounded balance on L. These results are generalizations of results of Culik II and Karhumäki (1983).

1. Introduction

In recent years a lot of research has been done to study the problems of morphism equivalence and existence of a test set for families of languages. A survey of the results in this area may be found in [3, 12].

Given a language L, a finite set T, $T \subseteq L$, is called a *test set* for L if for each two morphisms g, h we have $g(x) = h(x)$ for each x in T if and only if $g(x) = h(x)$ for each x in L. The notion of a test set is closely related to the problem of morphism equivalence. If, for a family of languages $\mathscr{L}$, each $L \in \mathscr{L}$ has effectively a test set (i.e., L has a test set and there exists an algorithm to find it), then the problem of morphism equivalence is decidable for $\mathscr{L}$, i.e., given $L \in \mathscr{L}$ and two morphisms g, h it is decidable whether $g(x) = h(x)$ for each x in L.

Ehrenfeucht conjectured [13, Problem 108] that for every language L there exists a test set. It is known that a test set cannot *effectively* exist for each context sensitive language, since the problem of morphism equivalence on these languages is undecidable [8]. However, the existence and effective existence of test sets have been shown for various families of languages. In [9] it has been shown that every language over a binary alphabet has (not effectively) a test set. A simpler proof to this result is given in [10], where the effective existence of a test set in the binary case is also shown for some families of languages. It is clear from arguments in [8] that a test set can be effectively constructed for each regular language, and this has been extended to context-free languages in [1]. Turning to L-systems, the existence of a test set is open for all families of languages between D0Ls and indexed languages.

Recently, two families of languages which have 'fair distribution of letters' have been shown to have a test set: the family of languages with 'bounded deviation' and 'fair distribution of letters', and the family of languages which are generated by 'positive' D0L systems [6]. For more results on the Ehrenfeucht Conjecture the reader is referred to the survey paper of Karhumäki [12].

In this paper we continue studying families of languages with 'fair distribution of letters', and generalize the two above mentioned results of Culik II and Karhumäki [6].

In Section 2 some definitions and notations are given. The concept of fair distribution of letters is introduced in Section 3. A language L has fair distribution (of letters) if there exists a $c>0$ such that in every substring v of L whose length is larger than c all the letters of Σ occur. Introducing this notion, the 'connection' between fair distribution and test sets is discussed. In Section 4 we present our results. The first result is that every D0L language with fair distribution has a test set. The second result shows that every language L with fair distribution has a test set relative to morphisms g, h which have bounded balance on L. The proofs of these results are discussed in Sections 5 and 6. Finally, in Section 7, some conclusions are given.

2. Preliminaries

In this section we give some definitions and notations. Some background material and additional definitions may be found in [15, 11].

A free monoid generated by a finite alphabet Σ is denoted by Σ^*. The elements of Σ^* are *words* or *strings* and its subsets are *languages*. The identity element of Σ^*, the empty word, is denoted by Ω, and $\Sigma^+ = \Sigma^* - \{\Omega\}$.

Throughout this paper let L be a language over Σ^*, where $\Sigma = \{a_1, a_2, \ldots, a_t\}$.

Let $w \in \Sigma^*$. The length of w is denoted by $|w|$, and the number of a_i's in w is denoted by $|w|_{a_i}$. The Parikh-mapping $\Psi: \Sigma^* \to N^t$ is defined by $\Psi(w) = (|w|_{a_1}, \ldots, |w|_{a_t})$. Consequently, the *Parikh vector* of a word w is denoted by $\Psi(w)$. The set of letters occurring in w is denoted by alph(w). A word w is *primitive* if the equation $w = z^n$ implies that $n = 1$ and $z = w$.

For $w \in \Sigma^*$, pref(w) denotes the set of all prefixes of w, and p-pref(w) the set of all prefixes of w whose length is less than the length of w. Similarly, suf(w) and p-suf(w) are defined with respect to the suffixes of w. For $L \subseteq \Sigma^*$, pref(L) = $\{\text{pref}(w) \mid w \in L\}$. We say that y is a *subword* of w if $w = w_1 y w_2$ for some words w_1 and w_2. The set of all subwords of words in L is denoted by sub(L).

The central notion in this paper is a morphism of a free monoid. Throughout the paper g and h denote morphisms from Σ^* to Δ^* (where Δ may be Σ) and f denotes a morphism from Σ^* to Σ^*. For a language L, $g(L) = \{g(x) \mid x \in L\}$. The size of a morphism g, denoted by $\|g\|$, is $\max\{|g(a)| \mid a \in \Sigma\}$. We say that g and h *agree* on L, in symbols $g =^L h$, if $g(w) = h(w)$ for all w in L. For a word w, the *balance* of

w with respect to g, h, in symbols $\beta_{g,h}(w)$, is defined by $\beta_{g,h}(w) = |g(w)| - |h(w)|$. We say that a pair (g, h) has *bounded balance* on L if there exists a constant k such that $|\beta_{g,h}(w)| \leq k$ for all $w \in \text{pref}(L)$. Otherwise, (g, h) has *unbounded balance* on L. (Note that this definition relates to the language pref(L) rather than to L.) For a language L, let $B(L) = \{(g, h) | (g, h)$ have bounded balance on $L\}$ and $\text{UB}(L) = \{(g, h) | (g, h)$ have unbounded balance on $L\}$. Moreover, let $H_l(L) = \{(g, h) | |g(x)| = |h(x)|$ for each x in $L\}$.

Given a language L and a set of pairs of morphisms D, we say that L has a *test set for* D if there exists a finite set T, $T \subseteq L$, such that for each pair $(g, h) \in D$ we have $g \equiv^T h$ if and only if $g \equiv^L h$. We say that L has a *test set* if L has a test set for the set of all pairs of morphisms (g, h). The *Ehrenfeucht Conjecture* states: Every language has a test set. We say that a family of languages $\mathscr{L}$ has *effectively* a test set if each L in $\mathscr{L}$ has a test set and there exists an algorithm which, given L in $\mathscr{L}$, finds its test set. A finite set V, $V \subseteq L$, is called a *length test set* for L if for each (g, h) we have $(g, h) \in H_l(V)$ if and only if $(g, h) \in H_l(L)$.

The notion of a D0L system is also needed. A *D0L system* G is a triple (Σ, f, x), where Σ is a finite alphabet, $f: \Sigma^* \to \Sigma^*$ a morphism and $x \in \Sigma^+$. The sequence of G, $E(G)$, is the sequence of words x, $f(x)$, $f^2(x), \ldots$. The language $L(G) = \{f^n(x) | n \geq 0\}$ is the *D0L language* which is generated by G. We say that $E(G)$ is *strictly monotonic* if $|f^{i+1}(x)| > |f^i(x)|$ for each $i \geq 0$. A *decomposition* of a D0L system $G = (\Sigma, f, x)$ is a set of D0L systems G_j, $0 \leq j < n_0$, defined by $G_j = (\Sigma, f^{n_0}, f^j(x))$. Notice that $\bigcup_{j=0}^{n_0-1} L(G_j) = L(G)$. For a string $v \in \Sigma^+$, the language L_v, with respect to a fixed D0L system $G = (\Sigma, f, x)$, is $L(G_v)$, where $G_v = (\Sigma, f, v)$.

The following remarks concerning test sets are relevant.

The notion of 'test set for D' where D is a set of pairs of morphisms, turns out to be useful when proving that a language L has a test set: if $D_1 \cup D_2$ equals the set of all pairs of morphisms and L has test sets for D_1 and for D_2, then L has a test set.

Dealing with length test set, one can verify that every language $L \subseteq \Sigma^*$ has a length test set. (A maximal set of words $w_1, \ldots, w_k \in L$, such that $\Psi(w_1), \ldots, \Psi(w_k)$ are linearly independent, is a length test set.)

Throughout the paper, dealing with existence of a test set for $L \subseteq \Sigma^*$, we assume that $\Sigma \subseteq \text{sub}(L)$.

3. Fair distribution of letters and test sets

In this section we present the notion of fair distribution of letters and illustrate its connection to test sets. The results concerning these concepts are given in the following sections.

The notion of fair distribution of letters was presented in [6]. A language $L \subseteq \Sigma^*$ has *fair distribution* (of letters) if there exists a $c > 0$ such that, for each $v \in \text{sub}(L)$, if $|v| > c$, then $\text{alph}(v) = \Sigma$.

As this paper deals with test sets for languages with fair distribution, we try to show the connection between these two concepts.

As a matter of fact, the notion of balance is the one which connects test sets and languages with fair distribution.

