RAIRO. INFORMATIQUE THÉORIQUE

KAREL II CULIK
JUHANI KARHUMÄKI

On the Ehrenfeucht conjecture for DOL languages

RAIRO. Informatique théorique, tome 17, nº 3 (1983), p. 205-230
<http://www.numdam.org/item?id=ITA_1983__17_3_205_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 17, n° 3, 1983, p. 205 à 230)

ON THE EHRENFEUCHT CONJECTURE FOR DOL LANGUAGES (*) (**)

by Karel CULIK II (1)
and Juhani KARHUMÄKI (2)

Communicated by J. BERSTEL

Abstract. — *Ehrenfeucht conjectured that each language L over a finite alphabet Σ possesses a test set, that is a finite subset F of L such that every two morphisms on Σ^* agreeing on each string in F also agree on each string in L. We introduce the notion of deviation of a string with respect to a language and use it to give a sufficient condition for the existence of such a test set. Moreover, we prove that a test set effectively exists for each positive DOL language. The well known open problem whether this holds for every DOL language remains open.*

Résumé. — *Ehrenfeucht a énoncé la conjecture suivante : chaque langage L sur un alphabet fini Σ possède un ensemble de test, c'est-à-dire une partie finie F de L telle que deux morphismes quelconques sur Σ^*, qui coïncident sur les mots de F, coïncident aussi sur les mots de L. Nous introduisons la notion de déviation d'un mot par rapport à un langage et nous l'utilisons pour donner une condition suffisante à l'existence d'un ensemble de test. De plus, nous démontrons qu'un ensemble de test existe effectivement pour tout langage DOL positif. Le problème ouvert bien connu, de savoir si ceci est vrai pour tout langage DOL, reste ouvert.*

1. INTRODUCTION

Ehrenfeucht conjectured (Problem 108 in [11]) that for every language $L \subseteq \Sigma^*$ there exists a finite subset F of L such that for any pair of morphisms on Σ^*, $g(x)=h(x)$ for each x in L if and only if $g(x)=h(x)$ for each x in F. Such a finite subset F has been called a test set for L in [7] where it has been shown that Ehrenfeucht's conjecture holds for every language over a binary alphabet. It is clear from arguments in [6] that a test set can be effectively constructed for each regular language and this has been extended to context

(*) Received June 1981, revised February 1982.
(**) This research was supported by the Natural Sciences and Engineering Research Council of Canada, Grant A 7403.

(1) Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.
(2) Department of Mathematics, University of Turku, Turku, Finland.

free languages in [1]. The effective existence of a test set for a language L clearly implies that we can test whether any given morphisms g, h on Σ^* agree on L, i. e., whether or not $g(x)=h(x)$ for each $x \in L$. Therefore a test set cannot effectively exist for each context sensitive language since the testing of morphism equivalence for them has been shown to be undecidable in [6].

Both the existence of a test set and the decidability of morphism equivalence are open for all families of languages between DOL and indexed languages, *cf.* [3] where positive answers are conjectured. The proof of these conjectures is not expected to be easy since already the weakest one of them, the decidability of morphism equivalence on DOL languages, implies the decidability of the HDOL sequence equivalence problem, *cf.* [3], a longstanding open problem.

Our main purpose is to provide a partial result in the direction of these open problems, namely we show that a test set effectively exists for each positive DOL language. A DOL system is positive if each letter can be derived from every other letter in one step.

In section 3 we introduce the deviation of a string with respect to a language. It is a generalization of weighted difference from [7], which for any pair of morphisms is linearly proportional to the balance of the considered string. However, the situation in the case of an arbitrary finite alphabet is essentially more complicated than in the binary case. We show that every language L with bounded prefix deviation and fair distribution of letters possesses a test set.

In the next section we show that it is decidable whether a given DOL language L has the above properties, and if so, that a test set for L can be effectively constructed. For positive DOL languages the case covered in section 4 is also covered in section 5, but we have included it since the arguments in the case of bounded prefix deviation are more intuitive (generalization of bounded weighted difference in [7]) and the effective existence of a test set is, unlike in section 5, shown independently of [5].

In section 5 we construct for a positive DOL language a "partial" test set covering all pairs of morphisms agreeing on the language with bounded balance. The part of a test set covering the pairs of morphisms agreeing with unbounded balance is constructed in section 6.

In the last section we obtain our main result, the effective existence of a test set for each positive DOL language, by combining the partial test sets from the previous two sections. This immediately implies the decidability of morphism equivalence for positive DOL languages.

2. PRELIMINARIES

This paper deals with basic properties of free monoids from the point of view of formal language theory. As a general reference we mention [9]. The basic properties and more background material on DOL systems as well as DTOL systems can be found in [13].

A free monoid generated by a finite alphabet Σ is denoted by Σ^*. For the notational convenience we fix $\Sigma=\{a_1, \ldots, a_t\}$ if not explicitly mentioned otherwise. The elements of Σ^* are *words* or *strings* and its subsets *languages*. The identity element of Σ^*, called *empty word*, is denoted by λ, and $\Sigma^+=\Sigma^*-\{\lambda\}$.

The *length* of a word x and the *cardinality* of a finite set A is denoted by $|x|$ and $|A|$, respectively. For $w\in\Sigma^*$, the number of a's in w is denoted by $|w|_a$. When $\Sigma=\{a_1, \ldots, a_t\}$ we usually write $|x|_i$ instead of $|x|_{a_i}$. The *Parikh mapping* $\psi : \Sigma^* \to \mathbb{N}^t$ is defined by $\psi(x)=(|x|_1, \ldots, |x|_t)$. Consequently, the *Parikh vector* of a word x is denoted by $\psi(x)$. We call words x and y *Parikh equivalent* if $\psi(x)=\psi(y)$. For a word x, $\mathrm{alph}(x)$ denotes the set of letters occurring in x.

For x, y in Σ^*, the *left* (*right*) *quotient* of x by y is denoted by $y^{-1}x$ (xy^{-1}). It is undefined if y is not a prefix (suffix) of x. If x is a prefix of y we write x pref y, while x Pref y means that either x pref y or y pref x holds. By $\mathrm{pref}_n(x)$ we mean the prefix of x of length n. By definition, if $|x|<n$ then $\mathrm{pref}_n(x)=x$. For a word x (resp. language L) $\mathrm{pref}(x)$ [resp. $\mathrm{pref}(L)$] denotes the set of all prefixes of x (resp. all prefixes of words in L). Similarly for suffixes if "pref" is replaced by "suf". We say that y is a *subword* of x if $x=x_1yx_2$ for some words x_1 and x_2. The set of all subwords of a language L is denoted by $\mathrm{sub}(L)$. The set of all such words of length n is denoted by $\mathrm{sub}_n(L)$. We say that y is a *sparse subword* of x if y is obtained from x by erasing some of its occurrences of letters.

Throughout this paper our central notion is a morphism of a free monoid. We say that a morphism $h : \Sigma^* \to \Delta^*$ is *λ-free* if $h(a)\neq\lambda$ for all $a\in\Sigma$. The *size* of a morphism h, denoted by $\|h\|$, is $\|h\|=\max\{|h(a)| \mid a\in\Sigma\}$. Let $h, g : \Sigma^* \to \Delta^*$ be two morphisms and L a language over Σ. We say that h and g *agree* (resp. *length-wise agree*) on L, in symbols $h \overset{L}{\equiv} g$ (resp. $h \overset{L}{\equiv}_l g$), if $h(x)=g(x)$ for all x in L [resp. $|h(x)|=|g(x)|$ for all x in L]. The set of all pairs of morphisms agreeing on L (resp. agreeing on L length-wise) is denoted by $\mathscr{H}(L)$ [resp. $\mathscr{H}_l(L)$]. We call a language L *rich* if $\mathscr{H}(L)=\{(h, h) \mid h : \Sigma^* \to \Delta^* \text{ is a morphism}\}$, i.e., only pairs with identical components agree on L. By a *test set* for a language L we mean any finite subset F of L satisfying: for any pair (h, g) of morphisms $h \overset{F}{\equiv} g$ implies $h \overset{L}{\equiv} g$. *Ehrenfeucht conjecture* states: Every language has a test set.

Let h and g be two morphisms $\Sigma^* \to \Delta^*$ and w a word. The *balance* of a word w with respect to (h, g), in symbols $\beta_{h,g}(w)$, or shortly $\beta(w)$ if h and g are known, is defined by:

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

cf. [3]. We say that a pair (h, g) has *bounded balance* on a language L if there exists a constant c such that $|\beta(w)| \leqq c$ for all $w \in \mathrm{pref}(L)$. Moreover, we say that (h, g) agree on L with bounded balance if $h \overset{L}{\equiv} g$ and (h, g) has bounded balance on L.

Next we introduce briefly DOL systems. A DOL *system* G is a triple (Σ, f, x), where Σ is a finite alphabet, f is a morphism $\Sigma^* \to \Sigma^*$ and x, called axiom of G, is a nonempty word of Σ^*. A DOL system G defines a sequence of words: $x, f(x), f^2(x), \ldots$ A language $L(G) = \{f^n(x) \mid n \geqq 0\}$ is the language generated by G. We call a DOL system *positive* if $a \in \mathrm{sub}(f(G))$ for each pair $(a, b) \in \Sigma \times \Sigma$, i. e., any letter of Σ is derived from any other letter in one step.

