JOURNAL OF COMPUTER AND SYSTEM SCIENCES 50, 11–19 (1995)

Polynomial Size Test Sets For Context-Free Languages*

JUHANI KARHUMÄKI

Department of Mathematics, University of Turku, SF-20500 Turku 50, Finland

AND

WOJCIECH PLANDOWSKI AND WOJCIECH RYTTER

Institute of Informatics, Warsaw University, ul. Banacha 2, 02-097 Warsaw, Poland

Received February 8, 1992; revised October 21, 1993

We prove that each context-free language possesses a test set of size $O(m^6)$, where m is the number of productions in a grammar-producing the language. A context-free grammar generating the test set can be found in polynomial time by a sequential algorithm. It improves the doubly exponential upper bound from [2] and single exponential one from J. Karhumäki, W. Rytter, and S. Jarominek (*Theoret. Comput. Sci.* **116** (1993), 305–316). On the other hand, we show that the lower bound for the problem is $\Omega(m^3)$ and that the lower bound for the size of a test set for a language defined over n-letter alphabet is $\Omega(n^3)$.

1. INTRODUCTION

For a long time the Ehrenfeucht conjecture has been one of the most important open problems in formal language theory. It states that every language over a finite alphabet has a finite test set. A test set T of a language L is a set of words being a subset of L which has the property that for each pair of morphisms (h, g) if they agree on every word of T (i.e., for every $x \in T$ we have $h(x) = g(x)$) then they agree on every word of L. The conjecture was proved to be true independently by Albert and Lawrence (see [3]) and by Guba (see comments to a Russian edition in [11]). However, the question about the size of a test set (measured in the number of words) remained open. Research went in two directions.

First, researchers looked for a dependence between the size of a test set and the size of alphabet. It can be easily seen that each nonempty language over a one-letter alphabet has a test set consisting of a single word. Every language over two-letter alphabet possesses a test set containing at most three elements (see [4]). Up to now there are no results for a three-letter alphabet. An obvious lower bound for the problem for an n-letter alphabet is n. We improve it to $\Omega(n^3)$.

The second approach was to find a dependence between the size of a test set and the size of a device generating a language. Suppose, for example, the language is defined by a grammar with m productions of a bounded length. In case of regular languages the known lower and upper bounds for the problem are of order $\Theta(m)$ (see [9]). For context-free languages, the known upper bounds were double exponential (see [2]) and single exponential (see [9]). Almost all of our paper is devoted to improve it to $O(m^6)$. We also show a $\Omega(m^3)$ lower bound for the problem.

For a review of results dealing with the Ehrenfeucht conjecture we refer the reader to [6, 7].

Our paper has the following structure. We start, in Section 3, by searching for a linear context-free language which is a test set for a given context-free language. Next, in Section 4, using a technical key lemma, we prove that every linear context-free language possesses a test set of size $O(m^6)$. From the transitivity of the relation of "being a test set" we obtain our main result. Section 5 contains a long and complicated proof of the key lemma. Difficulties in the proof are understandable in view of the fact that the key lemma from [2], whose proof in [2] spanned approximately seven pages, follows from a version of our key lemma. The last section is devoted to finding the lower bounds mentioned above.

2. PRELIMINARIES

We assume that the reader is familiar with the basic notions of formal language theory and combinatorics on words (see [5, 10]), in particular with the notions of free group, free monoid, and morphism.

All morphisms we deal with in the paper are from a free monoid Σ^* into a free monoid Δ^* where Σ, Δ are finite

* A preliminary version of this paper was presented at ICALP'92 (see [8]).

571/50/1-2

0022-0000/95 $6.00

alphabets. Similarly all the languages considered are subsets of Σ^*.

We start with a formal definition of the basic notion in our considerations, the notion of a test set.

DEFINITION. A set of words T is called a test set for the language L iff T is a subset of L and for every pair of morphisms h and g we have

$$(\forall x \in T, h(x) = g(x)) \quad \Leftrightarrow \quad (\forall x \in L, h(x) = g(x)). \quad (1)$$

Note that the implication from right to left in the definition always holds, since T is a subset of L.

Recall that two morphisms h and *agree* on a language L iff for all x in L $h(x) = g(x)$. Hence condition (1) in the definition is equivalent to

$$h, g \text{ agree on } T \quad \Rightarrow \quad h, g \text{ agree on } L.$$

We define a grammar by a 4-tuple (N, Σ, P, S), where N is a set of nonterminals, Σ is a terminal alphabet, P is a set of productions, and S is a starting symbol. Recall that each production in a linear context-free grammar has at most one nonterminal on its left-hand side and therefore it is either in form $A \to xBy$, or in form $A \to x$, where A, B are nonterminals and x, y are words over the terminal alphabet. Let $L(G)$ be a language generated by a grammar G.

In our paper a crucial role is played by the language L_4 generated by linear context-free grammar G_4 which is given by

$$\begin{aligned} A_4 &\to a_4 A_3 \bar{a}_4, & A_4 &\to b_4 A_3 \bar{b}_4 \\ A_3 &\to a_3 A_2 \bar{a}_3, & A_3 &\to b_3 A_2 \bar{b}_3 \\ A_2 &\to a_2 A_1 \bar{a}_2, & A_2 &\to b_2 A_1 \bar{b}_2 \\ A_1 &\to a_1 \bar{a}_1, & A_1 &\to b_1 \bar{b}_1. \end{aligned}$$

The convention is that the capital letters here are nonterminals and the lower case letters are terminals. The starting nonterminal is A_4, so that the language L_4 consists of 16 words. We consider also languages L_3, L_2 which are produced by the same grammar but the starting nonterminals are A_3, A_2, respectively. They contain 8 and 4 words, respectively. Denote by T_4 the language L_4 without the word $b_4 b_3 b_2 b_1 \bar{b}_1 \bar{b}_2 \bar{b}_3 \bar{b}_4$. Similarly, we define languages T_3, T_2 as languages L_3, L_2 without words $b_3 b_2 b_1 \bar{b}_1 \bar{b}_2 \bar{b}_3$, $b_2 b_1 \bar{b}_1 \bar{b}_2$, respectively.

The following lemma dealing with the language L_4 is the key point in our considerations. We postpone its proof to Section 5; however, we will use it earlier to find test sets for linear context-free languages.

LEMMA 1. *T_4 is a test set for L_4.*

Basic tools in the proof of this lemma are combinatorial properties of words.

3. TEST SETS FOR GENERAL CONTEXT-FREE LANGUAGES

In this section we present a method of finding test sets for context-free languages. To simplify the construction we assume that a language is given by a grammar in Chomsky normal form which is reduced, i.e., without useless nonterminals. This, however, is not essential, since each context-free grammar containing m productions of bounded length can be transformed into Chomsky normal form grammar having $O(m)$ productions. The test set we produce is given by a linear context-free grammar.