The balance is useful when dealing with morphism equivalence and test sets. Actually, when proving most (if not all) of the results concerning these problems, the notion of balance is crucial. In particular, some results only deal with pairs of morphisms which have bounded balance on a given language L. For example, it follows from [8, Theorem 2.1] that given a D0L language L and $(g, h) \in B(L)$, it is decidable whether $g \equiv^L h$. Note that for an arbitrary pair of morphisms the decidability of this problem is open.

On the other hand, the notion of fair distribution is related to bounded balance The balance measures the *difference* between $|g(w)|$ and $|h(w)|$ for words w. When L has fair distribution, it turns out that the *ratio* between $|g(w)|$ and $|h(w)|$ for $w \in \text{sub}(L)$ is bounded. This bound does not imply a bounded balance. Yet it distinguishes languages with fair distribution from arbitrary languages, as the property of 'bounded ratio' does not hold for arbitrary languages. The 'bounded ratio' property is shown in the following lemma, which appears in [6, proof of Theorem 6.1].

Lemma 3.1 (Culik II and Karhumäki [6]). *Let $L \subseteq \Sigma^*$ be a language with fair distribution, and let c be the constant of distribution. There exists a $k \geq 1$ which depends only on L such that the following is satisfied: for each $(g, h) \in H_I(L)$ and every $w \in \text{sub}(L)$ where $|w| \geq c$ we have $|g(w)| \leq k|h(w)|$ and $|h(w)| \leq k|g(w)|$.*

The fair distribution is necessary in this claim. Consider, for example, the following language L and morphisms $(g, h) \in H_I(L)$:

$$L = \{a^n b^n \mid n \geq 0\}, \quad g(a) = d, \quad g(b) = \Omega, \quad h(a) = \Omega, \quad h(b) = d.$$

One can verify that no constants c and k satisfy the requirements of Lemma 3.1.

A sketch of the proof of Lemma 3.1, which sheds some light on languages with fair distribution, is given below.

Proof of Lemma 3.1. The following claims, which are not difficult to verify, are needed to prove this lemma.

Claim 1. *Let $z \in \Sigma^*$ such that $\text{alph}(z) = \Sigma$. There exists a $k_1 \geq 1$, which depends only on z, such that the following is satisfied: for each g, h such that $|g(z)| = |h(z)|$, we have $k_1\|g\| \geq \|h\|$ and $k_1\|h\| \geq \|g\|$.*

Claim 2. *Let $L \subseteq \Sigma^*$ be a language with fair distribution. There exists a $0 < k_2 \leq 1$, which depends only on L, such that for each morphism h and for each $w \in \text{sub}(L)$ with $\text{alph}(w) = \Sigma$, the following is satisfied: $k_2|w| \cdot \|h\| \leq |h(w)|$.*

Turning to the proof of Lemma 3.1, let $z \in \Sigma^+$ such that $\text{alph}(z) = \Sigma$, and let k_1 and k_2 be the numbers which are guaranteed by Claims 1 and 2, respectively. Set k as k_1/k_2. To show that k satisfies the requirements of Lemma 3.1, let $(g, h) \in H_l(L)$ and $w \in \text{sub}(L)$ where $|w| \geq c$. We have

$$|g(w)| \leq |w| \cdot \|g\| \leq |w| \cdot k_1 \cdot \|h\| \leq \frac{1}{k_2} \cdot k_1 \cdot |h(w)|.$$

Similarly for $|h(w)|$. Since $k_1 \geq 1$ and $0 < k_2 \leq 1$, it follows that $k \geq 1$, which completes the proof. □

Note that there exists an $L \subseteq \Sigma^*$ with fair distribution and morphisms g, h which agree on L such that $(g, h) \in \text{UB}(L)$. A simple example is the following language and morphisms: $L = \{(ab^2)^n(ba^2)^n \mid n \geq 0\}$, and $g, h: \{a, b\}^* \to \{d\}^*$ such that $g(a) = \Omega$, $g(b) = d$, $h(a) = d$, $h(b) = \Omega$.

4. Main results

In this section we present two theorems which were proved by Culik II and Karhumäki [6]. Then we give our generalizations to these results. We also try to show the contribution of our results to the study of test sets for languages with fair distribution. The proofs of these results are discussed in the next two sections.

To present the results of [6] the following definition is needed A D0L system $G = (\Sigma, f, x)$ is *positive* if for each $b \in \Sigma$, $\text{alph}(f(b)) = \Sigma$. A D0L language L is a *positive language* if there exists a positive D0L system G such that $L = L(G)$. It is easy to verify that if L is a positive D0L language, then it has fair distribution.

The following definition is also needed. A language $L \subseteq \Sigma^*$ has a *bounded prefix deviation* if for each $(g, h) \in H_l(L)$ we have $(g, h) \in B(L)$. (Note that this definition is equivalent to the definition which is introduced in [6]. We do not give the original definition since it requires some additional concepts.)

The following theorems are proved by Culik II and Karhumäki in [6].

Theorem 4.1 ([6]). *Let $L \subseteq \Sigma^*$ be a language which has bounded prefix deviation and fair distribution. Then L has a test set.*

Theorem 4.2 ([6]). *Let $G = (\Sigma, f, x)$ be a positive D0L system. Then $L(G)$ has a test set.*

Theorem 4.2 is a result of the following theorems of [6].

Theorem 4.2.1 ([6]). *Let $G = (\Sigma, f, x)$ be a positive D0L system. Then $L(G)$ has a test set for $B(L(G))$.*

Theorem 4.2.2 ([6]). *Let $G = (\Sigma, f, x)$ be a positive D0L system. Then $L(G)$ has a test set for $\text{UB}(L(G))$.*

Theorems 4.2.1 and 4.1 have much in common, as both deal mainly with morphisms with bounded balance on a given language. Yet these results are incomparable: there are, of course, languages which have bounded prefix deviation and fair distribution, and which are not D0Ls; and one can find a positive D0L system G such that $L(G)$ does not have bounded prefix deviation (see [6, Example 5.1]).

Our first result is the following theorem.

Theorem 4.3. *A language $L \subseteq \Sigma^*$ which has fair distribution has a test set for $B(L)$.*

Theorem 4.3 generalizes Theorem 4.2.1, but it does not generalize Theorem 4.1. To obtain a theorem which generalizes both Theorems 4.1 and 4.2.1, one can add to the test set which is guaranteed by Theorem 4.3 a length test set (see Section 2), and get the following theorem.

Theorem 4.3'. *A language $L \subseteq \Sigma^*$ which has fair distribution has a test set T for $B(L)$ which satisfies the following: for each g, h, $g \equiv^T h$ implies that $(g, h) \in H_l(L)$.*

One can verify that Theorem 4.3' generalizes both Theorems 4.1 and 4.2.1.

The proof of Theorem 4.3, which is a generalization of the proofs of Theorems 4.1 and 4.2.1, is discussed in Section 5. The following lemma, which is useful in this proof, is given here, as we believe that it has importance of its own.

Lemma 4.4. *Let $L \subseteq \Sigma^*$. There exists a finite set U, $U \subseteq \text{pref}(L)$, such that for each $(g, h) \in B(L)$ we have*

$$\{\beta_{g,h}(w) \mid w \in \text{pref}(L)\} = \{\beta_{g,h}(w) \mid w \in U\}.$$

The above lemma shows the existence of a finite set U which 'presents' the balances on $\text{pref}(L)$ of *all* pairs of morphisms (g, h) in $B(L)$. Such a set may be useful when dealing with test sets for 'larger' families of languages in the 'bounded balance case'.

Theorem 4.3 may be viewed as a step towards proving that a 'large' family $\mathcal{L}$ of languages which have fair distribution has a test set. To show that such an $\mathcal{L}$ has a test set one has to prove that there exists a test set for $UB(L)$ for each $L \in \mathcal{L}$. The technique of dividing proofs concerning test sets and morphism equivalence into the 'bounded balance case' and 'unbounded balance case' is useful. This technique is explicitly used in [6, 9]. In many results it is used implicitly, as the 'unbounded balance case' is reduced to the 'bounded balance case' (see, for example, [4, 8]). Note that the 'unbounded balance case' turns out to be the more difficult one in these proofs.

Our second result, which deals with D0L languages, is the following theorem.

Theorem 4.5. *Let $G = (\Sigma, f, x)$ be a D0L system such that $L(G)$ has fair distribution. Then $L(G)$ has a test set.*

Theorem 4.5 shows a property of *languages* which causes a D0L language to have a test set, while Theorem 4.2 gives a property of D0L *systems* which causes the generated languages to have a test set. One can see that Theorem 4.5 is a strict generalization of Theorem 4.2. For example, consider the D0L system $G_0 = (\{a, b\}, f, a)$, where $f(a) = aba$ and $f(b) = b$. This system is not positive. Yet, Theorem 4.5 implies that $L(G_0)$ has a test set, as $L(G_0)$ has fair distribution. (It is easy to verify that $L(G_0) \subseteq (ab)^*a$.)