Finally, we need some terminology concerning vectors over rational numbers $\mathbb{Q}$ and nonnegative integers $\mathbb{N}$. For two vectors z and z' in $\mathbb{Q}^t$, $z \leqq z'$ means that z is componentwise smaller or equal than z'. If $z \leqq z'$ and $z \neq z'$, we write $z < z'$. By the absolute value of a vector $z = (z_1, \ldots, z_t)$ we mean the number $|z| = \sum_{i=1}^{t} |z_i|$.

Let $M \subseteq \mathbb{Q}^t$. The *vector space* over $\mathbb{Q}$ generated by M is denoted by $\langle M \rangle$. When $M \subseteq \mathbb{N}^t$ we call an element z of M *minimal* if there does not exist in M any element z' such that $z' < z$. The *set of minimal elements* of M is denoted by $\mathrm{Min}(M)$. By the well-known König Infinity Lemma, *cf.* [9], $\mathrm{Min}(M)$ is always finite. If M is a finite set of numbers we denote the smallest and the largest number of M by $\min(M)$ and $\max(M)$, respectively.

3. DEVIATION

In this section we define and study our central notion: deviation of a word with respect to a language. This notion is closely related to the notion of balance of a word with respect to two morphisms, however, our new notion depends on the considered language only.

Let L be a language over $\{a_1, \ldots, a_t\}$. We define a subset of $\mathbb{N}^t$ induced by L, in symbols $\operatorname{sp}(L)$, by setting:

$$\operatorname{sp}(L)=\psi^{-1}\{\langle\psi(L)\rangle \cap \mathbb{N}^t\}.$$

Since $\psi(\operatorname{sp}(L))$ is a subtractive submonoid of the additive monoid $\mathbb{N}^t$ we have, *see* [8].

LEMMA 3.1: *For each language L over $\{a_1, \ldots, a_t\}$, $\psi(\operatorname{sp}(L))$ is finitely generated submonoid of* $(\mathbb{N}^t, +)$.

By Lemma 3.1, there exists a finite set β of vectors in $\mathbb{N}^t$, say $\beta=\{e_1, \ldots, e_p\}$, such that:

$$\psi(\operatorname{sp}(L))=\left\{\sum_{i=1}^{p} n_i e_i \,\middle|\, n_i \in \mathbb{N}, \quad \text{for} \quad i=1, \ldots, p\right\}.$$

Now, we state our basic definition.

DEFINITION 3.1: Let L be a language over $\Sigma=\{a_1, \ldots, a_t\}$ and $w \in \Sigma^*$. The *deviation of w with respect to L*, in symbols $d_L(w)$ or briefly $d(w)$ when L is known, is the set:

$$d_L(w)=\operatorname{Min}\{z \in \mathbb{N}^t \mid \psi(w) \in \psi(\operatorname{sp}(L))+z\}.$$

Example 3.1: Let $L=ab^*c$. Then:

$$\operatorname{sp}(L)=\{x \in \{a, b, c\}^* \mid \ |x|_a=|x|_c\},$$

and, in terms of Lemma 3.1:

$$\psi(\operatorname{sp}(L))=\{n(1,0,1)+m(0,1,0) \mid n, m \in \mathbb{N}\}.$$

Further for each proper prefix ab^i of a word in L, $d(ab^i)=\{(1,0,0)\}$.

Roughly speaking $d(w)$ tells how far w is from the language $\operatorname{sp}(L)$. By the König Infinite Lemma, *see* [9], $d_L(w)$ is always finite. The relation between the deviation and the balance is as follows. For every pair $(h, g) \in \mathscr{H}_l(L)$ and every word w:

$$|\beta_{h,g}(w)| \leq \min\{|z| \ \ |z \in d(w)\} \max\{\|h\|, \|g\|\}. \tag{1}$$

We also have the following important lemma.

LEMMA 3.2: *Let L be a language and (h, g) a pair of morphisms in $\mathscr{H}_l(L)$. If u and w are words such that $\psi(u) \in d_L(w)$, then $\beta_{h,g}(u)=\beta_{h,g}(w)$.*

Proof: Immediate, since $\psi(w)-\psi(u)\in\psi(\mathrm{sp}(L))$ and h and g agree lengthwise on $\mathrm{sp}(L)$.

We continue with the following observation.

THEOREM 3.1: *Every language L over $\{a_1, \ldots, a_t\}$ containing t linearly independent Parikh-vectors is rich.*

Proof: In this case $\psi(\mathrm{sp}(L))=\mathbb{N}^t$, and hence for any pair $(h,g)\in\mathscr{H}_l(L)$, $|h(a_i)|=|g(a_i)|$ for $i=1, \ldots, t$. Consequently, for any pair $(h,g)\in\mathscr{H}(L)$, $h(a_i)=g(a_i)$ holds true for $i=1, \ldots, t$, too.

The problem of whether we can effectively find $\mathrm{sp}(L)$ for a given language, or as a special case effectively decide whether L is rich, depends, of course, on the way how L is given. For DOL languages, which we are particularly interested in, this can be done by:

LEMMA 3.3: *Let $G=(\Sigma, f, x)$ be a* DOL *system. There exists an integer $k<|\Sigma|$ such that $\psi(L(G))$ is included in the vector space generated by $\{\psi(f^i(x))\,|\,i\leqq k\}$.*

Lemma 3.3, as well as Lemma 3.4, follows easily from the properties of vector spaces.

LEMMA 3.4: *Let L be a* DOL *language generated by a* DOL *system* (Σ, f, x). *If $u\in\mathrm{sp}(L)$, then also $f(u)\in\mathrm{sp}(L)$.*

DEFINITION 3.2: Let L and L' be languages over the same alphabet. We say that L has *bounded prefix deviation with respect to L'* if there exists a constant C such that for every prefix w of a word in L:

$$\min\{|z|\ \ |z\in d_{L'}(w)\}\leqq C.$$

If the above is satisfied for $L=L'$ we say that L has *bounded prefix deviation.*

It follows from (1) that if L has bounded prefix deviation, then each pair (h,g) of morphisms in $\mathscr{H}_l(L)$ has bounded balance on L. However, the bound depends on the pair. On the other hand, a pair (h,g) may have bounded balance on such a language which does not have bounded prefix deviation, *see* Example 5.1.

Our notions of the deviation and the bounded prefix deviation are generalizations of those of the weighted difference and the bounded prefix difference defined in [7]. We can also generalize some arguments of [7] to yield the following theorem. To be able to state it we still need one notion. We say that a language L has a *fair distribution of letters* if there exists a constant q such that every subword in L with the length of least q contains all letters of the alphabet of L.

THEOREM 3.2: *Every language L over $\{a_1, \ldots, a_t\}$ with bounded prefix deviation and fair distribution of letters has a test set.*

Proof: Let the prefix deviation of L be bounded by C and let q be a constant giving a fair distribution of letters for L. We first prove:

Claim: There exists a constant N such that for any $uv \in \mathrm{pref}(L)$, with $|v| \geqq N$, the following holds true: for any pair (h, g) in $\mathscr{H}_l(L)$:

$$\min\{|h(uv)|, |g(uv)|\} \geqq \max\{|h(u)|, |g(u)|\}.$$

The claim is proved as follows. Let z be a vector in $d(u)$ such that $|z| \leqq C$. We start by showing that there exist a constant D and a vector z_1 in $\psi(\mathrm{sp}(L))$ such that:

$$z + D\eta \geqq z_1 \geqq z, \tag{2}$$

where $\eta = (1, \ldots, 1)$, i.e. all components of η equal 1. According to Lemma 3.1 let $\psi(\mathrm{sp}(L))$ be generated by $\{e_1, \ldots, e_p\}$. We set:

$$D = C + C \sum_{i=1}^{p} |e_i| \qquad \text{and} \qquad z_1 = C \sum_{i=1}^{p} e_i.$$

Then:

$$z + D\eta > C\left(\sum_{i=1}^{p} |e_i|\right)\eta = |z_1|\eta > z_1$$

and:

$$z < |z|\eta \leqq C\eta < z_1,$$

where the last inequality follows since each letter a_i occurs in a word of L. Hence (2) has been proved.

Now, let $N = Dq$. Since $|v| \geqq N$, v contains as a sparse subword a word v' such that $\psi(v') \geqq D\eta$. Assuming, without loss of generality, that $|h(u)| > |g(u)|$ we should show that $|g(uv)| \geqq |h(u)|$. For a vector y in $\mathbb{N}^t$ let $\bar{y}$ denote a word such that $\psi(\bar{y}) = y$. Then, by Lemma 3.2 and the above, we obtain:

$$|g(uv)| - |h(u)| = |g(\bar{z}v)| - |h(\bar{z})| \geqq |g(\bar{z}v')| - |h(\bar{z})|$$
$$\geqq |g(\bar{z}\,\bar{\eta}^D)| - |h(\bar{z})| \geqq |g(\bar{z}^1)| - |h(\bar{z}_1)| = 0.$$

Thus, the proof of the claim is completed and we return to the proof of the theorem.