We use the following simple fact to prove the main result of this section:

FACT 1. *If two morphisms agree on words $u_1 v_1$, $u_2 v_1$, $u_1 v_2$ then they agree on the word $u_2 v_2$.*

Proof. Take two morphisms g, f. Let $\mathrm{FG}(A)$ be a free group generated by an alphabet A. Morphisms g, f can be uniquely extended to morphisms g', f': $\mathrm{FG}(\Sigma) \to \mathrm{FG}(\Delta)$. Note that $w = u_2 v_2 = u_2 v_1 (u_1 v_1)^{-1} u_1 v_2$ belongs to a subgroup of $\mathrm{FG}(\Sigma)$ generated by $u_2 v_1, u_1 v_1, u_1 v_2$. Hence $g'(w) = f'(w)$ and, finally, $g(w) = f(w)$. ∎

Now, take a context-free language L generated by a context-free grammar $G = (N, \Sigma, P, S)$ in Chomsky normal form.

For each nonterminal A let w_A be one of the shortest terminal words which can be generated in the grammar G from the symbol A. Let P' be a set of productions produced from the set P by replacing each production of the form $A \to BC$, where B and C are nonterminals, by three productions $A \to w_B C$, $A \to B w_C$, $A \to w_B w_C$.

Let $\mathrm{lin}(G) = (N, T, P', S)$. Clearly, $\mathrm{lin}(G)$ is a linear context-free grammar. We will prove that the language L_{lin} generated by this grammar is a test set for L.

In our considerations the *height* of a node in the derivation tree is its distance (the number of vertices) from the root, the height of the root being zero. Let L_d be the language consisting of all words x in L which has a derivation tree such that all productions corresponding to nodes at height at most d are in grammar $\mathrm{lin}(G)$ and all other productions are in $\mathrm{lin}(G)$ or G. Thus $L_0 = L$.

For a language X denote by $\mathrm{rest}_i(X)$ a set of all words in X whose length is at most i.

LEMMA 2. (a) *$\mathrm{rest}_i(L_{d+1})$ is a test set for $\mathrm{rest}_i(L_d)$, for $i \geq d+1 \geq 1$*

(b) $\mathrm{rest}_i(L_{\mathrm{lin}})$ is a test set for $\mathrm{rest}_i(L)$, for $i \geq 1$.

Proof. (a) Assume that morphisms h, g agree on $\mathrm{rest}_i(L_{d+1})$. Let z be a word belonging to $\mathrm{rest}_i(L_d) \setminus \mathrm{rest}_i(L_{d+1})$. Then z is in form $uvxw$ and it has a derivation tree of the form presented in Fig. 1. The node corresponding to A in this figure has height d. Consequently, productions

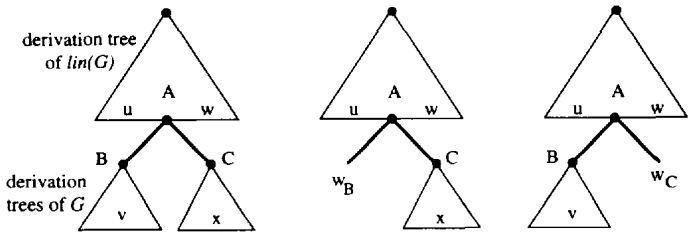

FIG. 1. The derivation trees for $z = uvxw$, uw_Bxw, uvw_Cw.

used to generate uAw are in $\mathrm{lin}(G)$, and so the language $\mathrm{rest}_i(L_{d+1})$ contains the words

$$uw_Bxw,\ uvw_Cw,\ uw_Bw_Cw.$$

Since morphisms h, g agree on these words, they agree, by Fact 1, on $z = uvxw$ as well.

(b) It follows from (a) that $\mathrm{rest}_d(L_d)$ is a test set for $\mathrm{rest}_d(L) = \mathrm{rest}_d(L_0)$. Now, since G is in Chomsky normal form every derivation tree of G generating a word of length at most d contains only internal nodes of heights at most d. Hence, $\mathrm{rest}_d(L_d) = \mathrm{rest}_d(L_{\mathrm{lin}})$ and the result follows. ∎

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

LEMMA 3. *Let G be a context-free grammar in Chomsky normal form. Then $L_{\mathrm{lin}} = L(\mathrm{lin}(G))$ is a test set for $L(G)$.*

Proof. We use the following property of test sets: Assume that we are given two sequences of languages H_i and W_i, such that for each i the language H_i is a test set for W_i. Then the union of the languages H_i is a test set for the union of the languages W_i.

Taking $H_i = \mathrm{rest}_i(L_{\mathrm{lin}})$ and $W_i = \mathrm{rest}_i(L)$ and applying Lemma 2 we obtain the thesis of the lemma. ∎

4. TEST SETS FOR LINEAR CONTEXT-FREE LANGUAGES

A linear context-free grammar $G = (N, \Sigma, P, S)$ can be viewed as a multigraph named graph(G), whose nodes correspond to nonterminal symbols and whose edges correspond to productions. One special node t called the *sink* is added. The graph has also one distinguished *source node* S corresponding to the starting nonterminal. The edges of the graph are derived in the following way from the set of productions P:

- each production of the form $A \to xBy$ corresponds to an edge from A to B labelled by a pair of terminal words (x, y),
- each production of the form $A \to x$ corresponds to an edge from A to t labelled by (u, v), where u, v are defined by any factorization $x = uv$.

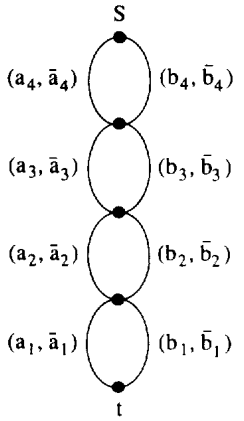

FIG. 2. graph(G_4).

Many distinct edges leading from A to B are possible. Clearly, the size of the graph equals the size of the grammar. A graph graph(G_4) for the grammar G_4 is presented in Fig. 2.

With each path π of graph(G) we associate a word denoted by $[\pi]$ and defined as follows: if (x_1, y_1), (x_2, y_2), ..., (x_k, y_k) is a sequence of labels of consecutive edges of the path π, then $[\pi] = x_1x_2 \cdots x_ky_k \cdots y_2y_1$.

The following fact is an easy consequence of the definitions:

FACT 2. *$L(G) = \{[\pi]: \pi$ is a path from S to t in* graph(G)$\}$.