Theorem 4.5 generalizes Theorem 4.2 even when we turn to the families of languages, because the family of positive D0L languages is strictly contained in the family of D0L languages with fair distribution. To verify it, consider a D0L system $G = (\Sigma, f, x)$ such that $L(G)$ has fair distribution, and where the following conditions are satisfied: (i) $E(G)$ is strictly monotonic, and (ii) there exists an $i_0 \geq 0$ for which $|f^{i_0+1}(x)| < |\Sigma| \cdot |f^{i_0}(x)|$. (For example, let $G = (\{a, b, c\}, f, a)$ where $f(a) = bc$, $f(b) = abc$, and $f(c) = abc$.) The following arguments show that if $L = L(G')$ where $G' = (\Sigma, f', x')$ is a D0L system, then G' is not positive. Assume, for the sake of contradiction, that such a G' is positive. Since $f'(d) \neq \Omega$ for each $d \in \Sigma$, and $E(G)$ is strictly monotonic, it follows that $E(G) = E(G')$. Therefore, $|f'^{i_0+1}(x)| < |\Sigma| \cdot |f'^{i_0}(x)|$, which contradicts the assumption that G' is positive.

The above arguments fail if we allow decomposition: if G may be decomposed into a finite set of positive D0L systems, then we can use Theorem 4.2 and conclude that $L(G)$ has a test set. The following lemma implies that this technique is not applicable for the D0L languages with fair distribution.

Lemma 4.6. *There exists an infinite D0L language L_0 with fair distribution that contains no infinite positive D0L language.*

It follows from Lemma 4.6 that there are no finite t and positive D0L languages $L_1, L_2, \ldots, L_t$ satisfying $L_0 = \bigcup_{i=1}^{t} L_i$. In particular, if $L_0 = L(G_0)$, G_0 may not be decomposed into positive D0L systems.

The main idea in the proof of Theorem 4.5 is to generalize the concept of a positive D0L as follows. An *almost positive* D0L system is one in which the condition $\text{alph}(f(b)) = \Sigma$ must hold only for symbols b which generate infinite languages. It is shown that, given an D0L system G such that $L(G)$ has fair distribution, G may be decomposed into almost positive D0L systems. Then, generalizing the proof of Theorem 4.2 for positive D0Ls [6], it is shown that an almost positive D0L system has a test set, which proves Theorem 4.5. The proof of Theorem 4.5 is discussed in Section 6, along with a proof of Lemma 4.6.

Dealing with test sets for D0L languages, the following result of Culik II and Karhumäki is important.

Lemma 4.7 ([5]). *If a D0L language has a test set, then it has effectively a test set.*

This result implies that Theorems 4.2 and 4.5 may be strengthened to show effective existence of a test set. Thus, introducing the proofs concerning D0Ls, no effort is made to show effective existence of a test set.

5. Proof of Theorem 4.3

In this section we discuss the proof of Theorem 4.3. The proof is deeply based on the proofs of Theorems 4.1 and 4.2.1, which are due to Culik II and Karhumäki [6]. These proofs are similar, and are, in turn, a generalization of another proof which deals with test sets in the 'bounded balance case' for languages over a binary alphabet, which appears in [9]. We first sketch the proofs of Theorems 4.1 and 4.2.1 and then we turn to the proof of Theorem 4.3.

Analyzing the proofs of Theorems 4.1 and 4.2.1 in [6], it follows that there exist two properties which imply existence of a test set for a given language L: 'existence of representatives' and 'overlap'. These two concepts, together with some notations, are given below. (Note that these definitions do not appear explicitly in [6].)

For a set of words $X \subseteq \Sigma^*$, and morphisms g, h, let $\beta_{g,h}(X) = \{\beta_{g,h}(x) | x \in X\}$. Let M be a language and D a set of pairs of morphisms. We say that M *has representatives for* D if there exists a finite set U, $U \subseteq M$, such that for each $(g, h) \in D$ we have $\beta_{g,h}(U) = \beta_{g,h}(M)$. We say that M *has overlap for* D if there exists a constant N such that for each $uv \in \text{pref}(M)$ with $|v| \geq N$, the following holds: For any pair $(g, h) \in D$, we have $|h(v)| \geq |\beta_{g,h}(u)|$ and $|g(v)| \geq |\beta_{g,h}(u)|$.

Note that the property of overlap is deeply connected to fair distribution and to existence of representatives. This is illustrated in the following lemma, which may be considered as a restatement of [6, Claim 1 in the proof of Theorem 5.1].

Lemma 5.1 ([6]). *Let L be a language with fair distribution and such that* pref(L) *has representatives for $B(L)$. Then L has overlap for $B(L) \cap H_l(L)$.*

The following lemma of Culik II and Karhumäki [6], which deals with representatives, overlap and test sets, is crucial in the proofs of Theorems 4.1 and 4.2.1.

Lemma 5.2 ([6]). *Let $L \subseteq \Sigma^*$ and let D be a set of pairs of morphisms. Assume that the following is satisfied*: (i) *Each subset of* pref(L) *has representatives for D, and* (ii) *L has overlap for D. Then L has a test set for D.*

Note that if, instead of condition (i), it is known that each subset of pref(L) has representatives for D_1, where $D_1 \supseteq D$, then the lemma still holds (i.e., L has a test set for D).

The proof of Lemma 5.2 may be found in [6]. (Note that this lemma does not appear explicitly in [6], but it is proved when proving [6, Theorem 3.2].)

The usefulness of Lemma 5.2 in proving [6, Theorems 4.1 and 4.2.1] may be described as follows.

Dealing with a language L with bounded prefix deviation (see Theorem 4.1), it is shown that each subset of pref(L) has representatives for $H_I(L)$. Using the fair distribution, it is shown that such L has overlap for $H_I(L)$. Now, by Lemma 5.2, L has a test set for $H_I(L)$. Adding a length test set, it follows that L has a test set, which proves Theorem 4.1.

Turning to Theorem 4.2.1, let L be a D0L language. It is shown that a 'large enough' (but partial) set of subsets of pref(L) has representatives for $B(L) \cap H_I(L)$. This proof is based on a result of Culik II [2], which roughly shows that there exist a vector v and matrices $M_1, \ldots, M_t$, M such that

$$\psi(\text{pref}(L)) = \{v \cdot M_{i_1} \cdot M_{i_2} \cdot \ldots \cdot M_{i_k} \cdot M \mid k \geqslant 0,\ 1 \leqslant i_j \leqslant t\}.$$

In addition, results of Mandel and Simon [14], which deal with matrices, are used. Now, using the fair distribution, it is shown that a positive D0L language L has overlap for $B(L) \cap H_I(L)$. Appealing to Lemma 5.2 again and adding a length test set, it follows that L has a test set for $B(L)$, which proves Theorem 4.2.1.

Note that the above-mentioned proofs concerning existence of representatives are deeply based on properties of the families of languages in consideration.

Our result, which is crucial in proving Theorem 4.3, is the following.

Lemma 5.3. *For an arbitrary language L, each subset of* pref(L) *has representatives for $B(L)$.*

Notice that Lemma 5.3, together with Lemmas 5.1 and 5.2, implies Theorem 4.3, i.e., the existence of a test set for $B(L)$ for an arbitrary language L with fair distribution. To verify this, consider a language L and let $D_1 = B(L)$ and $D = B(L) \cap H_I(L)$. By the remark which appears after Lemma 5.2, it follows that L has a test set for $B(L) \cap H_I(L)$. Adding a length test set we achieve a test set for $B(L)$. Thus, in order to prove Theorem 4.3, it suffices to prove Lemma 5.3.

Lemma 5.3 is a corollary of the following lemma.

Lemma 5.4. *Let $M \subseteq \Sigma^*$. Then M has representatives for $\{(g, h) \mid \beta_{g,h}(M)$ is a finite set$\}$.*

To verify that Lemma 5.4 implies Lemma 5.3, let M be a subset of pref(L) and $(g, h) \in B(L)$. Then the set $\beta_{g,h}(M)$ is a finite set, which, by Lemma 5.4, implies Lemma 5.3.

Proof of Lemma 5.4. Assume that $\Sigma = \{a_1, \ldots, a_t\}$. For a pair of morphisms g, h, let $\eta_{g,h}$ be

$$(|g(a_1)| - |h(a_1)|, \ldots, |g(a_t)| - |h(a_t)|).$$

Let $A = \{(g_1, h_1), (g_2, h_2), \ldots, (g_l, h_l)\}$ be a finite set of pairs of morphisms which satisfies the following conditions: (i) for each i, $1 \leq i \leq l$, $\beta_{g_i,h_i}(M)$ is a finite set, and (ii) for each (g, h) such that $\beta_{g,h}(M)$ is a finite set, $\eta_{g,h}$ is linearly dependent on $\eta_{g_1,h_1}, \ldots, \eta_{g_l,h_l}$.