We divide L into two parts F and $L-F$ by setting $F=\{w \in L \mid |w| \leqq 3N\}$. Moreover, for every w in $L-F$ we choose a fixed decomposition:

$$w=u_1 \dots u_m \qquad \text{with} \quad N \leqq |u_j| \leqq 2N. \tag{3}$$

For each such decomposition and for each $j=1, \dots, m$ we define pairs (z_j, u_j), where z_j is a fixed vector in $d(u_1 \dots u_{j-1})$ satisfying $|z_j| \leqq C$. Such pairs are called *pieces*. Clearly, the number of different pieces is finite. We say that two pieces (z, x) and (z', x') *occur consecutively in* L if there exists in L a word w such that x and x' occur consecutively in its decomposition (3), say $x=u_k$ and $x'=u_{k+1}$, and moreover $z \in d(u_1 \dots u_{k-1})$ and $z' \in d(u_1 \dots u_k)$. Now, we choose a finite subset L' of L such that for any pair of pieces if they occur consecutively in L they occur consecutively already in L'.

Finally, obviously there exists a finite subset F' of L such that $\mathrm{sp}(F \cup L' \cup F') = \mathrm{sp}(L)$. We infer that $F \cup L' \cup F'$ is a test set for L. We should show that for any pair (h, g) of morphisms $h \overset{F \cup L' \cup F'}{\equiv} g$ implies $h \overset{L}{\equiv} g$. Let $(h, g) \in \mathscr{H}(F \cup L' \cup F')$ and w be an arbitrary word in $L-(F \cup L' \cup F')$. Let the decomposition of w according to (3) be $w=u_1 \dots u_m$. Since $(h, g) \in \mathscr{H}(F \cup L' \cup F')$ and $\mathrm{sp}(L)=\mathrm{sp}(F \cup L' \cup F')$, h and g agree lengthwise on L and therefore by the claim and the choice of (3):

$$\min\{|h(u_1 \dots u_i)|, |g(u_1 \dots u_i)|\} \geqq \max\{|h(u_1 \dots u_{i-1})|, |g(u_1 \dots u_{i-1})|\}$$

for $i=1, \dots, m$. Consequently, the choice of L' and the fact $h \overset{L'}{\equiv} g$ imply that if $h(u_1 \dots u_{i-1})$ Pref $g(u_1 \dots u_{i-1})$ then also $h(u_1 \dots u_i)$ Pref $g(u_1 \dots u_i)$. So we derive inductively that $h(w)=g(w)$ which completes the proof of the theorem.

We note that not only the assumption that L has bounded prefix deviation but also the assumption that L has fair distribution of letters is essential for our above proof, i. e. for the piece construction. This is seen as follows.

Example 3.1 (Continued): As already mentioned the language $L=ab^*c$ has bounded prefix deviation. However, the pairs (h_k, g_k) of morphisms, for $k \geqq 1$, defined by:

$$h_k: \begin{cases} a \to a(ba)^k, \\ b \to ba, \\ c \to ba, \end{cases} \qquad g_k: \begin{cases} a \to ab, \\ b \to ab, \\ c \to (ab)^k a, \end{cases}$$

show that the claim in the proof of Theorem 3.2 does not hold true for L. Despite of that we, of course, believe that the theorem is true without the assumption of fair distribution of letters. Indeed, $\{ac, abc\}$ is a test set for L.

4. DOL LANGUAGES WITH BOUNDED PREFIX DEVIATION

Whether the assumptions of Theorem 3.2 imply the effective existence of a test set depends, of course, on how L is given. In this section we show that it is decidable whether a given DOL language satisfies the assumptions of Theorem 3.2 and, moreover, if this is the case, that a test set for it can be effectively found.

LEMMA 4.1: *Given a* DOL *language L, it is decidable whether it has fair distribution of letters. Moreover, if this is the case a constant q such that any subword u of L, with $|u|\geqq q$, contains all letters of L can be effectively found.*

Proof: Let $L=L(G)$ for a DOL system $G=(\Sigma, f, x)$ satisfying $\Sigma\subseteq\mathrm{sub}(L(G))$. For each a in Σ let $G_a=(\Sigma, f, a)$. We divide Σ into two disjoint parts Σ_f and Σ_i by setting $\Sigma_f=\{a\in\Sigma \mid L(G_a) \text{ is finite}\}$ and $\Sigma_i=\Sigma-\Sigma_f$. If $\Sigma_i\neq\emptyset$, i. e., $L(G)$ is finite, we are done.

So, assume that $\Sigma_i\neq\emptyset$. We claim that L has a fair distribution of letters, if and only if, the following two conditions are satisfied:

(i) there exists an n_0 such that for every a in Σ_i $\mathrm{alph}(f^n(a))=\Sigma$ for $n\geqq n_0$, and

(ii) the language $\Sigma_f^*\cap\mathrm{pref}(L(G_a))$ and $\Sigma_f^*\cap\mathrm{suf}(L(G_a))$ are finite for every a in Σ.

Clearly, the conditions (i) and (ii) are necessary for a fair distribution of letters in L. They are also sufficient since (ii) rules out the possibility that L would contain arbitrarily long subwords from Σ_f^* and after that (i) guarantees that any long enough subword contains all letters from Σ. Now, the first sentence of the lemma follows, since the validity of (i) and (ii) for a DOL language can easily be checked. Furthermore, if L satisfies the conditions (i) and (ii) then a bound giving a fair distribution for L can be effectively found.

LEMMA 4.2: *Given a* DOL *language L, it is decidable whether it has bounded prefix deviation. Moreover, if this is the case an upper bound for it can be effectively found.*

Proof: Let $L=L(G)$ for a DOL system $G=(\Sigma, f, \omega)$ with $\Sigma=\{a_1, \ldots, a_t\}$. By Lemma 3.3, we can effectively find $\mathrm{sp}(L)$. Let $F:\Sigma^*\to\mathbb{N}$ be a mapping defined by:

$$F(w)=\sum_{i=1}^{t} n_i|w|_i \qquad \text{for some} \quad n_i\in\mathbb{Z}$$

and satisfying:

$$F(w)=0 \qquad \text{if and only if} \quad w\in\mathrm{sp}(L). \tag{1}$$

Such an F can be defined, e. g., *via* a linear functional $\mathbb{Q}^t \to \mathbb{Q}$ having $\langle \psi(\mathrm{sp}(L)) \rangle$ as its kernel. Consequently, F can be computed from L. Let h and g be morphisms of Σ^* satisfying $|h(a_i)|-|g(a_i)|=n_i$. Therefore $F(w)=\beta_{h,g}(w)$ for all $w \in \Sigma^*$.

We claim that L has bounded prefix deviation if and only if the pair (h, g) has bounded balance on L. The implication "bounded prefix deviation implies bounded balance" is clear, *see* equation (1) in Section 3. So assume that (h, g) has bounded balance on L, i. e. $F(x)$ is bounded on $\mathrm{pref}(L)$. We show that:

$$F^{-1}(m) \cap \left(\bigcup_{w \in \mathrm{pref}(L)} d(w) \right), \tag{2}$$

is finite for each $m \in \{ F(v) \mid v \in \mathrm{pref}(L) \}$. If this is not the case, then, by the König Infinite Lemma, *cf.* [9], there exist words w_1 and w_2 in $\mathrm{pref}(L)$ such that $F(w_1)=F(w_2)$, $\psi(w_1)<\psi(w_2)$ and $\psi(w_1), \psi(w_2) \in \bigcup_{w \in \mathrm{pref}(L)} d(w)$. Let $w' \in \psi^{-1}(\psi(w_1)-\psi(w_2))$. Then $F(w')=0$ and, hence, by (1), $w' \in \mathrm{sp}(L)$. Consequently, $\psi(w_2)$ cannot be in $d(w)$ for any w_1, a contradiction. So (2) is always finite, and therefore L has bounded prefix deviation.

Now, the first sentence of Lemma 4.2 follows. Indeed, in [2] it has been shown that it is decidable whether an arbitrary pair of morphisms has bounded balance on a DOL language.