Let L be a linear context-free language generated by a linear context-free grammar G. For each node v of graph(G) construct a subgraph of graph(G) which forms a tree named tree(v) rooted at v. Its nodes are all nodes reachable from v. Contrary to graph(G), tree(v) is a simple graph; there is at most one directed edge between any pair of nodes. Let $\lambda = (v_1, w_1), (v_2, w_2), \ldots, (v_k, w_k)$ be any sequence of edges of graph(G), not necessarily forming a path, where v_i, w_i are endnodes of the edges. We associate with the sequence λ a

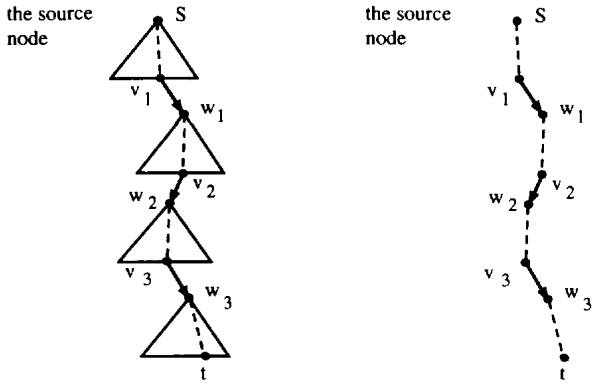

FIG. 3. The structure of the path corresponding to the sequence of three edges $(v_1, w_1), (v_2, w_2), (v_3, w_3)$.

path named path(λ) which starts at the starting node S and ends in the sink t. The path goes from S to v_1 by the unique path in tree(S); then it goes through the edge v_1, w_1), follows from w_1 to v_2 by the unique path in tree(w_1), traverses (v_2, w_2), and continues in tree(w_2). At the end it traverses (v_k, w_k) and goes in tree(v_k) in a unique way from w_k to the sink node t; see Fig. 3 for the case $k = 3$. Note that with each sequence of edges at most one path is associated, since in every tree there is at most one path from the root to a specified vertex. It can happen that for some sequence such a path is undefined. The edges of λ are called *special edges*

Recall that for any path π in the graph(G), $[\pi]$ denote the word corresponding to this path.

Let $F_k(G) = \{[\text{path}(\lambda)]$: λ is a sequence of at most k special edges in graph(G)$\}$

LEMMA 4. *Assume that G is a linear context-free grammar describing L(G). Then $F_6(G)$ is a test set for L(G).*

Proof. The main point is an application of the key Lemma 1. Assume that we are given a sequence $v_1, v_2, ..., v_5$ of five nodes and for each $i < 5$ we have two not necessarily distinct, subpaths from v_i to v_{i+1}. Let D be the multiset consisting of $2^4 = 16$ paths from v_1 to v_5 formed by combining these subpaths in all possible ways (if two distinct combinations give the same path then we consider the resulting path as two distinct copies of the same path). Let D' be a set resulting by removing a single path from D.

Take two morphisms h, g. Then the following claim is a direct consequence of Lemma 1.

CLAIM. *If h, g agree on all words corresponding to paths in D', then they agree on all words corresponding to paths in D as well.*

Proof. Since for each word w from $L(G)$ there is a finite number k such that $w \in F_k(G)$, it is enough to prove that $F_k(G)$ is a test set for $F_{k+1}(G)$ for $k \geqslant 6$.

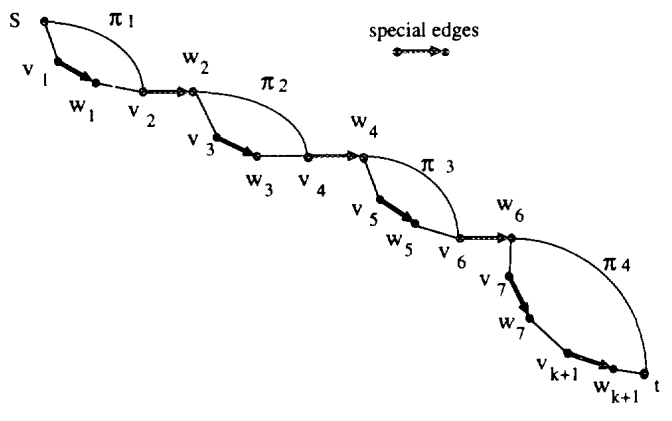

FIG. 4. All paths except the lowest one belong to $F_k(G)$.

The proof of the claim is illustrated in Fig. 4. Take any $k \geqslant 6$. Let π be the path determined by $k+1$ special edges $(v_1, w_1), (v_2, w_2), ..., (v_{k+1}, w_{k+1})$. Node S is connected to the vertex v_2, and, hence, there is a subpath π_1 in the tree(S) which starts at S and ends at v_2. In the same way we conclude that there are subpaths π_2 from w_2 to v_4 in tree(w_2), π_3 from w_4 to v_6 in tree(w_4), and π_4 from w_6 to the sink t in tree(w_6). The morphisms h and g agree on the words corresponding to paths containing at least one of subpaths $\pi_1, \pi_2, \pi_3, \pi_4$ since they are determined by at most k edges. For example, the path $\pi_1, v_2, w_2, v_3, w_3, v_4, w_4, \pi_3, v_6, w_6, \pi_4$ is associated with the sequence of four edges $(v_2, w_2), (v_3, w_3), (v_4, w_4), (v_6, w_6)$. Observe here that the same path can correspond to distinct sequences of edges. Consequently, by the claim, h and g agree on the word corresponding to π. ∎

Now we can state our main result.

THEOREM 3. *Assume that G is a context-free grammar in Chomsky normal form generating L. Then there is a test set F for L containing $O(m^6)$ words, where m is the number of productions in the grammar. A context-free grammar generating F can be constructed in polynomial time.*

Proof. By Lemma 3, $L(\text{lin}(G))$ is a test set of L. Now we use Lemma 4 to conclude that $F_6(\text{lin}(G))$ is a test set of $L(\text{lin}(G))$. Hence, $F_6(\text{lin}(G))$ is a test set of L. Since lin(G) contains at most $3m$ productions, $F_6(\text{lin}(G))$ contains at most $\sum_{i=0}^{6} (3m)^i$ words. Since context-free grammars producing words w_A can be constructed in polynomial time, we can construct a context-free grammar for $F_6(\text{lin}(G))$ in polynomial time as well. ∎

Remark. A brute force algorithm derivable from the proof of the theorem can be improved to produce a grammar containing $O(nm)$ productions and to work in $O(nm)$ worst-case time, where n is the number of nonterminals in the grammar. Observe here that a polynomial time algorithm generating all the words of a test set for a context-free language does not exist, because some context-free grammars describe languages in which the shortest word has exponential length with respect to the number of productions in a grammar.

5. PROOF OF THE KEY LEMMA

The proof is similar to the proof of the key lemma in [2]. Its idea is to show that certain systems of equations on words imply other equations. The key tool is the periodicity in words. Some properties of periodic and primitive words can be found in [10]. Below, we present the basic notions that we use in the proof.

Let $|x|$ be the length of a word x. We write $x \leqslant_p y$ ($x \leqslant_s y$) iff x is a prefix (respectively suffix) of y. The longest common prefix (suffix) of words x and y we denote by $x \wedge_p y$ ($x \wedge_s y$,

respectively). We say, that a word w is a *period* of a word x iff it is in form w^*p, where p is a prefix of w. If p is empty then the word w is called a *full period* of x. The shortest full period of x is denoted by $\mathrm{root}(x)$. Word x is said to be *primitive* iff x is nonempty and $\mathrm{root}(x) = x$. We denote the empty word by ε. A word x_1 is a *conjugate* of x_2 iff $x_1 = vu$, $x_2 = uv$ for some words u, v.

Let us fix morphisms h, g and write x'' instead of $h(x)$ and x' instead of $g(x)$ for each word x. The assumption of the key lemma is equivalent to the following two conditions:

$$a_4'' x'' \bar{a}_4'' = a_4' x' \bar{a}_4' \quad \text{for each } x \in L_3, \tag{2}$$

$$b_4'' x'' \bar{b}_4'' = b_4' x' \bar{b}_4' \quad \text{for each } x \in T_3. \tag{3}$$

Denoting $\mu = b_3 b_2 b_1 \bar{b}_1 \bar{b}_2 \bar{b}_3$, the key lemma can now be restated.