Such a set always exists.

For a word $w \in M$, let $\mathrm{vec}_A(w) = (\beta_{g_1,h_1}(x), \ldots, \beta_{g_l,h_l}(x))$. By condition (i), $\mathrm{vec}_A(M)$ is finite, where $\mathrm{vec}_A(M) = \{\mathrm{vec}_A(w) \mid w \in M\}$. Therefore, there exists a finite set U, $U \subseteq M$, such that $\mathrm{vec}_A(M) = \mathrm{vec}_A(U)$.

Claim. For each (g, h) such that $\beta_{g,h}(M)$ is a finite set, we have $\beta_{g,h}(U) = \beta_{g,h}(M)$.

Proof of the Claim. Let (g, h) be a pair of morphisms such that $\beta_{g,h}(M)$ is a finite set. The choice of A (condition (ii)) implies that there exist numbers $k_1, \ldots, k_l$ such that

$$\eta_{g,h} = k_1 \cdot \eta_{g_1,h_1} + \cdots + k_l \cdot \eta_{g_l,h_l}.$$

Therefore, for each $z \in \Sigma^*$,

$$\begin{aligned}\beta_{g,h}(z) &= \Psi(z) \cdot \eta_{g,h} = \Psi(z) \cdot [k_1 \cdot \eta_{g_1,h_1} + \cdots + k_l \cdot \eta_{g_l,h_l}] \\ &= k_1 \cdot \beta_{g_1,h_1}(z) + \cdots + k_l \cdot \beta_{g_l,h_l}(z) = (k_1, \ldots, k_l) \cdot \mathrm{vec}_A(z). \qquad (1)\end{aligned}$$

Now, let $w \in M$. The fact that $\mathrm{vec}_A(M) = \mathrm{vec}_A(U)$ implies that there exists a w' in U such that $\mathrm{vec}_A(w) = \mathrm{vec}_A(w')$. Therefore, using (1), we have

$$\beta_{g,h}(w) = (k_1, \ldots, k_l) \cdot \mathrm{vec}_A(w) = (k_1, \ldots, k_l) \cdot \mathrm{vec}_A(w') = \beta_{g,h}(w').$$

This implies that $\beta_{g,h}(M) \subseteq \beta_{g,h}(U)$, which completes the proof of the above claim and the proof of Lemma 5.4. □

6. Proof of Theorem 4.5

In this section we sketch the proof of Theorem 4.5. As it is deeply based on the proof of Theorem 4.2, which is due to Culik II and Karhumäki [6], we first sketch this proof, and then present the required generalizations. Note that the proof of Theorem 4.2 is complicated, and we believe that its sketch contributes to better understanding it.

Theorem 4.2 is a result of Theorems 4.2.1 and 4.2.2. The proof of Theorem 4.2.1 has been described in Section 5. Therefore, we sketch here only the proof of Theorem 4.2.2.

Sketch of the proof of Theorem 4.2.2

Let $G = (\Sigma, f, x)$ be a positive D0L system and $L = L(G)$. We have to show that L has a test set for UB(L).

Main idea

The main idea in this proof is 'periodicity'. One chooses a 'large enough' but finite set T, $T \subseteq L$, such that the following is satisfied. For each $(g, h) \in \mathrm{UB}(L)$, if $g \equiv^T h$, then there exists a primitive word p such that $g(L)$ and $h(L)$ are contained in $(\mathrm{sub}(p^*))^k$, where $k = |x|$ (ignoring a finite set of words of L).

Note that if it is known that $g(L)$ and $h(L)$ are contained in $\mathrm{sub}(p^*)$ (i.e., $k = 1$), then the task of proving that $g(w) = h(w)$ for each $w \in L$ (which means proving that T is a test set for L) becomes easier: For $w \in L$, it happens that $g(w) = p_1 p^i p_2$ and $h(w) = p_1' p^{i'} p_2'$ where $p_1, p_1' \in \text{p-suf}(p)$, $p_2, p_2' \in \text{p-pref}(p)$ and $i, i' \geq 0$. One only has to show that $p_1 = p_1'$, $i = i'$ and $p_2 = p_2'$. Similar information is useful when it is known that $g(L)$ and $h(L)$ are contained in $(\mathrm{sub}(p^*))^k$.

Structure of the proof and main claims

The main claims of the proof are given below. A discussion concerning the validity of these claims is given later. Note that the claims in [6] are presented differently.

The first step in the proof is to choose a 'large enough' number M_0 and let $T = \{x, f^1(x), \ldots, f^{M_0}(x)\}$. The set T is chosen such that it includes a length test set for L. It is claimed that T is a test set for UB(L). To prove this, let g, h be a fixed pair of morphisms of UB(L) such that $g \equiv^T h$. The following claims show that $g \equiv^L h$, which implies that T is a test set for UB(L).

Claim 6.1. *There exist words w and p, where p is primitive, and an integer $i < M_0$, such that the following is satisfied:*

(i) *for each $c, d \in \Sigma$ such that $cd \in \mathrm{sub}(\bigcup_{a \in \Sigma} L_a)$, $f^i(cd) \in \mathrm{sub}(w)$,*

(ii) *$g(w) \in \mathrm{sub}(p^*)$ and $h(w) \in \mathrm{sub}(p^*)$, and*

(iii) *$|g(f^i(a))| \geq |p|$ and $|h(f^i(a))| \geq |p|$ for each $a \in \Sigma$.*

From Claim 6.1 one derives the following.

Claim 6.2. *Let, for each $a \in \Sigma$, $\bar{L}_a = L_a - \{f^j(a) \mid j < i\}$. Then $g(\bar{L}_a) \subseteq \mathrm{sub}(p^*)$ and $h(\bar{L}_a) \subseteq \mathrm{sub}(p^*)$ for each $a \in \Sigma$.*

Recall that $L_a = L(G_a)$, where $G_a = (\Sigma, f, a)$.

From Claim 6.2 it follows that for $w = f^r(x) \in L$, where $r \geq i$, $g(w) \in (\mathrm{sub}(p^*))^k$ and $h(w) \in (\mathrm{sub}(p^*))^k$, where $k = |x|$. This information, the choice of T, and the assumption that $g \equiv^T h$ are sufficient to imply the following.

Claim 6.3. $g \equiv^L h$.

Relevant properties of positive D0Ls

The proofs of the above claims are based on the assumption that $L = L(G)$ where G is a positive D0L system. Relevant properties of positive D0Ls are listed below. Some intuition concerning these properties is given later.

Property 1—relatively small balances: Let $G = (\Sigma, f, x)$ be a positive D0L system, and r an integer. There exists an integer N such that the following is satisfied: for every pair of morphisms $(g, h) \in H_l(L(G))$, $n \geq N$ and $b \in \Sigma$, we have

$$|g(f^n(b))| \geq r \operatorname{MAX}\{\beta_{g,h}(w) \mid w \in \operatorname{pref}(f^m(x)), 0 \leq m < n\}.$$

Roughly, this property says that the balances are small with respect to the lengths of words in L.

Property 2—density of pairs of letters: A positive D0L system may be decomposed into positive D0L systems which have 'density of pairs of letters'. A D0L system $G = (\Sigma, f, x)$ has *density of pairs of letters*, if for each $c, d \in \Sigma$ such that $cd \in \bigcup_{a \in \Sigma} L_a$, it is the case that $cd \in \operatorname{sub}(f(b))$ for each $b \in \Sigma$.

Note that, proving that $L(G)$ has a test set where G is a positive D0L system, the first step (before choosing the number M_0) is to decompose G into systems with density of pairs of letters. Then a test set is found to each one of these languages, and the union of these test sets is a test set for $L(G)$.

Property 3—different values of balance: Let $G = (\Sigma, f, x)$ be a D0L system, and r an integer. There exists an $n_0 \geq 1$, which depends only on G and r, such that the following is satisfied: for each $(g, h) \in \operatorname{UB}(L(G))$ there exists an n, $r < n \leq r + n_0$, and a string $u \in \operatorname{pref}(f^n(x))$, such that the balance on u is 'new'. By 'new' we mean that

$$\beta_{g,h}(u) \notin \{\beta_{g,h}(w) \mid w \in \operatorname{pref}(f^m(x)), 0 \leq m < n\}.$$

Note that, given a D0L system $G = (\Sigma, f, x)$, an integer number r, and $(g, h) \in \operatorname{UB}(L(G))$, there exist n and $u \in \operatorname{pref}(f^n(x))$, such that $\beta_{g,h}(u)$ is 'new' in the above sense, where n depends on G, r and g, h. This is an immediate consequence of the assumption that $(g, h) \in \operatorname{UB}(L(G))$. However, Property 3 gives a range for this n, which is valid for all pairs of morphisms $(g, h) \in \operatorname{UB}(L(G))$.