Knowing that the prefix deviation of L is bounded, an upper bound for it can be effectively found as follows. Let $xa \in \mathrm{pref}(L)$, with $a \in \Sigma \cup \{\lambda\}$. We associate to xa a pair $(\bar{d}(x), a)$ where $\bar{d}(x)$ is a fixed element in $d(x)$. Let L_0 be the set of all such pairs. For each pair $(\bar{d}(x), a)$ we define a finite set $S(\bar{d}(x), a)$ of pairs as follows. Let $yb \in \mathrm{pref}(f(a))$, where $b \in \Sigma$ or if $f(a)=\lambda$ then $b=\lambda$, and let x' be a fixed word in $f(\psi^{-1}(\bar{d}(x)))$. $S(\bar{d}(x), a)$ contains all pairs $(\bar{d}(x'y), b)$ where again $\bar{d}(x'y)$ denotes a fixed vector in $d(x'y)$. Let the set of all pairs thus obtained be L_1' and let $L_1=L_0 \cup L_1'$. We proceed inductively to define the sets L_i for $i \geqq 0$. Now, the important observation is that all the deviations (or more precisely a representative of all the deviations) of prefixes of words in $\{ h^i(w) \mid i \leqq n \}$ are obtained as first components of elements of L_n. This follows easily from Lemma 3.4 by induction on n. From the definition of L_i-sets it follows that $L_0 \subseteq L_1 \subseteq L_2 \subseteq \ldots$ Moreover, since L has bounded prefix deviation we finally find an i_0 such that $L_{i_0+1}=L_{i_0}$, and consequently, assuming that the fixation of the value of deviation is always done in the same way, we have $L_i=L_{i_0}$ for each $i \geqq i_0$. Hence, a bound for the prefix deviation has been found.

Now, we are ready for the main result of this section.

THEOREM 4.1: *Given a* DOL *language L, it is decidable whether L has bounded prefix deviation and fair distribution of letters, and if this is the case, then a test set for L can be effectively found.*

Proof: Let $L=L(G)$ for a DOL system $G=(\Sigma, f, \omega)$. The first part of the theorem is proved in Lemmas 4.1 and 4.2. The second part is deduced from the proof of Theorem 3.2 as follows. Now, instead of using pieces where the lengths of the second components are between N and $2N$ it is preferable to use pieces of the length between N and $2KN$, where K is a constant satisfying: if $u \in \mathrm{sub}(L)$, with $|u| \geqq KN$, then $|h^n(u)| \geqq N$ for each $n \geqq 0$. Such a constant K clearly exists. Namely, this makes it possible to generate the "piece decomposition of L", i.e., L with the information how its words are decomposed according to (3) in Theorem 3.2 into pieces, as a DOL language. Let $G_p=(\Sigma_p, f_p, x_p)$ be such a system. Consequently, Σ_p consists of all second components of pieces of L as well as short words, i.e., words in F, specified in the proof of Theorem 3.2.

We continue by showing that we can incorporate into each occurrence of Σ_p in L also the information about what is the deviation at the beginning of this *occurrence* of a letter. More precisely, let $y\,x\,y'$ be a word in L such that x corresponds to a piece. We want to put into x the information about $d(y)$. This can be done as follows. First, we recall that the constant N was selected in the proof of Theorem 3.2 such that if $u \in \mathrm{sub}(L)$, with $|u| \geqq N$, then for all $w \in \mathrm{pref}(L)$ there exists z in $d(w)$ such that $\psi(u) \geqq z$. Consequently, we can incorporate the information about $d(y)$ into x, for example, by using barred letters. (Observe that for short words $d(y)=\bar{0}$.) But can the sequence still be generated by a DOL system? The answer is "yes", since, as we have already pointed out, $\mathrm{sp}(L)$ is closed under f (Lemma 3.4), and consequently the deviation at the beginning of an occurrence of a piece obtained from x by applying f_p can be computed from $f(x)$ and $d(y)$, i.e., from x and the barred letters of x. So a new morphism, and also a DOL system, say $\bar{G}_p=(\bar{\Sigma}_p, \bar{f}_p, \bar{x}_p)$ can be defined in such a way that it contains the entire information about how words of L are decomposed into pieces.

The construction of a test set for L is now easy. The requirement for L' in the proof of Theorem 3.2 is surely fulfilled if we take from $L(\bar{G}_p)$ a finite subset L_p such that it contains all the subwords of $L(\bar{G}_p)$ of the length two, and choose L' equal to a finite subset of L corresponding to L_p. By the definition of L_p, we can effectively find an n_0 such that $L_p \subseteq \{\bar{f}_p^{\,n}(\bar{x}_p) \mid n \leqq n_0\}$. Consequently, a finite set $\{f^n(x) \mid n \leqq n_0\}$ is a test set for L.

COROLLARY 4.1: *Given a positive* DOL *language L with bounded prefix deviation a test set for L can be effectively found.*

Proof: Clearly, positive DOL languages have fair distribution of letters.

5. MORPHISMS AGREEING ON A POSITIVE DOL LANGUAGE WITH BOUNDED BALANCE

In this section we consider the case when two morphisms agree on a given positive DOL language L with bounded balance. We show that there exists a finite subset F of L such that any pair of morphisms with bounded balance on L agrees on L if and only if it agrees on F. Thus the considerations of this section yields an alternate proof for the existence of a test set (and hence also for the effective existence of a test set, *cf.* Section 7) for positive DOL language with bounded prefix deviation (*cf.* Corollary 4.1). Moreover, this section takes also care of morphisms agreeing on a positive DOL language with bounded balance although the language itself has unbounded deviation. The reason why we included Section 4 is that the considerations therein are, we believe, more intuitive and neater.

Example 5.1: Let G be a positive DOL system defined by the morphism:

$$f: \begin{cases} a \to aaabcd, \\ b \to abcbcd, \\ c \to acbcbd, \\ d \to acbddd \end{cases}$$

and the axiom $abcd$. Clearly, $\psi(L(G)) \subseteq \{(k, k, k, k) \mid k \geqq 1\}$ and therefore $\psi(\mathrm{sp}(L(G))) = \{(k, k, k, k) \mid k \geqq 1\}$. We claim that, for each $n \geqq 1$, $x_n = \mathrm{pref}_{6^n} f^n(abcd)$ satisfies $|x_n|_a - |x_n|_d \geqq 2^n$. Since $x_1 = aaabcd$ the claim is true for $n=1$. So the claim follows from the relation $x_{n+1} = f(x_n)$ by induction on n. The claim immediately implies that $L(G)$ has unbounded prefix deviation. Consequently, a positive DOL language may possess unbounded prefix deviation.

Consider now two morphisms defined by:

$$h: \begin{cases} a \to ab, \\ b \to a, \\ c \to bab, \\ d \to abab, \end{cases} \qquad g: \begin{cases} a \to ab, \\ b \to aba, \\ c \to b, \\ d \to abab. \end{cases}$$

Clearly, h and g agree on the language $L = \{a, bc, cb, d\}^*$ with bounded balance (in fact, with balance 2). Since $L(G) \subseteq L$, (h, g) also agrees on $L(G)$

with bounded balance. It is also easy to give (periodic) pairs of morphisms agreeing on L with unbounded balance.

To cover the cases like in the above example, we have to prove:

THEOREM 5.1: *Let L be a positive* DOL *language. There exists a finite subset F of L such that F is a test set for all pairs (h, g) of morphisms having bounded balance on L, i. e., for any pair (h, g), $h \overset{F}{\equiv} g$ implies that either $h \overset{L}{\equiv} g$ or (h, g) has unbounded balance on L.*

Proof: Let L be generated by a positive DOL system $G=(\Sigma, f, x)$ with $\Sigma=\{a_1, \ldots, a_t\}$. As shown in [2] we can construct a DTOL system G' and a morphism τ such that:

$$\operatorname{pref}(L)=\tau(L(G')).$$

Consequently, $\psi(\operatorname{pref}(L))$ has a matrix representation, i. e., there exist matrices $M_1, \ldots, M_k, M$ and a vector π over $\mathbb{N}$ such that $\psi(\operatorname{pref}(L))$ coincides with the range of the function $F:\{1, \ldots, k\}^* \to \mathbb{N}^{|\Sigma|}$ defined by:

$$F(i_1 \ldots i_q)=\pi M_{i_1} \ldots M_{i_q} M \qquad \text{for} \quad q \geqq 0, \quad i_j \in \{1, \ldots, k\}.$$

Moreover:

$$\psi(\operatorname{pref}(f^n(x)))=\{F(y) \mid \; |y|=n+1\}. \tag{1}$$

Now, let h and g be two morphisms of Σ^*. Clearly:

$$\{\beta_{h, g}(w) \mid w \in \operatorname{pref}(L)\}=\{F(y) \eta_{h, g} \mid y \in\{1, \ldots, k\}^*\}, \tag{2}$$

where $\eta_{h, g}=(|h(a_1)|-|g(a_1)|, \ldots, |h(a_t)|-|g(a_t)|)$. We assume that (2) is finite, i. e., (h, g) has bounded balance on L, and apply results of Mandel and Simon, *cf.* [12] Section 5, in the following form. There exists a constant n_G such that all the values of (2) are obtained when y ranges over $\{y \in\{1, \ldots, k\}^* \mid \; |y|<n_G\}$. Moreover, n_G can be chosen independently of $\eta_{h, g}$, i. e., independently of (h, g). Consequently, by (1), for any pair (h, g) of morphisms having bounded balance on L, all possible values of the balance on L are already obtained on the finite language $L'=\{f^n(x) \mid n \leqq n_G\}$.

Next we establish an analogy to the claim of the proof of Theorem 3.2.