LEMMA 5. *If* (2) *and* (3) *hold then* $b_4'' \mu'' \bar{b}_4'' = b_4' \mu' \bar{b}_4'$.

Before we present the proof of the lemma we prove some facts and lemmas which are used in the proof. Our first fact is well known.

FACT 4. *Let $v, \bar{v}$ be fixed words. Then the equality $vx = x\bar{v}$ is equivalent to the following property of x: there exist words p, q such that $v = pq$, $\bar{v} = qp$, $x \in (pq)^*p$.*

In our paper we use a sharper version of this fact which can be formulated in the following way.

FACT 5. *Let $v, \bar{v}$ be fixed words. Then the equality $vx = x\bar{v}$ is equivalent to the following property of x: there exist unique words $p \neq \varepsilon, q$ such that $\mathrm{root}(v) = pq$, $\mathrm{root}(\bar{v}) = qp$, $x \in (pq)^*p$, and p, q are determined by v and $\bar{v}$ independently of x.*

Another fact can be as follows.

FACT 6. *Assume that w is primitive and that $|z| \geq |w|$. If w is a period of both ux and vx (not necessarily a full period) then there exists a prefix w_p of w such that $u, v \in w^*w_p$.*

Proof. Take the prefix x of z of length $|w|$. Then x is a primitive word due to the property that a conjugate of primitive word is also primitive. We can assume that u and v are proper prefices of w; otherwise we cut off some number of occurrences of w.

We may also assume that $|u| \geq |v|$, so that vx is a prefix of ux. It is enough to show that $u = v$. Assume that $u \neq v$. Then we have the situation presented in Fig. 5. Both x and y are periods of the indicated segment. Due to periodicity lemma, see [10], x has a period of size $\gcd(|x|, |u| - |v|)$. Since x is primitive this is possible only if $|u| - |v| = 0$. ∎

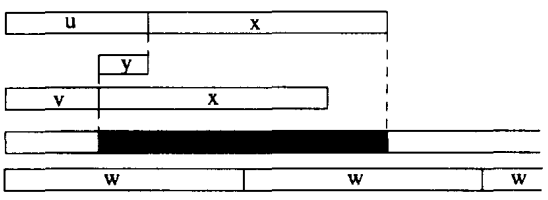

FIG. 5. Both x and y are periods of the indicated segment.

FACT 7. *Let x be a word satisfying conditions: $|x| \geq |pr|$ and $vx\bar{v}, wx\bar{w} \in p(rp)^*$. Then for each word z the following equivalence holds: $vz\bar{v} \in p(rp)^* \Leftrightarrow wz\bar{w} \in p(rp)^*$.*

Proof. If pr is primitive then the proof of this fact follows directly from Fact 6 and its dual formulation for suffices.

If rp is not primitive then we consider the root ρ of rp. Then $rp = \rho^j$ and $p = \sigma\rho^i$ for some i, j, and a proper suffix σ of ρ. Assuming the left-hand side of the proposed equivalence, we have for some $k, l, m \geq 0$:

$$vx\bar{v} = \sigma\rho^{i+jk}, \quad wx\bar{w} = \sigma\rho^{i+jl}, \quad vz\bar{v} = \sigma\rho^{i+jm}.$$

We also have that $wz\bar{w} \in \sigma\rho^*$, since ρ is primitive. However, $|wz\bar{w}| = |wx\bar{w}| + |vz\bar{v}| - |vx\bar{v}| = |\sigma\rho^i| + (k+m-l)\,|\rho^j|$, where $k + m - l \geq 0$. Hence, $wz\bar{w} \in \sigma\rho^i(\rho^j)^* = p(rp)^*$. ∎

LEMMA 6. *Let $p, q, v, \bar{v}, w, \bar{w}, z$, and $\bar{z}$ be fixed words. Then*

(a) *if $x' \in (pq)^*p$ for each $x \in T_3$ then $\mu' \in (pq)^*p$;*

(b) *if $vx' = x'\bar{v}$ for each $x \in T_3$ then $v\mu' = \mu'\bar{v}$;*

(c) *if $wx'\bar{w} = zx'\bar{z}$ for each $x \in T_3$ then $w\mu'\bar{w} = \bar{z}\mu'\bar{z}$.*

Proof. It is enough to prove part (a). Two remaining parts follow from it easily. Part (b) follows from it and Fact 5. Part (c) is an easy consequence of (b). For example, if $|w| > |z|$ then we can write $w = zt$, $\bar{z} = \bar{t}\bar{w}$, and $tx' = x'\bar{t}$ for all $x \in T_3$ and, hence, part (b) become applicable.

To prove (a) assume that $x' \in (pq)^*p$ for each $x \in T_3$. Consider words $y_1 = a_2' b_1' \bar{b}_1' \bar{a}_2'$ and $y_2 = a_2' a_1' \bar{a}_1' \bar{a}_2'$.

Case 1. $y_1 = y_2$. In that case, $b_1'\bar{b}_1' = a_1'\bar{a}_1'$. Thus, $\mu' = b_3' b_2' b_1' \bar{b}_1' \bar{b}_2' \bar{b}_3' = b_3' b_2' a_1' \bar{a}_1' \bar{b}_2' \bar{b}_3' \in (pq)^*p$.

Case 2. $y_1 \neq y_2$. Assume $|y_1| > |y_2|$, so that $|y_1| > |pq|$. Let $\mu_1 = b_2' b_1' \bar{b}_1' \bar{b}_2'$. From the assumption of the lemma we have $a_3' y_1 \bar{a}_3'$, $b_3' y_1 \bar{b}_3'$, $a_3' \mu_1 \bar{a}_3' \in (pq)^*p$. By Fact 7 we get $\mu' = b_3' \mu_1 \bar{b}_3' \in (pq)^*p$. ∎

FACT 8. *Let γ, β, q be three words such that $|\gamma \wedge_s q| < |q|$ and $|q \wedge_p \beta| = 0$. Then for any distinct words x, y we have: If $vx\bar{v}$, $vy\bar{v}$, $wx\bar{w}$, $wy\bar{w}$, $vz\bar{v} \in \gamma q^*\beta$ then $wz\bar{w} \in \gamma q^*\beta$.*

Proof. Denote $p = \gamma \wedge_s q$. Take r, α such that $q = rp$, $\gamma = \alpha p$. Since $s \neq y$ we may assume that $|x| > |y|$ and, hence, $|x| - |y| \geq |pr|$. Consequently, $|x| \geq |pr|$. Let $s = vx\bar{v} \wedge_s vy\bar{v}$, $t = wx\bar{w} \wedge_s wy\bar{w}$, $f = vx\bar{v} \wedge_p vy\bar{v}$, $g = wx\bar{w} \wedge_p wy\bar{w}$. Clearly, $|s| \geq |\bar{v}|$, $|t| \geq |\bar{w}|$, $|f| \geq |v|$, $|g| \geq |w|$. Words $vx\bar{v}$, $vy\bar{v}$ are both in $\alpha(pr)^*p\beta$ and they have different lengths. By the choice of α word, s is in form $(pr)^*p\beta$. Similarly we obtain:

$t \in (pr)^* p\beta$, $f \in \alpha(pr)^* p$, $g \in \alpha(pr)^* p$. We distinguish two cases:

Case 1. $|s| < |y\bar{v}|$ or $|t| < |y\bar{w}|$ or $|f| < |vy|$ or $|g| < |wy|$. Assume $|t| < |y\bar{w}|$ (other cases are symmetric). There exist two nonempty words x_1, y_1 such that $x_1 t = x\bar{w}$, $y_1 t = y\bar{w}$ and the last letters of x_1, y_1 are different. From inequality $|t| \geq |\bar{w}|$ words x, y have a common suffix u such that $x_1 u = x$, x, $y_1 u = y$, and $t = u\bar{w}$. It follows that $u = x \wedge_s y$ and, hence, $s = u\bar{v}$.