Note that Property 3 holds for any D0L system (not only for positive systems).

On the proofs of Claims 6.1, 6.2 and 6.3

Sketch of the proof of Claim 6.1. Requirement (i) in Claim 6.1 is, roughly, a result of the density of pairs of letters in positive D0Ls (see Property 2).

To prove Claim 6.1(ii) one shows that there exist two words $v_1 w u_1$ and $v_2 w u_2$ in T (i.e., words with a common substring w) such that w is 'long enough' to guarantee Claim 6.1(i), and such that $\beta_{g,h}(v_1) \neq \beta_{g,h}(v_2)$. Since, by our assumption, $g =^T h$, the situation may be illustrated as in Fig. 1.

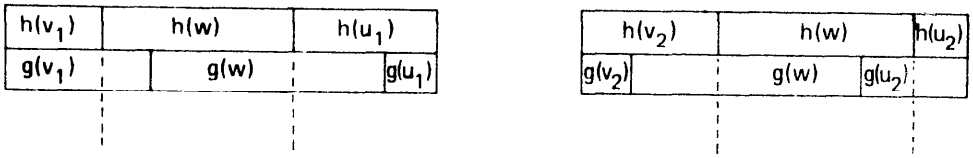

Fig. 1.

If the relations between the lengths of the strings are as in Fig. 1, then $h(w) = z_1 w_1 = w_2 z_2$, where $w_1 \in \text{pref}(g(w))$, $w_2 \in \text{suf}(g(w))$ and z_1, z_2 are two strings, as is illustrated in Fig. 2.

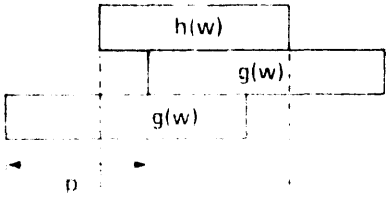

Fig. 2.

Denoting a prefix of $g(w)$ by p as in Fig. 2, one can show that $g(w) \in \text{sub}(p^*)$ and $h(w) \in \text{sub}(p^*)$. Moreover, $|p| = |\beta_{g,h}(v_1)| + |\beta_{g,h}(v_2)|$. Using some length arguments one can show that, even when the relations between the lengths of the above strings are different, there exists a 'long enough' string $\bar{w}$ (which is a substring of w) such that the relevant p satisfies $g(\bar{w}) \in \text{sub}(p^*)$, $h(\bar{w}) \in \text{sub}(p^*)$ and $|p| \leq |\beta_{g,h}(v_1)| + |\beta_{g,h}(v_2)|$. These length arguments show, among other properties, that in Fig. 2 the two occurrences of $g(w)$ are really laid out on each other (a property which implies the periodicity). To show that these strings are laid out on each other, one shows that $|g(w)| \geq |\beta_{g,h}(v_1)| + |\beta_{g,h}(v_2)|$, a length relation which is a consequence of the relatively smal' balances in positive D0Ls (see Property 1).

Note that we may assume that $g(w)$, $h(w) \in \text{sub}(p^*)$ where p is a primitive word, by considering $\bar{p}$ in the case that $p = \bar{p}^l$ for $l > 0$.

An important point in the above arguments is that the chosen words $v_1 w u_1$ and $v_2 w u_2$ where $\beta_{g,h}(v_1) \neq \beta_{g,h}(v_2)$ are included in T. The ability to define T such that it includes such words for *all* the pairs of morphisms $(g, h) \in \text{UB}(L)$ is guaranteed by Property 3. This property enables us, given $(g, h) \in \text{UB}(L)$, to consider a word $v_1 w u_1 = f^{m_1}(x)$, and to find a word $v_2 w u_2 = f^{m_2}(x)$ which satisfies: (i) $\beta_{g,h}(v_1) \neq \beta_{g,h}(v_2)$, and (ii) $|m_2 - m_1| \leq n_0$ for some n_0 which only depends on L and m_1. This situation enables us, given a language L, to define T such that the relevant words $v_1 w u_1$ and $v_2 w u_2$ may be found in T for *all* pairs of morphisms $(g, h) \in \text{UB}(L)$.

The above arguments sketch the proof of Claim 6.1(ii).

Turning to Claim 6.1(iii), the idea is to take i to be 'large enough'. Trying to choose i, the following problem arises. The number i is required to be less than M_0, where M_0 is chosen a priori and depends only on L; meanwhile, Claim 6.1(iii) presents a condition which involves both i and a pair of morphisms g, h. One may

overcome this problem by using, again, the property of relatively small balances of positive D0Ls (Property 1). One chooses M_0 'large' and i less than M_0 but such that $|f^i(a)|$ is 'big' for each $a \in \Sigma$. Using the property of relatively small balances, one can derive that $|g(f^i(a))| \geq |\beta_{g,h}(v_1)| + |\beta_{g,h}(v_2)|$, and similarly for h. But, as was noted before, $|\beta_{g,h}(v_1)| + |\beta_{g,h}(v_2)| \geq |p|$, which implies that Claim 6.1(iii) holds true. □

Sketch of the proof of Claim 6.2. To prove Claim 6.2, consider $y = f^m(x) \in \bar{L}_a$ for $a \in \Sigma$. Since $m \geq i$ we have $y = f^i(z)$ for some $z = z_1 z_2, \ldots, z_r$ where $z_j \in \Sigma$. Consider a pair $z_j z_{j+1}$ for some j, $1 \leq j \leq r-1$. Since $z_j z_{j+1} \in \text{sub}(\bigcup_{a \in \Sigma} L_a)$ one can apply Claim 6.1 and derive that $g(f^i(z_j z_{j+1})) \in \text{sub}(p^*)$. Hence $g(f^i(z_1 z_2)) \in \text{sub}(p^*)$, $g(f^i(z_2 z_3)) \in \text{sub}(p^*)$, and so on. To prove that $g(f^i(z_1 z_2, \ldots, z_r)) \in \text{sub}(p^*)$ (i.e., $g(y) \in \text{sub}(p^*)$), it is enough to show that, for each z_j, $g(f^i(z_j))$ has exactly one representation as a substring of p^* (i.e., if $g(f^i(z_j)) = p_1 p^l p_2 = p_1' p^{l'} p_2'$ for p_1, $p_1' \in \text{p-suf}(p)$ and p_2, $p_2' \in \text{p-pref}(p)$, then $p_1 = p_1'$, $l = l'$ and $p_2 = p_2'$). One can verify that if $x \in \text{sub}(p^*)$ for a primitive word p, and $|x| \geq |p|$, then x has exactly one representation as a substring of p^*. But, by Claim 6.1, $|g(f^i(z_j))| \geq |p|$ for each j and p is primitive, which completes the proof. □

Sketch of the proof of Claim 6.3. Since $i < M_0$ (see Claim 6.1), it is enough to show that $g(f^m(x)) = h(f^m(x))$ for $m \geq i$. Consider such m. By Claim 6.2 we have

$$g(f^m(x)) = g(f^m(a_1)) \ldots g(f^m(a_k)) = p_1 p^{i_1} p_1' p_2 p^{i_2} p_2' \ldots p_k p^{i_k} p_k',$$

where $p_j \in \text{p-suf}(p)$, $p_j' \in \text{p-pref}(p)$, $i_j \geq 0$ and $x = a_1 a_2 \ldots a_k$, $a_j \in \Sigma$. Similarly for $h(f^m(x))$. Our aim is to prove that $g(f^m(x)) = h(f^m(x))$. The main idea here is that there is a finite number of possible combinations for the strings $p_1, \ldots, p_k, p_1', \ldots, p_k'$ for g and for h over all the words in L, because $|p_j|, |p_j'| < |p|$. Using this property, one can choose a large enough but finite set $U \subseteq L$ such that all the possible combinations of $p_1, \ldots, p_k, p_1', \ldots, p_k'$ in both $g(f^m(x))$ and $h(f^m(x))$ occur in words $f^m(x)$ which are included in U. Refinement of this idea enables choosing T such that if $g \equiv^T h$, then $g(f^m(x)) = h(f^m(x))$ for each m.