Claim I: There exists a constant N such that for any $uv \in \operatorname{pref}(L)$, with $|v| \geqq N$, the following holds true: for any pair (h, g) in $\mathscr{H}_l(L)$ having bounded balance on L:

$$\min\{|h(uv)|, |g(uv)|\} \geqq \max\{|h(u)|, |g(u)|\}.$$

Claim I is proved as follows. Let (h, g) be a pair of morphisms satisfying the above assumptions and let $K=\max\{|x| \mid x\in L'\}$. Then:

$$|\beta_{h,g}(w)|\leqq K\max\{\|h\|, \|g\|\} \quad \text{for every } w \text{ in } \quad \operatorname{pref}(L).$$

Consequently, if we show that there exists a constant N such that for every $v\in \operatorname{sub}(L)$, with $|v|\geqq N$:

$$\min\{|h(v)|, |g(v)|\}\geqq K\max\{\|h\|, \|g\|\}, \tag{3}$$

then Claim I follows. To prove (3) we apply the length argument to a fixed word of L containing all letters of Σ, i. e. we obtain that:

$$\sum_{i=1}^{t} n_i|h(a_i)|=\sum_{i=1}^{t} n_i|g(a_i)|$$

for some positive values of $n_1, \ldots, n_t$. Therefore:

$$|h(z)|\geqq\|g\| \qquad \text{and} \qquad |g(z)|\geqq\|h\| \tag{4}$$

whenever $\psi(z)\geqq(n_1, \ldots, n_t)$. Now, we use the positiveness of G. This yields a constant N such that if $v\in\operatorname{sub}(L)$, with $|v|\geqq N$, then $\psi(v)\geqq K(n_1, \ldots, n_t)$. Thus, (3) and also Claim I follows from (4).

To complete the proof of Theorem 5.1 we have to show how Claim I implies the existence of a finite subset of L such that it tests whether arbitrary pair of morphisms having bounded balance on L agrees on L. First we recall a result mentioned already in the proof of Theorem 4.1: there exists a DOL system $G_p=(\Sigma_p, f_p, x_p)$, where $\Sigma_p=\bigcup_{i=1}^{N'}\Sigma^i$ for some $N'>N$, such that the letters in $\bigcup_{i=1}^{N-1}\Sigma^i$ occur only in a finite subset of $L(G_p)$ and $\psi(L(G_p))=L$, where ψ is the morphism mapping each element of Σ_p into a corresponding word of Σ^*.

We make another claim.

Claim II: Let $w', w''\in\Sigma_p$ and (h, g) be a pair of morphisms in $\mathscr{H}_l(L)$ having bounded balance on L. There exists a finite language $L''\subseteq L(G_p)$, independent of (h, g), such that:

$$\{\beta_{h,g}(\psi(w_1 w'))\mid w_1 w' w'' w_2\in L'' \quad \text{for some} \quad w_1, w_2\in\Sigma_p^*\}$$
$$=\{\beta_{h,g}(\psi(w_1 w'))\mid w_1 w' w'' w_2\in L(G_p) \quad \text{for some} \quad w_1, w_2\in\Sigma_p^*\}.$$

The proof of the Claim II is as follows. It is a simple modification of the construction presented in [2] to see that there exist a DTOL system G_1 and a morphism τ_1 such that:

$$\operatorname{pref}(L(G_p)) \cap \Sigma_p^* w' w'' = \tau_1(L(G_1)).$$

Consequently, the ideas of the beginning of this proof become applicable, and prove Claim II.

Now, we are ready to finish the proof of Theorem 5.1. Indeed, Claims I and II guarantee that the arguments of the proof of Theorem 3.2, e. g. the piece construction, can be modified in a obvious way to complete the proof of Theorem 5.1.

Note that we do not require that F in Theorem 5.1 is found effectively.

6. MORPHISMS AGREEING ON A POSITIVE DOL LANGUAGE WITH UNBOUNDED BALANCE

Now, we turn to consider the case when two morphisms agree on a positive DOL language L with unbounded balance. Necessarily, this means that the DOL language must have unbounded prefix deviation. We shall prove an analogy to Theorem 5.1 for pairs of morphisms having unbounded balance on L. In doing this we use ideas, especially the "shifting argument", presented in [4].

LEMMA 6.1. *Let $G=(\Sigma, f, x)$ be a positive* DOL *system. For each $\varepsilon>0$ there exists an integer n_ε such that:*

$$|d(w)|_{\min} \leqq \varepsilon |f^n(x)|$$

for every $n \geqq n_\varepsilon$ and $w \in \operatorname{pref}(f^n(x))$, where $|d(w)|_{\min} = \min\{|z| \mid z \in d(w)\}$.

Proof : Let v be a word in $L(G)$ such that $\operatorname{alph}(v)=\Sigma=\{a_1, \ldots, a_t\}$. Since G is positive we find a constant s such that for all a in Σ:

$$f^s(a) = \alpha_a \beta_a \gamma_a \qquad \text{with} \quad \psi(\alpha_a) \geqq \psi(v) \qquad \text{and} \quad \psi(\gamma_a) \geqq \psi(v). \tag{1}$$

Now, for each a in Σ, we fix v_a to be a word obtained from $f^s(a)$ by erasing from it a word Parikh-equivalent to v, and we define $\bar{f}: \Sigma^* \to \Sigma^*$ by $\bar{f}(a) = v_a$. This means that for each word y $\psi(f^s(y)) - \psi(\bar{f}(y))$ belongs to

$\psi(\mathrm{sp}(L))$. Let q be a constant satisfying:

$$\psi(v^{|f^q(x)|}) \geqq \psi(f^s(a)) \quad \text{for all } a \in \Sigma. \tag{2}$$

We set $\bar{x}_0 = f^r(x)$ where r satisfies:

$$\psi(f^r(x)) \geqq (|f^q(x)|+1)\,\psi(v) \tag{3}$$

and define, for $i=0, \ldots, s-1$, DOL systems:

$$G_i = (\Sigma, \bar{f}, \bar{x}_i) \qquad \text{where} \quad \bar{x}_i = f^i(\bar{x}_0).$$

We first claim that for every prefix $w \in \mathrm{pref}(f^{ns+r+i}(x))$ there exists a vector z in $d(w)$ such that:

$$\psi(\bar{f}^n(\bar{x}_i)) \geqq z \qquad \text{for} \quad n \geqq 0. \tag{4}$$

We fix an i and prove (4) by induction on n. The case $n=0$ is clear since $f^0(\bar{x}_i) = f^{r+i}(x)$. So let $w \in \mathrm{pref}(f^{(n+1)s+r+i}(x))$, i.e., $w = w_1 w_2$ where $w_1 = f^s(w_1')$ for some word w_1' and $w_2 \in \mathrm{pref}(f^s(b))$ for some b in Σ. By induction hypothesis, there exists a vector z_1 in $d(w_1')$ such that $\psi(\bar{f}^n(\bar{x}_i)) \geqq z_1$. Now, since $a \in \mathrm{sub}(\bar{f}(a))$ for each a, we conclude from (3) that there exist a constant k and a word u, with $\psi(u) = z_1$, such that:

$$\psi(\bar{f}^n(\bar{x}_i)) \geqq \psi(uv^k) \qquad \text{where} \quad |uv^k| \geqq |f^q(x)|. \tag{5}$$

By (2), (5) and the definition of w_2, $|uv^k|\,\psi(v) - \psi(w_2)$ contains only positive components. Moreover, by the definition of $\bar{f}$, the same holds true for all vectors $\psi(\bar{f}(y)) - |y|\,\psi(v)$ where $y \in \Sigma^*$. Consequently, we obtain:

$$\begin{aligned}
\psi(\bar{f}(uv^k)) &\geqq \psi(\bar{f}(uv^k)) - |uv^k|\,\psi(v) + \psi(w_2) \\
&= \psi(\bar{f}(u)) - |u|\,\psi(v) + \psi(f(v^k)) - |v^k|\,\psi(v) + \psi(w_2) \\
&\geqq \psi(\bar{f}(u)) - |u|\,\psi(v) + \psi(w_2).
\end{aligned}$$

Because of the relation $\psi(\bar{f}(u)) - |u|\,\psi(v) \geqq 0$ there exists in $d(\bar{f}(u))$ a vector, say z_2, such that $\psi(\bar{f}(u)) - |u|\,\psi(v) \geqq z_2$. Now, remember that $\psi(u) \in d(w_1')$. This implies, since $\mathrm{sp}(L)$ is closed under f (*cf.* Lemma 3.4) and hence also under $\bar{f}$, that there exists in $d(w_1)$ a vector, say z_3, such that $z_2 \geqq z_3$. In conclusion, we have:

$$\psi(\bar{f}(uv^k)) \geqq z_3 + \psi(w_2) \qquad \text{where} \quad z_3 \in d(w_1),$$

which, by (4) and the identity $w = w_1 w_2$, completes the induction.