Suppose first that s, t have different lengths. Since u is a common prefix of t and s, it follows from their form and the assumption $|q \wedge_p \beta| = 0$, that u is a prefix of the word in form $(pr)^* p$. Hence we can find a word k such that k is a proper suffix of rp and words $\bar{w}, \bar{v}$ are both in $k(rp)^*\beta$. Consequently, there exists a prefix n of pr such that vx, $wx \in \alpha(pr)^* n$ and x is a common suffix of them. Since the only common suffix of α and r is the empty word, we have either $v = w$ or v, $w \in \alpha(pr)^* m$, where m is a prefix of pr. In either case, if $vz\bar{v} \in \alpha(pr)^* p\beta$, then $wz\bar{w} \in \alpha(pr)^* p\beta$.

Assume that s, t have equal lengths. Now, $s = t$ and consequently $\bar{w} = \bar{v}$. Therefore, $y\bar{w}$ is a suffix of both $wy\bar{w}$ and $vy\bar{v}$, so that either $v = w$ or $y\bar{w}$ is a suffix of a word in $p(rp)^*\beta$. Since $|t| > |\beta|$ and $|y\bar{w}| > |t|$, we have $|y\bar{w}| > |\beta|$ and finally there is a word n such that v, $w \in \alpha(pr)^* n$. Hence $\bar{w} = \bar{v}$ and either $v = w$ or v, $w \in \alpha(pr)^* n$. In both cases, if $vz\bar{v} \in \alpha p(rp)^*\beta$, then $wz\bar{w} \in \alpha p(rp)^*\beta$.

Case 2. $|s| \geq |y\bar{v}|$ and $|t| \geq |y\bar{w}|$ and $|f| \geq |vy|$ and $|g| \geq |wy|$. The conditions for this case imply that $|\alpha| \leq |v|$, $|\alpha| \leq |w|$, $|\beta| \leq |\bar{v}|$, $|\beta| \leq |\bar{w}|$. Now v cannot be in $\alpha(pr)^* pd$, where d is a nonempty prefix of β because $|\beta| \leq |\bar{v}|$. Hence, $v \in \alpha(pr)^* h$, where h is a prefix of pr. Similarly, we obtain that $w \in \alpha(pr)^* l$, $\bar{v} \in m(rp)^*\beta$, $\bar{w} \in n(pr)^*\beta$, where l is a prefix of pr and n, m are suffices of rp. Thus hxm, $lxn \in (pr)^* p$ which, together with $|x| \geq |pr|$ and $hzm \in (pr)^* p$, imply that $lzn \in (pr)^* p$. It follows from Fact 7. Consequently, $wz\bar{w} \in \alpha(pr)^* p\beta$. ∎

FACT 9. *Let S be set of pairs of words (x, y) consisting of at least two elements and u, z, $\bar{u}$, $\bar{z}$ be fixed words, such that $u \neq z$. If each pair (x, y) of words in S satisfies the equalities $zx = y\bar{z}$ and $ux = y\bar{u}$, then one can find words α, q, β such that*

(a) zx, $y\bar{z}$, ux, $y\bar{u} \in \alpha q^*\beta$ for all pairs $(x, y) \in S$ and

(b) $|q \wedge_s \alpha| < |q|$ and $|\beta \wedge_p q| = 0$.

Proof. Case 1. $z \leqslant_p u$ and $\bar{z} \leqslant_s \bar{u}$. Let us prove first that there exist words α_1, β_1, q_1 which satisfy condition (a).

Let w, $\bar{w}$ be such that $u = zw$ and $\bar{u} = \bar{w}\bar{z}$. Assume that $w \neq \bar{w}$. If they are equal the proof is similar. Let ρ be the root of w and consider a factorization $\rho = pr$, where p, r are fixed later. Now, we take shortest words α_1, β_1 such that $zp \in \alpha_1(rp)^*$, $\bar{z} \in (rp)^*\beta_1$ and set $q_1 = rp$. Let (x, y) be a fixed pair from S.

If $|y| \geq |z|$ then there is a word t satisfying the equality $y = zt$. Consequently, our assumptions yield the equalities $x = t\bar{z}$ and $wt = t\bar{w}$. Now, we apply Fact 5 to the latter identity and fix our factorization: $\mathrm{root}(w) = pr$, $\mathrm{root}(\bar{w}) = rp$, and $t \in (pr)^* p$. Since $w \neq \bar{w}$, p and r are nonempty words. Hence, $y \in z(pr)^* p = \alpha_1 q_1^*$ and $x \in (pr)^* p\bar{z} = pq_1^*\beta_1$ and, consequently, $y\bar{z}$, zx, $y\bar{u}$, $ux \in z(pr)^* p = \alpha_1 q_1^*\beta_1$, as required.

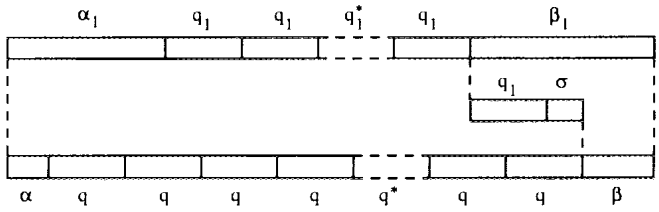

FIG. 6. Construction of the words α, q, β. $|q| = |q_1|$, $\sigma\beta \wedge_p q_1 = \sigma$, $|\alpha \wedge_s q| < |q|$.

If $|y| < |z|$ then there exists a word t such that $yt = z$. Hence, $tx = \bar{z}$ from the second equality and $tw = \bar{w}t$ from the first one. By Fact 5 and $p \neq \varepsilon$ we get $t \in (rp)^* r$ and, therefore $z \in yr(pr)^*$ and $\bar{z} \in r(pr)^* x$. Consequently, the construction of α_1, β_1 yields $y \in \alpha_1 q_1^*$, $z \in \alpha_1 q_1^* r$, $rx \in q_1^*\beta_1$, $\bar{z} \in q_1^*\beta_1$. This implies the required result.