The main problem in choosing a number M_0 where $T = \{x, f^1(x), \ldots, f^{M_0}(x)\}$ is that M_0 is chosen a priori, while the string p depends on g, h. One can overcome this problem using Claim 6.1(iii), which gives a connection between the number i which is less than M_0, and the length of p, a connection which holds for every relevant string p. □

Before turning to the properties of positive D0Ls, the following remark is in order. Consider Claim 6.1(i) and notice that $\text{sub}_2(\bigcup_{a \in \Sigma} L_a) \subset \text{sub}_2(L)$, where $\text{sub}_2(U) = \text{sub}(U) \cap \Sigma^2$ and the inclusion may be strict. If Claim 6.1(i) were true for each $cd \in \text{sub}_2(L)$, then we could derive that $g(L)$ and $h(L)$ are contained in $\text{sub}(p^*)$ (ignoring a finite set of words of L). Having Claim 6.1(i) only for $cd \in \text{sub}_2(\bigcup_{a \in \Sigma} L_a)$, implies that only $g(\bar{L}_a)$ and $h(\bar{L}_a)$ are contained in $\text{sub}(p^*)$ for each $a \in \Sigma$. Therefore,

one can only conclude that $g(L)$ and $h(L)$ are contained in $(\mathrm{sub}(p^*))^k$ where $k = |x|$ (ignoring a finite set of words of L).

On the properties of positive D0Ls

Property 1—*relatively small balances*: The intuition behind this property may be roughly explained as follows.

Consider, first, a language $L \subseteq \Sigma^*$ with fair distribution. Let g, h be a pair of morphisms, and assume that there exists a $z \in \Sigma^+$ such that $\beta_{g,h}(z) = 0$ (i.e., $|g(z)| = |h(z)|$). The fair distribution implies that there exists a $c' > 0$ such that for every $v \in \mathrm{sub}(L)$, if $|v| \geq c'$, then $\psi(v) \geq \psi(z)$. Let $w \in \mathrm{pref}(L)$, $|w| = c'l + t$, where $l \geq 0$ and $0 \leq t < c'$. The choice of c' implies that $\psi(w) = l\psi(z) + \psi(u)$ for some string u. Since $\beta_{g,h}(z) = 0$, $\beta_{g,h}(w) = \beta_{g,h}(u)$. This shows that if w is 'long', then there are many letters in w on which the total balance is zero. This, of course, causes the balance to be small.

In order to prove Property 1 for $L(G)$, where $G = (\Sigma, f, x)$ is a positive D0L system, one has to refine these arguments. This refinement is a consequence of the positiveness. Considering $f(b)$ for $b \in \Sigma$, all the letters of Σ occur in it, and, for each letter $a \in \mathrm{sub}(f(b))$, all the letters of Σ occur in $f(a)$, and so on. One can show that this 'rapid growth' implies that, for $w \in \mathrm{pref}(f^n(x))$, $\psi(w) = \psi(w') + \psi(w'')$, where $\beta_{g,h}(w') = 0$ for each $(g, h) \in H_l(L)$, and where $|w''|$ is 'small'. Actually, it turns out that

$$|w''| / |f^n(x)| \xrightarrow{n \to \infty} 0.$$

This fact shows that, for $(g, h) \in H_l(L)$, $\beta_{g,h}(w)$ is 'small' with respect to $|f^n(x)|$, which enables to prove Property 1.

Property 2—*density of pairs of letters*: This property is a fairly easy consequence of the definition of a positive D0L system.

Property 3—*different values of balance*: This result, which relates to an arbitrary D0L language $L(G)$, is combinatorial in nature, and uses the fact that $\mathrm{pref}(L(G)) = \tau(L(G'))$ for a DT0L G' and a morphism τ, a result of Culik II [2].

On the proof of Theorem 4.5

Turning to Theorem 4.5 we have to prove that a D0L language L with fair distribution has a test set. The existence of a test set for $B(L)$ is a result of Theorem 4.3. (Actually, it is also a result of the proof in [6] of Theorem 4.2.1.) Therefore, it suffices to prove that L has a test set for $\mathrm{UB}(L)$.

The following characterization of D0Ls with fair distribution, which appears in [6, Lemma 4.1], motivates our proof in the unbounded case.

Lemma 6.4 ([6]). *Let $G = (\Sigma, f, x)$ be a D0L system such that $L(G)$ is infinite. Let $\Sigma_f = \{a \in \Sigma \mid L_a \text{ is finite}\}$ and $\Sigma_i = \Sigma - \Sigma_f$. $L(G)$ has fair distribution if and only if the following two conditions are satisfied:*

(i) *there exists an integer* n_0 *such that, for every* $a \in \Sigma_i$, $\text{alph}(f^n(a)) = \Sigma$ *for* $n \geq n_0$, *and*

(ii) *the languages* $\Sigma_f^* \cap \text{pref}(L_a)$ *and* $\Sigma_f^* \cap \text{suf}(L_a)$ *are finite for every* $a \in \Sigma$.

Adapting the partition of Σ into Σ_i and Σ_f, we define the notion of 'almost positiveness'. A D0L system $G = (\Sigma, f, x)$ is said to be *almost positive* if the following conditions are satisfied:

(i) $L(G)$ has fair distribution,

(ii) for each $a \in \Sigma_i$, $\text{alph}(f(a)) = \Sigma$, and

(iii) $E(G)$ is strictly monotonic.

A D0L language L is an *almost positive language* if there exists an almost positive D0L system G such that $L = L(G)$.

Notice that a positive D0L system $G = (\Sigma, f, x)$ is an almost positive D0L system where $\Sigma_f = \phi$ (unless $|\Sigma| = 1$ and $L(G)$ is finite, in which case monotonicity is not satisfied).

The following observation is crucial in the proof of Theorem 4.5.

Lemma 6.5. *Let* G *be a D0L system such that* $L(G)$ *is an infinite language with fair distribution. Then* G *may be decomposed into a finite set of almost positive D0L systems* $G_1, \ldots, G_n$ *such that* $\bigcup_{i=1}^{n} L(G_i) = L - V$ *for some finite language* V.

Corollary 6.5.1. *Proving Theorem* 4.5 *it is enough to prove that every almost positive D0L language has a test set.*

Proof of Lemma 6.5. It is easy to verify that, given a D0L system $G' = (\Sigma', f', x')$ which satisfies conditions (i) and (ii) of the definition of an almost positive system, G' may be 'decomposed' into almost positive D0L systems $G'_1, \ldots, G'_m$ such that $\bigcup_{i=1}^{m} L(G_i) = L(G') - V'$, where V' is a finite set of words. Hence it suffices to prove that the D0L system G of Lemma 6.5 may be decomposed into D0L systems which satisfy conditions (i) and (ii) of the definition of almost positive. Now, let $G = (\Sigma, f, x)$, let n_0 be the number which is guaranteed by Lemma 6.4, and consider the D0L systems $G_j = (\Sigma, f^{n_0}, f^j(x))$ where $0 \leq j < n_0$. It is easy to verify that these systems satisfy the above-mentioned conditions (i) and (ii), which completes the proof. □

By Corollary 6.5.1, and since the existence of a test set for $B(L)$ is guaranteed by Theorem 4.3, the following theorem implies Theorem 4.5.

Theorem 6.6. *Let* $G = (\Sigma, f, x)$ *be an almost positive D0L system. Then* $L(G)$ *has a test set for* $\text{UB}(L)$.

The proof of Theorem 6.6 is based on the proof of Theorem 4.2.2. Intuitively, arguments which are similar to those of the proof of Theorem 4.2.2 are useful, as almost positive D0L systems are 'similar' to positive systems. One can say that the set Σ_i is the one which 'determines the nature of a D0L language $L(G)$' when

dealing with test sets. This and the fact that the letters of Σ_i satisfy the requirement of positiveness (i.e., for $a \in \Sigma_i$, alph$(f(a)) = \Sigma$), imply that almost positive D0Ls are 'similar' to positive D0Ls.

However, trying to generalize the proof of Theorem 4.2.2 to deal with almost positive D0Ls, some problems arise which require some modifications in the proof. We first give the properties of almost positive D0Ls and the claims which prove Theorem 6.6. Then we discuss the changes that have been made with respect to the proof of Theorem 4.2.2, and discuss the validity of the 'new' properties and claims.

Structure of the proof of Theorem 6.6 and main claims

The following definition is needed. Let $G = (\Sigma, f, x)$ be a D0L system. The *set of blocks of* L, BL(L), is $\{z \in \text{sub}(L) \mid z \in \Sigma_f^* \Sigma_i \Sigma_f^*\}$, i.e., the substrings of L in which a letter of Σ_i occurs exactly once. Notice that if $L(G)$ has fair distribution, then BL(L) is a finite set.

Turning to the proof of Theorem 6.6, let $G = (\Sigma, f, x)$ be an almost positive D0L system, $L = L(G)$. The first step in the proof is, again, to choose M_0, and let $T = \{x, f^1(x), \ldots, f^{M_0}(x)\}$. To prove that T is a test set for UB(L), let $(g, h) \in \text{UB}(L)$ such that $g \equiv^T h$.