By (4), to complete the proof of the lemma it is enough to show that, for $i = 0, \ldots, s-1$:

$$\lim_{n \to \infty} \frac{|\bar{f}^n(\bar{x}_i)|}{|\bar{f}^{ns}(x_i)|} \to 0. \tag{6}$$

Let M_1 and M_2 denote the growth matrices of G and G_1, respectively, *cf.* [13]. By the definition of $\bar{f}$, we have $M_1^s \geqq M_2 + I$, where $\geqq$ denotes the natural componentwise order. Let $\pi = \psi(\bar{x}_i)$ and $\eta = (1, \ldots, 1)^T$. We have:

$$0 \leqq \frac{|\bar{f}^n(\bar{x}_i)|}{|f^{ns}(\bar{x}_i)|} = \frac{\pi M_2^n \eta}{\pi M_1^{sn} \eta} \leqq \frac{\pi M_2^n \eta}{\pi (M_2 + I)^n \eta} \leqq \frac{\pi M_2^n \eta}{\pi (n M_2^{n-1}) \eta} \leqq \frac{C t^2}{n},$$

where C is an upper bound for the values of entries in M_2. So (6) and hence also Lemma 6.1 follows.

We also need another lemma, a lemma on formal power series (as a general reference of the topic we mention [14]).

LEMMA 6.2: *Let $F : \Sigma^* \to \mathbb{Z}$ be a $\mathbb{Z}$-rational formal power series and N a constant. There exists a constant n_0, depending on the cardinality of Σ and N only, such that F is unbounded if and only if there exists a word u such that $N < |u| \leqq N + n_0$ and $F(u) \notin \{F(w) \mid |w| \leqq N\}$.*

Proof: The proof of Lemma 6.2 can be derived as an application of the theory of Hankel matrices, e. g. by using Corollary II. 3.4 in [14].

Next we prove an analogy of Theorem 5.1.

THEOREM 6.1: *Let $G = (\Sigma, f, x)$ be a positive* DOL *system and $L = L(G)$. There exists a finite subset F' of L such that F' is a test set for all pairs of morphisms having unbounded balance on L, i. e., for each pair (h, g), $h \overset{F'}{\equiv} g$ implies that $h \overset{L}{\equiv} g$ or (h, g) has bounded balance on L.*

Proof: By the standard decomposition technique, *cf.* [12], we may decompose G into a finite number of systems such that each such system $(\overline{\Sigma}, \overline{f}, \overline{x})$ satisfies: $\mathrm{sub}_2(\overline{f}(a)) = \mathrm{sub}_2(\overline{f}(b))$ for all $(a, b) \in \overline{\Sigma} \times \overline{\Sigma}$. Consequently, we may assume that G shares this property.

We first assume that $x \in \Sigma$, say $x = a$. This means that $\mathrm{sub}_2(L) = \mathrm{sub}_2(f(G))$ for all $b \in \Sigma$. Let (h, g) be an arbitrary pair of morphisms having unbounded balance on L. We show that there exists an n_0 such that if h and g agree on $\{f^n(a) \mid n \leqq n_0\}$, then they agree on L, too. Since n_0 is shown to be independent of (h, g) the theorem follows for DOL languages generated by positive systems with the axiom of length 1.

From now on we consider a fixed, but arbitrary, pair of morphisms having unbounded balance on L and agreeing on a later specified finite language $F' \subseteq L$. Since $h(f(a)) = g(f(a))$, we have:

$$\sum_{a \in \Sigma} m_a |h(a)| = \sum_{a \in \Sigma} m_a |g(a)|$$

for some positive integers m_a. Consequently, there exists a constant q, independent of (h, g), such that:

$$\min\{\|h\|, \|g\|\} \geqq \frac{1}{q} \max\{\|h\|, \|g\|\}. \tag{7}$$

On the other hand, the positiveness of G implies the existence of a constant $K > 0$, again independently of (h, g), such that:

$$\left.\begin{aligned} &K|w| \ \|h\| \leqq |h(w)| \leqq \frac{1}{K}|w| \ \|h\|, \\ \text{and:} \\ &K|w| \ \|g\| \leqq |g(w)| \leqq \frac{1}{K}|w| \ \|g\| \end{aligned}\right\} \tag{8}$$

for every subword w of L containing all letters of Σ. Consequently, setting $K' = q/K^2$ we have:

$$\frac{1}{K'}|g(w)| \leqq |h(w)| \leqq K'|g(w)| \tag{9}$$

for $w \in \mathrm{sub}(L)$ with $\mathrm{alph}(w) = \Sigma$.

We choose a constant k such that:

$$|f^{k-1}(b)|_b \geqq K' + 1 \qquad \text{for each} \quad b \in \Sigma. \tag{10}$$

Let now $f^n(a)=uv$ for some words u and v and large enough n. Further let $|u|\leqq|v|$ (the other case is symmetric) and $\mathrm{pref}_1(v)=\alpha$. We search for ancestors of α, i.e., occurrences $\alpha_1, \alpha_2, \ldots$ of letters in L such that $f^i(\alpha_i)$ contains the above mentioned occurrence of α. Clearly, since G is positive, there exist α_i and α_j, $i<j$, and a constant $N>0$ such that $\alpha_i=\alpha_j$, their right neighbours in $L(G)$ are the same, say β, and moreover:

$$|f^{i-k}(b)|\geqq\frac{1}{N}|f^n(a)| \tag{11}$$

for all b in Σ, large enough n, and k defined in (10). Observe that constant N can be chosen independently of u, v and n, while α_i and α_j, of course, depend on u, v and n. This is because α_i and α_j can always be chosen from the uniformly bounded initial part of the sequence generated by G. (Here the assumption $|v|\geqq|u|$ is needed to guarantee the existence of β.)

Our next goal is to fix the integer n in the decomposition $f^n(a)=uv$. By (7), (8) and (11), we have:

$$\left.\begin{aligned}&|h(f^{i-k}(b))|\geqq\frac{K}{Nq}|f^n(a)|\max\{\|h\|, \|g\|\},\\ \text{and:}\quad&\\ &|g(f^{i-k}(b))|\geqq\frac{K}{Nq}|f^n(a)|\max\{\|h\|, \|g\|\}.\end{aligned}\right\} \tag{12}$$

On the other hand, by Lemma 6.1, for every $\varepsilon>0$ there exists n_ε such that:

$$|d(u)|_{\min}\leqq\varepsilon|f^n(a)| \qquad \text{for} \quad n\geqq n_\varepsilon,$$

and hence:

$$|\beta(u)|\leqq\varepsilon|f^n(a)|\max\{\|h\|, \|g\|\} \qquad \text{for} \quad n\geqq n_\varepsilon. \tag{13}$$

By (12) and (13), if n is large enough, then for all letters b in Σ:

$$\left.\begin{aligned}&|h(f^{i-k}(b))|\geqq 2|\beta_{\max}|,\\ \text{and:}\quad&\\ &|g(f^{i-k}(b))|\geqq 2|\beta_{\max}|,\end{aligned}\right\} \tag{14}$$

where $\beta_{\max}=\max\{|\beta(w)| \;\; | w\in\mathrm{pref}\,(f^m(a))$ for some $m\leqq n\}$, i.e., we can find for any decomposition $f^n(a)=uv$, with $|v|\geqq|u|$ and n large enough, α_i (and α_j) satisfying (14). So far we have not used the assumption that (h, g) has unbounded balance on L. Now we do so. We fix the decomposition $f^{n_0}(a)=uv$ requiring that n_0 is large enough to yield (14) and that the balance

$\beta_{h,g}(u)$ is different from the balances of the prefixes of $\{f^n(a) \mid n<n_0\} \cup \{u\}$, i. e., for any such prefix $w \neq u$, $|\beta_{h,g}(w)| \neq |\beta_{h,g}(u)|$. Observe here that we have two possibilities: either $|u| \leqq |v|$ (handled in detail above) or $|u| \geqq |v|$ (which is symmetric). Observe also that the above is the only point which makes n_0 dependent on (h, g). However, by Lemma 6.2 and the considerations of the beginning of the proof of Theorem 5.1, there exists a uniform upper bound for n_0. Consequently, n_0 can be after all chosen independently of (h, g). We further assume that $n_0 \geqq |\Sigma|$.

Now we set $F' = \{f^n(a) \mid n \leqq n_0\}$ and recall our assumption: $h \overset{F'}{\equiv} g$. We have:

$$f^{n_0}(a) = u_1 u' v' v_1,$$

$$f^{n_0-j+i}(a) = u_2 u' v' v_2,$$

where $u_1 u' = u$, $u' v' = f^i(\alpha_i \beta)$. The choice of α_i and α_j can be illustrated as in Figure 1.