We proved that the words we found satisfy condition (a). If they do not satisfy condition (b) we choose α, β, q which satisfy $\alpha q^*\beta = \alpha_1 q_1^*\beta_1$ and the condition (b) as follows, see Fig. 6. We simply extend period q_1 as far as possible in β_1 and take the last $|q_1|$ letters to form the period q. A word to the right of this place is set to be β. Then we "go" with the period q from right to left until we cannot find whole q inside α_1. The word which is placed to the left of this position is set to be α. By construction, the chosen words satisfy the required conditions. This completes the proof of Case 1.

Case 2. $z \not\leqslant_p u$ or $\bar{z} \not\leqslant_s \bar{u}$. Assume the first possibility, the proof for the other being similar. We can find words w, u_1, z_1 such that $z = wz_1$, $u = wu_1$ and the first letters of u_1 and z_1 are different. Take any pair (x, y) from the set S. Then $wz_1 x = y\bar{z}$ and $wu_1 x = y\bar{u}$ implies that $w = y\bar{u} \wedge_p y\bar{z} = y(\bar{u} \wedge_p \bar{z})$. Thus y is unique for all pairs of words from S. But also x is unique for all pairs from S, contradicting the fact that S contains at least two elements. ∎

LEMMA 7. (a) *Let f be a morphism and q, γ, β be fixed words such that $|q \wedge_s \gamma| < |q|$ and $|\beta \wedge_p q| = 0$. Then we have the implication: $uf(x)\bar{u} \in \gamma q^*\beta$ for each $x \in T_3 \Rightarrow uf(\mu)\bar{u} \in \gamma q^*\beta$.*

(b) *Let z, $\bar{z}$, u, $\bar{u}$ be fixed words and $v = b_2 b_1 \bar{b}_1 \bar{b}_2$. Then $(zx'' = x'\bar{z}$ for each $x \in T_2$ and $ux'' = x'\bar{u}$ for each $x \in L_2) \Rightarrow zv'' = v'\bar{z}$.*

Proof. (a) Assume that $uf(x)\bar{u} \in \gamma q^*\beta$ for each $x \in T_3$. Consider words $y_1 = f(a_2 b_1 \bar{b}_1 \bar{a}_2)$, $y_2 = f(a_2 a_1 \bar{a}_1 \bar{a}_2)$. There are two cases:

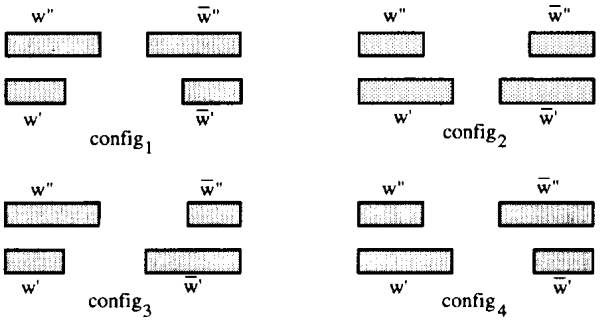

FIG. 7. The definition of configurations.

Case 1. $y_1 = y_2$. In that case $f(b_1\bar{b}_1) = f(a_1\bar{a}_1)$ and, hence, $uf(\mu)\bar{u} = uf(b_3b_2b_1\bar{b}_1\bar{b}_2\bar{b}_3)\bar{u} = uf(b_3b_2a_1\bar{a}_1\bar{b}_2\bar{b}_3)\bar{u} \in \alpha(pq)^*p\beta$.

Case 2. $y_1 \neq y_2$. Follows directly from Fact 8 with the choice $v = uf(a_3)$, $x = y_1$, $\bar{v} = f(\bar{a}_3)\bar{u}$, $w = uf(b_3)$, $y = y_2$, $\bar{w} = f(\bar{b}_3)\bar{u}$.

(b) If $u = z$ then $\bar{u} = \bar{z}$ and the result follows. So suppose that $u \neq z$. Define S at a set of words $\{(x'', x') : x \in T_2\}$. If S contains exactly one element then it is easy to prove that $(v'', v') \in S$ and we are done.

Assume that S contains at least two elements. Then we use Fact 9 to obtain zx'', $x'\bar{z}$, ux'', $x'\bar{u} \in \gamma q^*\beta$ for all $x \in T_2$. Now, set $u = \bar{u} = \varepsilon$ and define a morphism f by

$$f(a_3) = u, \quad f(b_3) = z, \quad f(\bar{a}_3) = \varepsilon, \quad f(\bar{b}_3) = \varepsilon,$$
$$f(t) = t'' \quad \text{for } t \in \{a_1, a_2, \bar{a}_1, \bar{a}_2, b_1, b_2, \bar{b}_1, \bar{b}_2\},$$

and apply part (a). Since $f(x) \in \gamma q^*\beta$ for all $x \in T_3$, we conclude that $f(\mu) = zv'' \in \gamma q^*\beta$.

Similarly, we see that $v'\bar{z} \in \gamma q^*\beta$. However, the lengths of zv'', $v'\bar{z}$ are equal, since $|zv''| = |uv''| - |ux''| + |zx''| = |v'\bar{u}| - |x'\bar{u}| + |x'\bar{z}| = |v'\bar{z}|$ for any $x \in T_2$. Thus, $zv'' = v'\bar{z}$, completing the proof. ∎

We are ready to proof the main lemma of this section.

Proof (of the Key Lemma). We define four conditions for a pair of words $w, \bar{w}$, see Fig. 7:

$\text{config}_1(w, \bar{w})$: $(w' \leqslant_p w'')$ and $(\bar{w}' \leqslant_s \bar{w}'')$;

$\text{config}_2(w, \bar{w})$: $(w'' \leqslant_p w')$ and $(\bar{w}'' \leqslant_s \bar{w}')$;

$\text{config}_3(w, \bar{w})$: $(w' \leqslant_p w'')$ and $(\bar{w}'' \leqslant_s \bar{w}')$;

$\text{config}_4(w, \bar{w})$: $(w'' \leqslant_p w')$ and $(\bar{w}' \leqslant_s \bar{w}'')$.

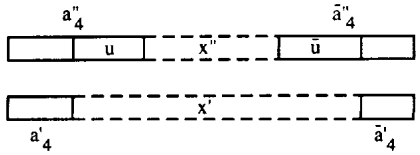

FIG. 8. Case 1 in the proof of the key lemma. Here x ranges over L_3.

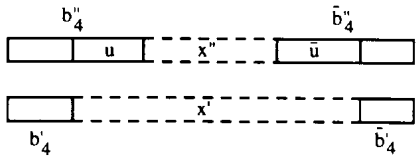

FIG. 9. Case 2 in the proof of the key lemma. Here x ranges over T_3.

The following crucial fact follows directly from the definitions:

FACT 10. *If the conditions* (2), (3) *are satisfied, then* $\text{config}_i(a_4, \bar{a}_4)$ *and* $\text{config}_j(b_4, \bar{b}_4)$ *hold for some* $1 \leqslant i \leqslant 4$, $1 \leqslant j \leqslant 4$.

Proof. Assume that conditions (2), (3) are satisfied. The proof of the key lemma proceeds by considering several cases depending on integers i, j in the fact above. Fortunately several pairs of them are symmetric. Our goal is to prove that $b''_4\mu''\bar{b}''_4 = b'_4\mu'\bar{b}'_4$. There are essentially four cases to be considered.