Claim 6.1'. *There exist words w and p, where p is primitive, and an integer $i < M_0$, such that the following is satisfied:*

(i) *for each $\alpha, \beta \in \text{BL}(L)$ such that $\alpha\beta \in \text{sub}(\bigcup_{a \in \Sigma_i} L_a)$, $f^i(\alpha\beta) \in \text{sub}(w)$,*

(ii) *$g(w) \in \text{sub}(p^*)$ and $h(w) \in \text{sub}(p^*)$, and*

(iii) *$|g(f^i(\alpha))| \geq |p|$ and $|h(f^i(\alpha))| \geq |p|$ for each $\alpha \in \text{BL}(L)$.*

Claim 6.2'. *For each $a \in \Sigma_i$, $g(\bar{L}_a) \subseteq \text{sub}(p^*)$ and $h(\bar{L}_a) \subseteq \text{sub}(p^*)$, where $\bar{L}_a = L_a - \{f^j(a) \mid j < i\}$.*

Claim 6.3'. $g \equiv^L h$.

Relevant properties of almost positive D0Ls

We claim that the following properties hold true for almost positive D0Ls. (A sketch of the proofs is given later.)

Property 3 holds true for almost positive D0Ls. (Actually, it holds for each D0L language.)

Let *Property* 1' be similar to Property 1, the only difference is that letters b in Σ_i are considered, instead of letters b in Σ. Property 1' holds true for almost positive D0Ls.

Instead of Property 2, the following *Property* 2' holds true: An almost positive D0L system may be decomposed into almost positive D0L systems which have 'density of pairs of blocks for Σ_i'. A D0L system $G = (\Sigma, f, x)$ has density of pairs of blocks for Σ_i, if for each $\alpha, \beta \in \text{BL}(L)$ such that $\alpha\beta \in \bigcup_{a \in \Sigma_i} L_a$, it is the case that $\alpha\beta \in \text{sub}(f(b))$ for each $b \in \Sigma_i$.

Note that the density of blocks generalizes, in some sense, the density of letters which is introduced in Property 2. Yet, Properties 1′ and 2′ are weaker than Properties 1 and 2 as they deal only with Σ_i, and no information is given concerning Σ_f.

The differences between the proofs of Theorem 4.2.2 and Theorem 6.6

The main problems in generalizing the proof of Theorem 4.2.2 is that Properties 1 and 2 do not hold true for each almost positive D0L language. Only weakened versions of these properties, Properties 1′ and 2′, which deal with Σ_i and ignore Σ_f, are satisfied.

The fact that Property 1 does not hold true motivates us to consider the blocks $BL(L)$ instead of letters. This technique solves the 'length problem': The estimation that $|g(f^m(b))|$ is 'larger than the balances' for $b \in \Sigma_i$, which appears in Property 1′ (and which does not hold for b in Σ_f), holds true for blocks, as in each block of $BL(L)$ a letter of Σ_i occurs.

The fact that Property 2 is not satisfied implies the weak version of Claim 6.1′(i): only blocks α, β in $\mathrm{sub}(\bigcup_{a \in \Sigma_i} L_a)$ are considered, and substrings of $\bigcup_{a \in \Sigma_f} L_a$ are ignored. This, in turn, implies the weak version of Claim 6.2′: languages $\bar{L}_a$ are considered only for $a \in \Sigma_i$. The 'lack of information' about $\bar{L}_a$ for $a \in \Sigma_f$ complicates the proof of Claim 6.3′, and, comparing its proof to the proof of Claim 6.3, some additional arguments are needed.

Validity of the properties and claims

Consider, first, Property 2′. It deals with density of pairs of blocks instead of letters, but it turns out that the same proof shows density of blocks. The almost positiveness causes this property to deal only with Σ_i, but, turning to Σ_i, the proof is exactly as the proof of Property 2.

Having density of blocks, it turns out that the proofs of Claims 6.1′ and 6.2′ are similar to those of Claims 6.1 and 6.2.

Turning to Claim 6.3′, the 'lack of information' concerning $\bar{L}_a$ for $a \in \Sigma_f$ complicates its proof. To see this, recall the proof of Claim 6.3. In this proof, letting x be $a_1 a_2 \dots a_k$, the fact that $g(\bar{L}_{a_j})$, $h(\bar{L}_{a_j}) \subseteq \mathrm{sub}(p^*)$ for $1 \le j \le k$ was crucial in the proof. Now, if $x = b_1 a_1 \dots a_k b_{k+1}$ for $a_j \in \Sigma_i$ and $b_j \in \Sigma_f^*$, then $g(\bar{L}_{a_j})$, $h(\bar{L}_{a_j}) \subseteq \mathrm{sub}(p^*)$ for each j, but this is not true for $\bar{L}_{b_1}, \bar{L}_{b_2}, \dots, \bar{L}_{b_{k+1}}$, where $\bar{L}_{b_j} = L(G_{b_j}) = \{f^m(b_j) \mid m < i\}$, $G_{b_j} = (\Sigma, f, b_j)$. The main idea in solving this problem is that the languages $L_{b_1}, \dots, L_{b_{k+1}}$ are finite languages (as $b_j \in \Sigma_f^*$ for each j). This implies that there exists a finite set W such that, for each $y \in L$, $g(y)$ equals $w_1 p_1 p'^{i_1} p'_1 w_2 p_2 p'^{i_2} p'_2 \dots w_{k+1}$, where $p_j \in \mathrm{p\text{-}suf}(p)$, $p'_j \in \mathrm{p\text{-}pref}(p)$, $i_j \ge 0$ and $w_j \in W \cup \{\Omega\}$. Similarly for $h(y)$. Recall that the arguments in the proof of Claim 6.3 are based on the fact that the number of different strings $p_1, \dots, p_k, p'_1, \dots, p'_k$ over all the words in L is finite. This finiteness implies that a finite number of checks suffices to guarantee that $g =^l h$, and one has to include these checks in T. Turning to Claim 6.3′, we use also the finiteness of W and take a larger but still finite set T which

includes all the necessary checks. Arguments which are similar to those of Claim 6.3 show that if $g \equiv^T h$, then $g \equiv^L h$, which completes the proof.

The above discussion 'explains' Claims 6.1', 6.2', 6.3' and Property 2'. Note that Property 3 deals with arbitrary D0Ls. So it is only left to sketch the proof of Property 1' for almost positive D0L languages.

Property 1' of relatively small balances only deals with Σ_i. It turns out that, limiting ourselves to Σ_i, the proof of Property 2' for almost positive D0Ls is similar to the proof of Property 2. The modifications which are needed in proving Property 2' are that some length arguments must ignore the letters of Σ_f. More precisely, we consider $|w|_{\Sigma_i}$ instead of $|w|$, where $|w|_{\Sigma_i}$ is the number of occurrences of letters of Σ_i in w. However, given an almost positive D0L system $G = (\Sigma, f, x)$, the fair distribution of $L(G)$ implies the existence of a constant $l_G \geq 1$ such that $l_G|w|_{\Sigma_i} \geq |w|$ for each $w \in \text{sub}(L(G))$. Therefore, $(1/l_G)|w| \leq |w|_{\Sigma_i} \leq |w|$, so that length arguments which ignore Σ_f are enough to derive Property 1'.

This completes our discussion concerning the proofs of Theorem 6.6 and Theorem 4.5.

To complete the discussion about D0Ls, Lemma 4.6 needs to be proved.

Proof of Lemma 4.6. Let $G_0 = (\{a, b\}, f, a)$, where $f(a) = aba$ and $f(b) = b$, and let $L_0 = L(G_0)$. Then $L_0 = \{(ab)^{2^n-1}a \mid n \geq 0\}$ is an infinite D0L language with fair distribution. We claim that L_0 contains no infinite positive D0L language.

Assume, for the sake of contradiction, that $L' \subseteq L_0$ is an infinite language which is generated by a positive D0L system $G' = (\{a, b\}, f', x')$. The positiveness of G' and the fact that $L' \subseteq (ab)^*a$ implies that $E(G')$ is strictly monotonic ($|f'(a)|, |f'(b)| \geq 2$). In addition, we claim that $f'(a) = (ab)^k a$ and $f'(b) = (ba)^l b$ for some $k, l \geq 0$. To verify this, consider first $f'(a)$. Since each word in $E(G')$ begins and ends with the letter a, and since $L' \subseteq (ab)^*a$, it follows that $f'(a) = (ab)^k a$ for $k \geq 0$ or $f'(a) = \Omega$. Since G' is positive, $f'(a) = (ab)^k a$. Now, the fact that $L' \subseteq (ab)^*a$ implies that $f'(b) = (ba)^l b$ for some $l \geq 0$.

The following claim proves that $f'(b) = b$, which contradicts the assumption that G' is positive.