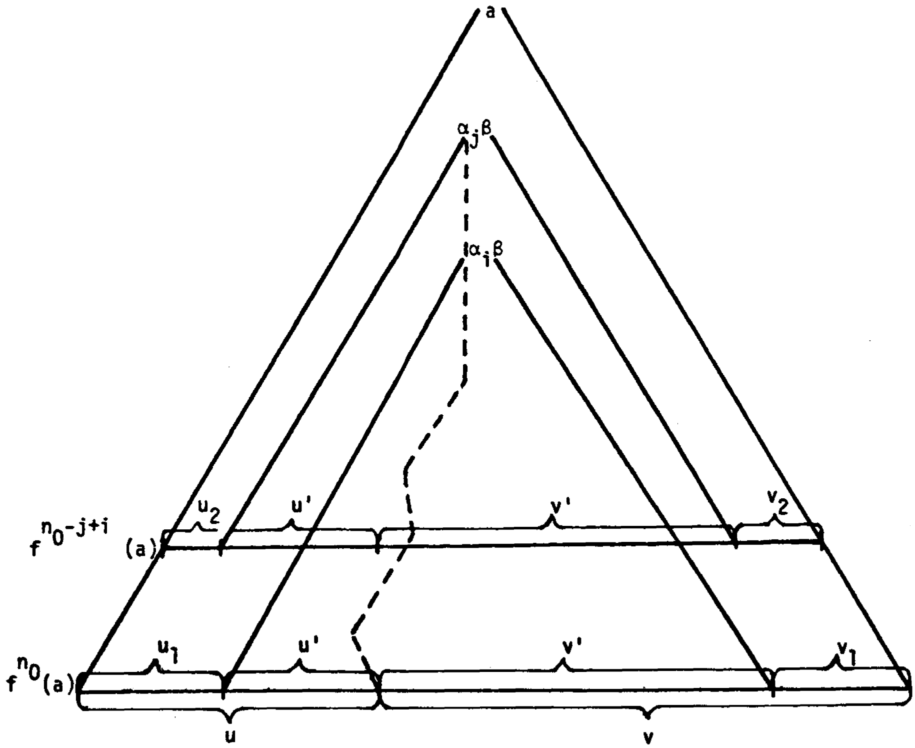

Figure 1

Since the above specified α is in $f^i(\alpha_i)$, $|f^i(\beta)| \leqq |v'|$.

So using (10), (14) and (9) we deduce:

$$\left.\begin{aligned}&|h(v')| \geqq |h(f^i(\beta))| = |h(f^{i-k}(f^k(\beta)))|\\ &\geqq 2|\beta_{\max}| + |h(f^{i-k}(f(\beta^{K'})))| \geqq 2|\beta_{\max}| + K'|h(f^{i-k}(f(\beta))|\\ &\geqq 2|\beta_{\max}| + \max\{|h(f^{i-k}(f(\beta)))|, |g(f^{i-k}(f(\beta)))|\},\end{aligned}\right\} \tag{15}$$

and that the same holds true when h and g are interchanged.

By our assumption $h(f^{n_0}(a)) = g(f^{n_0}(a))$. Therefore since $uv' \in \mathrm{pref}(f^{n_0}(a))$ there exists a word y such that $yh(v')$ Pref $g(v')$ or $h(v')$ Pref $yg(v')$ with $|y| = |\beta(u)|$. Similarly, since $h(f^{n_0-j+i}(a)) = g(f^{n_0-j+i}(a))$ there exists a word y' such that either $y'h(v')$ Pref $g(v')$ or $h(v')$ Pref $y'g(v')$ with $|y'| = |\beta(u_2 u')|$. Moreover, by the choice of $|\beta(u)|$, $|y| \neq |y'|$. Consequently, we have the situation illustrated in Figure 2 (where we assume that $h(v')$ pref $yg(v')$ and $y'h(v')$ pref $g(v')$; the other three possibilities can be handled with the very same manner).

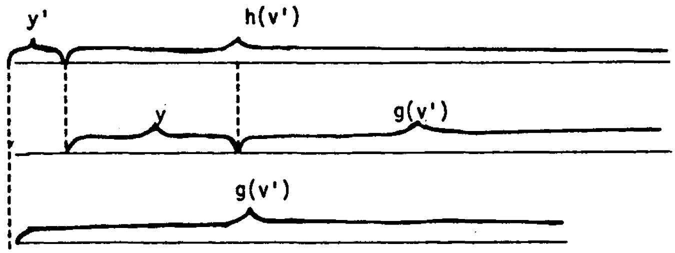

Figure 2

That is to say, we have three representations for a prefix of $h(v')$. Consequently, the prefix $\bar{w}$ of $h(v')$ with the length:

$$\min\{|h(v')|, |g(v')| - |\beta_{\max}|\} \tag{16}$$

is quasiperiodic with the period $p = yy'$, i.e., $\bar{w} \in \mathrm{pref}(p^*)$. Possibly by choosing p shorter we may assume that p is primitive, *cf.* [9].

Now let:

$$c_i = f^{i-k}(c) \qquad \text{for each} \quad c \in \Sigma.$$

By (14):

$$|h(c_i)| \geqq 2|\beta_{\max}| \geqq |p|. \tag{17}$$

Let $L_2=\{cd\in\Sigma^2 \mid cd\in\text{sub}(L(G))\}$. We claim that $h(c_i d_i)\in\text{sub}(p^*)$ for every $cd\in L_2$. Now, by (15), its symmetric form for g, (16) and the fact $L_2\subseteq\text{sub}(f(\beta))$ we conclude that $h(c_i d_i)\in\text{sub}(\bar{w})$ for every $cd\in L_2$. Thus $h(c_i d_i)\in\text{sub}(p^*)$. Now, by (17) and the primitiveness of p, we conclude that $h(f^{i-k}(y))\in\text{sub}(p^*)$ for every word y in Σ^* such that $\text{sub}_2(y)\subseteq L_2$. In particular:

$$h(f^{i-k}(f^n(a)))\in\text{sub}(p^*) \qquad \text{for} \quad n\geqq 0. \tag{18}$$

Symmetrically, we find a primitive word p' such that:

$$g(f^{i-k}(f^n(a)))\in\text{sub}(p'^*) \qquad \text{for} \quad n\geqq 0. \tag{19}$$

So, by the primitiveness of p and p' and by the fact $h(f^{i-k}(a))=g(f^{i-k}(a))$, we must have $p=p'$.

Finally, we are ready to finish the proof of Theorem 6.1 in the case of one letter axiom. Since $h\overset{F'}{\equiv}g$ and $n_0\geqq|\Sigma|$ we, by Lemma 3.3, conclude that $h\overset{L}{\equiv}_l g$. Moreover, $n_0\geqq|\Sigma|$ implies that if L contains a word starting with some letter in Σ, then also F' contains such a word. Consequently, (18) and (19) guarantee that $h\overset{L}{\equiv}g$.

The proof for the general case, i. e., for the case when x need not be of length one, is obtained as a modification of the above in the following way. Let:

$$L_2=\{cd\in\Sigma^2 \mid cd\in\text{sub}(L(G))\}$$

and:

$$L'_2=\{cd\in\Sigma^2 \mid cd\in\text{sub}(\bigcup_{a\in\Sigma}\{f^n(a) \mid n\geqq 0\})\}.$$

Now, we cannot require that, for each $b\in\Sigma$, $f(b)$ contains as subwords all words from L_2, but we can require, as we did, that this is true for words from L'_2. Hence, by the arguments above, there exists a primitive word p such that:

and:
$$\left.\begin{aligned} &h(f^{i-k}(f^n(b)))\in\text{sub}(p^*),\\ &g(f^{i-k}(f^n(b)))\in\text{sub}(p^*)\end{aligned}\right\} \tag{20}$$

for all $n\geqq 0$ and $b\in\Sigma$.

Let $x=a_1\ldots a_r$ with $a_j\in\Sigma$. As in the case $x=a$, we have $h\overset{F'}{\equiv} g$ and $h\overset{L}{\equiv}_l g$, and we should show that $h\overset{L}{\equiv} g$. This follows if we show that:

$$h(f^n(a_1\ldots a_j))\ \mathrm{Pref}\ g(f^n(a_1\ldots a_j)) \tag{21}$$

for $n\geqq 0$ and $j=1,\ldots,r$.

Let us consider (21) for $j=2$. We define, for $n\geqq 0$, $a_0(n)=\mathrm{pref}_1(f^n(a_1))$, $a_1(n)=\mathrm{suf}_1(f^n(a_1))$ and $a_2(n)=\mathrm{pref}_1(f^n(a_2))$. Clearly, the sequence $(a_0(n), a_1(n), a_2(n))_{n\geqq 0}$ is periodic, i.e., for some integers τ and ρ the following holds:

$$a_i(\tau+l+m\,\rho)=a_i(\tau+l+(m+1)\,\rho), \tag{22}$$

for $i=0, 1, 2$, $l=0,\ldots,\rho-1$ and $m\geqq 0$. We fix l and show that (21) holds for $n=\tau+l+\rho\,m$ with $m\geqq 0$.

For notational convenience let $f^{\tau+l+m\rho}(a_1)=\gamma(m)$ and $f^{\tau+l+m\rho}(a_2)=\delta(m)$. For $\tau+l+m\,\rho\leqq i-k$ we are done: the required equation is among our assumptions. So let m assume only values such that $\tau+l+m\,\rho\geqq i-k$. Observe that, by (20) and (22):

$$\left.\begin{aligned} h(\gamma(m))&\in p_1 p^* p_2,\\ h(\delta(m))&\in p_2'\ \mathrm{pref}(p^*),\\ g(\gamma(m))&\in p_1 p^* p_3,\\ g(\delta(m))&\in p_3'\ \mathrm{pref}(p^*),\end{aligned}\right\} \tag{23}$$

for some words $p_1, p_2', p_3'\in\mathrm{suf}(p)$ and $p_2, p_3\in\mathrm{pref}(p)$.