Case 1. $\text{config}_1(a_4, \bar{a}_4)$ holds. We have the situation presented in Fig. 8, where, by (2), $x' = ux''\bar{u}$ for each x in L_3, in particular, $\mu' = u\mu''\bar{u}$. Therefore, it is enough to show that $b''_4\mu''\bar{b}''_4 = b'_4u\mu''\bar{u}\bar{b}'_4$. Now, by (3), we known that $b''_4x''\bar{b}''_4 = b'_4x'\bar{b}'_4 = b'_4ux'\bar{u}\bar{b}'_4$. for each $x \in T_3$ so that Lemma 6(c) yields the required identity $b''_4\mu''\bar{b}''_4 = b'_4u\mu''\bar{u}\bar{b}'_4$.

Case 2. $\text{config}_1(b_4, \bar{b}_4)$ holds. In this case we have the situation of Fig. 9, where $x' = ux''\bar{u}$ for each x in T_3. By (2) we also have that $a''_4x''\bar{a}''_4 = a'_4x'\bar{a}'_4$ for each $x \in L_3$, implying that $a''_4x''\bar{a}''_4 = a'_4ux''\bar{u}\bar{a}'_4$ for each $x \in T_3$ and that $a''_4\mu''\bar{a}''_4 = a'_4\mu'\bar{a}'_4$. Now, Lemma 6(c) yields the identity $a''_4\mu''\bar{a}''_4 = a'_4u\mu''\bar{u}''\bar{a}'_4$ which, together with the previous identity, implies that $\mu' = u\mu''\bar{u}$. This, in turn, yields the required identity $b''_4\mu''\bar{b}''_4 = b'_4\mu'\bar{b}'_4$ by the assumptions of this case.

Case 3. $\text{config}_3(a_4, \bar{a}_4)$ and $\text{config}_4(b_4, \bar{b}_4)$ hold simultaneously. We have the situation presented in Fig. 10; in particular, we have $zux'' = zx'\bar{u} = x''\bar{z}\bar{u}$ for each $x \in T_3$.

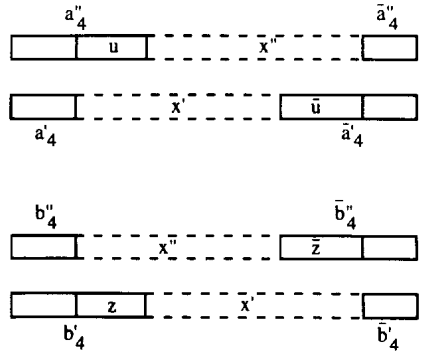

FIG. 10. Case 3 in the proof of the key lemma. Here in the first figure x ranges over L_3 and in the second one over T_3.

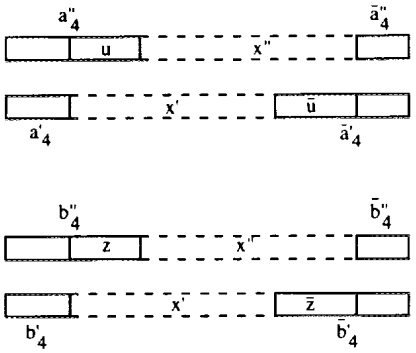

FIG. 11. Case 4 in the proof of the key lemma. Here in the first figure x ranges over L_3 and in the second one over T_3.

Now, Lemma 6 (b) implies the identity $z u \mu'' = \mu'' \bar{z} \bar{u}$ which, together with the equality $u\mu'' = \mu' \bar{u}$ (see assumption (2)), implies that $z\mu' \bar{u} = \mu'' \bar{z} \bar{u}$ or, equivalently, that $z\mu' = \mu'' \bar{z}$. Now, the required identity $b_4'' \mu'' \bar{b}_4'' = b_4' \mu' \bar{b}_4'$ follows from the assumption concerning $\text{config}_4(b_4, \bar{b}_4)$.

Case 4. $\text{config}_4(a_4, \bar{a}_4)$ and $\text{config}_4(b_4, \bar{b}_4)$ hold simultaneously. According to Fig. 11 we have the identity $ux'' = x'\bar{u}$ for all $x \in L_3$ and $zx'' = x'\bar{z}$ for all $x \in T_3$. Consequently, Lemma 7(b) yields $z\mu'' = \mu'\bar{z}$ and the required identity $b_4'' \mu'' \bar{b}_4'' = b_4' \mu' \bar{b}_4'$ follows directly from the assumption concerning $\text{config}_3(b_4, \bar{b}_4)$.

Two remaining cases are symmetric to the previous ones:

case "$\text{config}_2(a_4, \bar{a}_4)$" to Case 1, case "$\text{config}_2(b_4, \bar{b}_4)$" to Case 2;

case "$\text{config}_4(a_4, \bar{a}_4)$ and $\text{config}_3(b_4, \bar{b}_4)$ hold simultaneously" to Case 3;

case "$\text{config}_4(a_4, \bar{a}_4)$ and $\text{config}_4(b_4, \bar{b}_4)$ hold simultaneously" to Case 4.

This completes the proof of the key lemma as well as our main result. ∎

6. LOWER BOUND

In this section we present a family Ξ of finite languages which has the property that for each $L \in \Xi$ the only set for L is L itself. We start with a proof of an auxiliary lemma.

LEMMA 8. *T_3 is not a test set of L_3.*

Proof. It is enough to introduce two morphisms h and g which agree on T_3 and which do not agree on L_3, i.e., which do not agree on the word $\mu = b_3 b_2 b_1 \bar{b}_1 \bar{b}_2 \bar{b}_3$.

Let $\text{graph}(G_3)$ corresponding to language L_3 be as shown in Fig. 12. Then for a given morphism f, $\text{graph}_f(G_3)$ denotes a graph which is obtained from $\text{graph}(G_3)$ by replacing its each label $(x, \bar{x})$ by $(f(x), f(\bar{x}))$. It is straightforward to see that each path associated with a word w in $\text{graph}(G_3)$ has corresponding path in $\text{graph}_f(G)$ which is associated with a word $f(w)$.

Thus to check if two morphisms h and g agree on w, it is enough to compare words associated with appropriate paths in $\text{graph}_h(G_3)$ and $\text{graph}_g(G_3)$.