Claim. $l = 0$.

Proof. Let $E(G') = w_0, w_1, \ldots, w_n, \ldots$. Since $E(G')$ and $E(G_0)$ are strictly monotonic, it follows that for each $i \geq 0$ there exists an $n_i \geq 1$ such that $w_{i+1} = f^{n_i}(w_i)$, where f is the morphism of G_0.

Let $\Sigma = \{a, b\}$. Since $f'(a) = (ab)^k a$ and $f'(b) = (ba)^l b$, it follows that the matrix which is induced by G' is

$$\begin{pmatrix} k+1 & k \\ l & l+1 \end{pmatrix},$$

while the matrix which is induced by G_0 is

$$\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}.$$

Therefore, the following equation holds for each $i \geq 0$:

$$(|w_i|_a, |w_i|_b)\begin{pmatrix} k+1 & k \\ l & l+1 \end{pmatrix} = (|w_i|_a, |w_i|_b)\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}^{n_i}.$$

One can verify that

$$\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 2^n & 2^n - 1 \\ 0 & 1 \end{pmatrix}.$$

Since, for each i, $|w|_b = |w|_a - 1$, we can denote $|w_i|_a$ by m_i and achieve the following:

$$(m_i, m_i - 1)\left[\begin{pmatrix} k+1 & k \\ l & l+1 \end{pmatrix} - \begin{pmatrix} 2^{n_i} & 2^{n_i} - 1 \\ 0 & 1 \end{pmatrix}\right] = (0, 0).$$

Therefore,

$$(m_i, m_i - 1)\begin{pmatrix} k+1-2^{n_i} & k+1-2^{n_i} \\ l & l \end{pmatrix} = (0, 0),$$

which implies that

$$m_i(k+1-2^{n_i}) + (m_i - 1)l = 0.$$

We may assume that $m_0 > 1$. Therefore, for each $i \geq 0$,

$$2^{n_i} = l(m_i - 1)/m_i + k + 1. \tag{2}$$

This equality must be satisfied for fixed numbers $k, l \in N \cup \{0\}$, and an infinite sequence of pairs (m_i, n_i), where $m_i \to \infty$.

Now, by (2) we have $2^{n_i} \leq l + k + 1$, which implies that $\{n_i \mid i \geq 0\}$ is a finite set. Since l and k are fixed, and $\{m_i/(m_i - 1) \mid i \geq 0\}$ is an infinite set (as $m_i \to \infty$), the only possibility to satisfy (2) is that $l = 0$ (and $k + 1 = 2^{n_i}$ for each i), which completes the proof of the claim and of Lemma 4.6. □

7. Conclusions and open problems

The results of this paper may be viewed as another step towards proving the Ehrenfeucht Conjecture: the existence of test sets for languages. However, we regard these results as another step in the study of test sets for languages with fair distribution. We believe that this study may be fruitful, and that the notion of fair distribution is closely related to test sets.

By Theorem 4.3, given a family of languages $\mathcal{L}$ with fair distribution, if one proves that each $L \in \mathcal{L}$ has a test set for $UB(L)$, then it follows that L has a test set.

Turning to Theorem 4.5, its proof is based on the properties of D0Ls with fair distribution. Trying to isolate these properties from the claims for existence of a test set, the following may be derived.

Claim 7.1. *Let* $L \subseteq \Sigma^*$, $\bar{B}(L) \subseteq \mathrm{sub}(L)$ *and let* g, h *be a pair of morphisms. Assume that the following is satisfied:*

(1) *For each* $x \in L$, $x = x_1 x_2 \ldots x_k$ *for* $x_j \in \bar{B}(L)$.

(2) *There exist words* w *and* p, *where* p *is primitive, such that*: (i) *for each* α, $\beta \in \bar{B}(L)$ *such that* $\alpha\beta \in \mathrm{sub}(L)$ *we have* $\alpha\beta \in \mathrm{sub}(w)$, (ii) $g(w) \in \mathrm{sub}(p^*)$ *and* $h(w) \in \mathrm{sub}(p^*)$, *and* (iii) $|g(\alpha)| \geq |p|$ *and* $|h(\alpha)| \geq |p|$ *for each* $\alpha \in \bar{B}(L)$.

Then $h(L) \subseteq \mathrm{sub}(p^*)$ *and* $g(L) \subseteq \mathrm{sub}(p^*)$.

Note that this technique to show periodicity is taken from [6].

Showing 'periodicity' may be useful for finding a test set for L, as in Theorems 4.2.2 and 4.5. Note that a situation of 'periodicity' in some sense appears again and again in the 'unbounded balance case' of proofs of morphism equivalence and test sets (see, in addition to [6], the proofs in [8, 9]).

In [6] simple D0L systems are discussed. A D0L system $G = (\Sigma, f, x)$ is simple if, for each b, $c \in \Sigma$, b is generated from c in a number of steps (i.e., $b \in \mathrm{sub}(f^n(c))$ for some n). It is shown that, given a simple D0L system $G = (\Sigma, f, x)$ where $|x| = 1$, it may be decomposed into positive D0L systems $G_1, \ldots, G_n$, and hence $L(G)$ has a test set. Notice that $L(G) = \bigcup_{i=1}^{n} L(G_i)$, i.e., $L(G)$ is a union of a finite set of D0Ls with fair distribution.

The notion of a simple D0L system is a generalization of the notion of a positive D0L system. This leads to a conjecture of [6] that the technique of the proof of Theorem 4.2 may be useful in showing existence of test sets for simple D0Ls.

Notice that, given a simple D0L system $G = (\Sigma, f, x)$, $\Sigma_f = \Phi$ (like in a positive system), but $L(G)$ does not necessarily have fair distribution. For example, consider $G_2 = (\{a, b\}, f, ab)$ where $f(a) = bb$ and $f(b) = aa$. The language $L(G_2)$ does not have fair distribution. Moreover, one can verify that there exist no finite t and languages $L_1, \ldots, L_t$ with fair distribution such that $L(G_2) = \bigcup_{i=1}^{t} L_i$. In this sense, given a simple D0L system $G = (\Sigma, f, x)$, the length of x influences the distribution of letters of $L(G)$.

One can say that an important difference between positive and simple D0Ls is the lack of fair distribution in simple D0L languages. The fact that the technique of the proof of Theorem 4.2 was not yet used for simple D0Ls strengthens our belief that the property of fair distribution is crucial for existence of test sets for D0Ls, as well as for other families of languages.

Acknowledgment

The authors are grateful to David Maon for useful discussions on the topic of test sets, and for help in proving the results of this paper.

References

[1] J. Albert, K. Culik II and J. Karhumäki, Test sets for context free languages and systems of equations over a free monoid, *Inform. and Control* **52** (1982) 172–186.

[2] K. Culik II, The ultimate equivalence problem for D0L systems, *Acta Informatica* **10** (1978) 79–84.

[3] K. Culik II, Homomorphisms: Decidability, equality and test sets, in: R. Book, ed., *Formal Language Theory, Perspectives and Open Problems* (Academic Press, New York, 1980).

[4] K. Culik II and J. Fris, The decidability of the equivalence problem for D0L-systems, *Inform. and Control* **35** (1977) 20–39.

[5] K. Culik II and J. Karhumäki, Systems of equations over a free monoid and Ehrenfeucht Conjecture, *Discrete Math.* **43** (1983) 139–153.

[6] K. Culik II and J. Karhumäki, On the Ehrenfeucht Conjecture for D0L languages, *RAIRO* **17** (1983).

[7] K. Culik II and J.L. Richier, Homomorphism equivalence on ET0L languages, *Internat. J. Comput. Math. Sec. A* **7** (1979) 43–51.

[8] K. Culik II and A. Salomaa, On the decidability of homomorphism equivalence for languages, *JCSS* **17** (1978) 163–175.

[9] K. Culik II and A. Salomaa, Test sets and checking words for homomorphism equivalence, *JSCC* **20** (1980) 379–395.

[10] A. Ehrenfeucht, J. Karhumäki and G. Rozenberg, On binary equality sets and a solution to the Ehrenfeucht Conjecture in the binary case, *J. Algebra* **85** (1983).

[11] M.A. Harrison, *Introduction to Formal Language Theory* (Addison-Wesley, Reading, MA, 1978).

[12] J. Karhumäki, The Ehrenfeucht Conjecture: A compactness claim for finitely generated free monoids, *Theoret. Comput. Sci.*, **29** (1984) 285–308.

[13] M. Karpinski, ed., *New Scottish Book of Problems*, in preparation.

[14] A. Mandel and I. Simon, On finite semigroups of matrices, *Theoret. Comput. Sci.* **5** (1977) 101–111.

[15] G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems* (Academic Press, New York, 1980).