Now, we assume that $\beta_{h,g}(\gamma(m))$ assumes at least two different values, say $\beta_{h,g}(\gamma(m_1))\neq\beta_{h,g}(\gamma(m_2))$. Because $(\beta_{h,g}(\gamma(m)))_{m\geqq 0}$ is governed by a difference equation of order t, we may, possibly enlarging n_0, assume that $m_1, m_2\leqq n_0$. By (23), $|\beta_{h,g}(\gamma(m_1))-\beta_{h,g}(\gamma(m_2))|$ is a multiple of $|p|$. Let:

$$\mathrm{suf}_{|\beta_{\max}|}(h(\gamma(m_1)))=\gamma_1=\mathrm{suf}_{|\beta_{\max}|}(h(\gamma(m_2)))$$

and:

$$\mathrm{pref}_{|\beta_{\max}|}(h(\delta(m_1)))=\delta_1=\mathrm{pref}_{|\beta_{\max}|}(h(\delta(m_2))).$$

Since $h(f^{m_1}(x))=g(f^{m_1}(x)), h(f^{m_2}(x))=g(f^{m_2}(x))$ and $|\beta_{h,g}(f^{m_i}(a_1))|\leqq|\beta_{\max}|$, for $i=1, 2$, we have, by (23), the situation illustrated in Figure 3 (where we assume that $\beta_{h,g}(\gamma(m_1)) \geqq 0$ and $\beta_{h,g}(\gamma(m_2)) \leqq 0$; the other cases are similar):

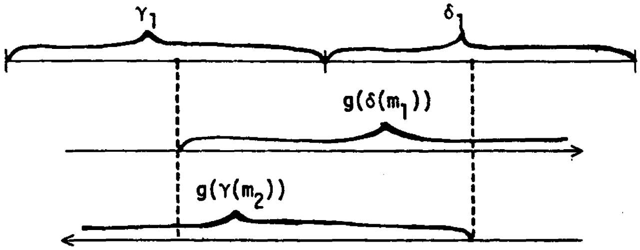

Figure 3

So it follows from (23), from the primitiveness of p and from the fact that $|\beta_{h,g}\gamma(m_1) - \beta_{h,g}\gamma(m_2)|$ is larger than $|p|$ that $p_3 p'_3 = p$. It also follows from Figure 3 that $p_2 p'_2 = p$. Consequently, by (23), the equation (21) follows in this case.

The other possibility, i. e., the case when $\beta_{h,g}(\gamma(m))$ assumes only one value is simpler. Clearly, (21) now follows from (23) and from the fact that $h(\gamma(m)\,\delta(m)) = g(\gamma(m)\,\delta(m))$ for some value of m, say m'.

Equation (21) for cases $j > 2$ can obviously be derived in the very same manner. Indeed, to prove (21) for some j, only the behaviour of h and g near the occurrences of subwords $\mathrm{suf}_1 f^n(a_{j-1})\,\mathrm{pref}_1 f^n(a_j)$ are needed. This, finally, completes our proof for Theorem 6.1.

7. TEST SETS FOR POSITIVE DOL LANGUAGES

Now, we are ready for our main result concerning DOL languages.

THEOREM 7.1: *Every positive* DOL *language* L *possesses a test set. Moreover, a test set for* L *can be effectively found.*

Proof: Let F and F' be subsets of L determined by Theorems 5.1 and 6.1. Clearly, $F \cup F'$ is a test set for L proving the first sentence of Theorem 7.1. The second sentence follows from Theorem 3.2 in [5], which shows that if a test set for a DOL language exists it can be effectively found.

In order to be able to state a corollary of Theorem 7.1 we need the following definition. Let $\mathscr{L}$ be a family of languages. *Morphism equivalence problem* for $\mathscr{L}$ is to decide whether two given morphisms agree string by string on a given language of $\mathscr{L}$.

COROLLARY 7.1: *Morphism equivalence problem for positive* DOL *languages is decidable.*

Proof : Immediate by Theorem 7.1.

As regards possibilities to generalize the above the following remark is in order. Let L be a positive DOL language and $(h, g) \in \mathscr{H}(L)$. By the proof of Theorem 6.1, either (h, g) agree on L with bounded balance or there exists a constant i [independent of (h, g)] and a word p such that:

$$h(f^n(b)), g(f^n(b)) \in \operatorname{sub}(p^*) \qquad \text{for} \quad n \geqq i \quad \text{and} \quad b \in \Sigma, \tag{1}$$

i. e., h and g are, in a sense, "very periodic on L". This is not true for arbitrary DOL languages as seen from

Example 7.1: Let G be the DOL system defined by the morphism:

$$f: \begin{cases} a \to abc, \\ b \to bb, \\ c \to c, \\ d \to d, \\ e \to ee, \\ f \to cef, \end{cases}$$

and the axiom abdef. Further let h and $g : \{a, b, c, d, e, f\}^* \to \{1, 2, 3, 4, 5\}^*$ be the morphisms defined by:

$$h: \begin{cases} a \to 1234, \\ b \to 2323, \\ c \to 4, \\ d \to 24, \\ e \to 32, \\ f \to 5, \end{cases} \qquad g: \begin{cases} a \to 1, \\ b \to 23, \\ c \to 4, \\ d \to 42, \\ e \to 3232, \\ f \to 4325. \end{cases}$$

It is straightforward to see that $h \overset{L(G)}{\equiv} g$, *cf.* [10]. It is also clear that (1) is not satisfied for G, h and g. However, (h, g) has unbounded balance on $L(G)$. In fact, for each $w \in L(G)$:

$$\beta_{h,g}(\operatorname{pref}_{(1/2)|w|-1}(w)) \geqq \frac{1}{3}|w|.$$

On the other hand, we believe that our considerations can be generalized to cover all simple DOL languages, *cf.* [3], i. e. languages generated by DOL systems satisfying: for each pair (a, b) of letters a is generated from b in a number of steps. Indeed, we have:

THEOREM 7.2: *Each simple* DOL *language containing a word of the length one has effectively a test set.*

Proof : A DOL system generating such a language can be decomposed, *cf.* [13], into a finite number of positive DOL systems. We leave the details for the reader.

We conclude with a simple observation which somewhat extends our main result.

LEMMA 7.1: *If a test set (effectively) exists for each language from L than the same holds also for the morphic closure of L.*

Proof: Obvious.

COROLLARY 7.2: *Every* HDOL *language based on a positive* DOL *language possesses (effectively) a test set.*

ACKNOWLEDGEMENTS

The authors are very grateful to the anonymous referee for reading the manuscript carefully and making the comments which improved the terminology and presentation.

REFERENCES

1. J. ALBERT, K. CULIK II and J. KARHUMAKI, *Tests Sets for Context Free Languages and Systems of Equations Over a Free Monoid*, Information and Control, Vol. 52, 1982, pp. 172-186.
2. K. CULIK II, *The Ultimate Equivalence Problem for* DOL *Systems*, Acta Informatica, Vol. 10, 1978, pp. 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, *On the Decidability of the Sequence Equivalence Problem for* DOL *Systems*, Theor. Comp. Science, Vol. 3, 1977, pp. 75-84.
5. K. CULIK II and J. KARHUMAKI, *Systems of Equations Over a Free Monoid and Ehrenfeucht's Conjecture*, Discrete Mathematics, Vol. 43, 1983, pp. 139-153.
6. K. CULIK II and A. SALOMAA, *On the Decidability of Homomorphism Equivalence for Languages*, J. Comput. Systems Sc., Vol. 17, 1978, pp. 163-175.
7. K. CULIK II and A SALOMAA, *Test Sets and Checking Words for Homomorphism Equivalence*, J. Comput. Systems Sc., Vol. 21, 1980, pp. 379-395.
8. S. EILENBERG and M. P. SCHÜTZENBERGER, *Rational Sets in Commutative Monoids*, J. of Algebra, Vol. 13, 1969, pp. 173-191.
9. M. A. HARRISON, *Introduction to Formal Language Theory*, Addison-Wesley, Reading, Massachusetts, 1978.
10. J. KARHUMÄKI and I. SIMON, *A Note on Elementary Homomorphisms and the Regularity of Equality Sets*, EATCS Bulletin, Vol. 9, 1979, pp. 16-24.
11. M. KARPINSKI, Ed., *New Scottish Book of Problems*, in preparation.
12. A. MANDEL and I. SIMON, *On Finite Semigroups of Matrices*, Theor. Comp. Science, Vol. 5, 1977, pp. 101-111.
13. G. ROZENBERG and A. SALOMAA, *The Mathematical Theory of L Systems*, Academic Press, New York, 1980.
14. A. SALOMAA and M. SOITTOLA, *Automata-Theoretic Aspects of Formal Power Series*, Springer Verlag, New York, 1978.