Now let h and g be morphisms for which $\text{graph}_g(G_3)$ and $\text{graph}_h(G_3)$ look as in Fig. 13. In other words, h and g are defined as

$$
\begin{aligned}
&h(a_3) = h(\bar{b}_3) = \varepsilon, && h(\bar{a}_3) = h(b_3) = p \\
&h(a_2) = h(\bar{b}_2) = \varepsilon, && h(\bar{a}_2) = h(b_2) = q \\
&h(a_1) = h(\bar{b}_1) = h(\bar{a}_1) = \varepsilon, && h(\bar{b}_1) = qp \\
&g(a_3) = g(\bar{b}_3) = q, && g(\bar{a}_3) = g(b_3) = \varepsilon \\
&g(a_2) = g(\bar{b}_2) = p, && g(\bar{a}_2) = h(b_2) = \varepsilon \\
&g(a_2) = g(\bar{b}_1) = h(\bar{a}_1) = \varepsilon, && g(\bar{b}_1) = qp.
\end{aligned}
$$

It is straightforward to check that h, g agree on each word in T_3. On the other hand, $h(\mu) = pqqp \neq qppq = g(\mu)$, proving our lemma. ∎

For $k \geq 1$ define a linear context-free grammar $G^k = (N^k, T^k, P^k, S^k)$ in the following way:

$N^k = \{A_1, A_2, A_3\}$,
$T^k = \{a_{i,j}, \bar{a}_{i,j} : 1 \leq i \leq 3, 1 \leq j \leq k\}$,
$S^k = A_3$,
$P_k = \{A_3 \to a_{3,j} A_2 \bar{a}_{3,j},\ A_2 \to a_{2,j},\ A_1 \bar{a}_{2,j},\ A_1 \to a_{1,j} \bar{a}_{1,j} : 1 \leq j \leq k\}$ for $k \geq 1$.

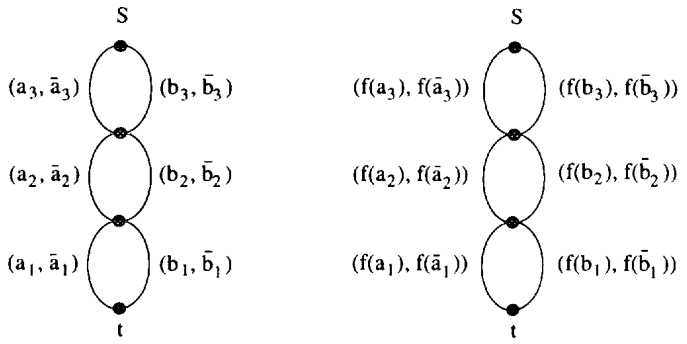

FIG. 12. $\text{graph}(G_3)$ and $\text{graph}_f(G_3)$.

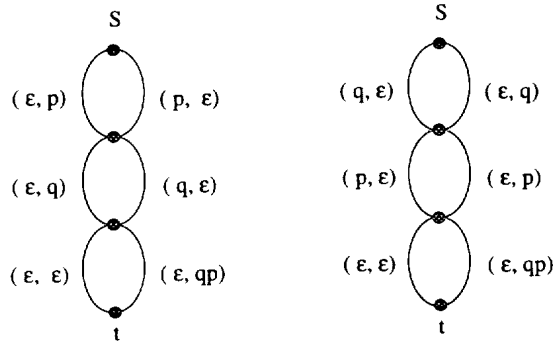

FIG. 13. $\text{graph}_h(G_3)$ and $\text{graph}_g(G_3)$.

Let $\Xi = \{L(G^k)\}_{k \geq 1}$. Our next lemma shows that Ξ has the interesting property we are looking for:

LEMMA 9. *$L(G^k)$ is the only test set for $L(G^k)$ for $k \geq 1$.*

Proof. Suppose that $L(G^k)$ possessess a test set $T \subset L(G^k)$ such that $T \neq L(G^k)$. Let $x = a_{1,j}{}^1 a_{2,j}{}^2 a_{3,j}{}^3 \bar{a}_{3,j}{}^3 \bar{a}_{2,j}{}^2 \bar{a}_{1,j}{}^1$ be any fixed word in $L(G^k) \backslash T$. We define two morphisms h', g' as

$$g'(a_{i,j}) = \begin{cases} g(a_i) & \text{if } j \neq j_i, \\ g(b_i) & \text{if } j = j_i, \end{cases} \quad \text{where } 1 \leq i \leq 3,\ 1 \leq j \leq k;$$

$$g'(\bar{a}_{i,j}) = \begin{cases} g(\bar{a}_i) & \text{if } j \neq j_i, \\ g(\bar{b}_i) & \text{if } j = j_i, \end{cases} \quad \text{where } 1 \leq i \leq 3,\ 1 \leq j \leq k;$$

$$h'(a_{i,j}) = \begin{cases} h(a_i) & \text{if } j \neq j_i, \\ h(b_i) & \text{if } j = j_i, \end{cases} \quad \text{where } 1 \leq i \leq 3,\ 1 \leq j \leq k;$$

$$h'(\bar{a}_{i,j}) = \begin{cases} h(\bar{a}_i) & \text{if } j \neq j_i, \\ h(\bar{b}_i) & \text{if } j = j_i, \end{cases} \quad \text{where } 1 \leq i \leq 3,\ 1 \leq j \leq k,$$

where g and h are morphisms from the previous lemma. It follows from the construction that h' and g' agree on all words from $L(G^k)$ except x. This contradicts our assumption that T is a test set for $L(G^k)$. ∎

Now we are ready to prove the main result of this section.

THEOREM 11. (a) *The lower bound for the size of a test set for languages from the family of linear context-free languages which are produced by grammars containing m productions is $\Omega(m^3)$.*

(b) *The lower bound for the size of a test set for languages from the family of finite languages which are defined over n-letters alphabet is $\Omega(n^3)$.*

Proof. The family Ξ is a subfamily of linear context-free languages and finite languages. $L(G^k)$ consists of k^3 words, it is produced by a linear context-free grammar with $3k$ productions and it is defined over a $6k$-letter alphabet. Hence the result follows. ∎

REFERENCES

1. J. Albert and K. Culik, Test sets for homomorphism equivalence on context-free languages, *Inform. and Control* **45** (1980), 273–284.
2. J. Albert, K. Culik II, and J. Karhumäki, Test sets for context-free languages and algebraic systems of equations, *Inform. and Control* **52** (1982), 172–186.
3. M. H. Albert and J. Lawrence, A proof of Ehrenfeucht's conjecture, *Theoret. Comput. Sci.* **41** (1985), 121–123.
4. A. Ehrenfeucht, J. Karhumäki, and G. Rozenberg, On binary equality sets and a solution to the test set conjecture in the binary case, *J. Algebra* **85** (1983), 76–85.
5. M. A. Harrison, "Introduction to Formal Language Theory," Addison–Wesley, Reading, MA, 1978.
6. J. Karhumäki, The Ehrenfeucht conjecture: a compactness claim for finitely generated free monoids, *Theoret. Comput. Sci.* **29** (1984), 285–308.
7. J. Karhmäki, On recent trends in formal language theory, *in* "Lect. Notes in Comput. Science, Vol. 267," pp. 136–162, Springer-Verlag, New York/Berlin, 1987.
8. J. Karhumäki, W. Plandowski, and W. Rytter, Polynomial size test sets for context-free languages, *in* "Proceedings, of ICALP'92," Lect. Notes in Comput. Science, Vol. 623, pp. 53–64, Springer-Verlag, New York/Berlin, 1992.
9. J. Karhumäki, W. Rytter, and S. Jarominek, Efficient construction of test sets for regular and context-free languages, *Theoret. Comput. Sci.* **116** (1993), 305–316.
10. M. Lothaire, "Combinatorics on Words," Academic Press, New York/Londen 1983.
11. A. Salomaa, "Jewels of Formal Languages Theory," Mir, Moscow, 1986. [Russian]