ON THE WORD PROBLEM FOR SYNTACTIC MONOIDS OF PIECEWISE TESTABLE LANGUAGES

KAMILLA KÁTAI-URBÁN, PÉTER PÁL PACH, GABRIELLA PLUHÁR, ANDRÁS PONGRÁCZ, AND CSABA SZABÓ

Abstract. Piecewise testable languages are widely studied area in the theory of automata. We analyze the algebraic properties of these languages via their syntactic monoids. In this paper a normal form is presented for 2- and 3-piecewise testable languages and a log-asymptotic estimate is given for the number of words over these monoids.

1. Introduction

The theory of formal languages goes back to natural languages. Linguists, e.g. Chomsky, gave mathematical definitions of natural concepts such as words, languages and grammars: given a finite set A, a word on A is simply an element of the free monoid on A, and a language is a set of words. The theory deals with languages, automata and semigroups, and nowadays it has some interesting connections with model theory in logic, symbolic dynamics and topology.

The foundation of the theory is based on Kleene's theorem: it proves that the class of recognizable languages (e.g. recognized by finite automata) coincides with the class of rational languages, which are given by rational expressions. Rational expressions are the generalization of polynomials involving three operations: union, product and star operation. A real break-through in the history of language theory is a work of Schützenberger: he established an equivalence between finite automata and finite semigroups. He showed that a finite monoid, called the syntactic monoid, can be assigned to each recognizable language; this is the smallest monoid recognizing the language. According to Eilenberg's theorem varieties of finite monoids are in one to one correspondence with classes of recognizable languages closed under product and boolean operations. For example, star-free languages correspond to aperiodic monoids. For more details, see [4].

Date: October 3, 2011.

Key words and phrases. semigroup, languages.

A large class of star-free languages is the family of piecewise testable languages, which has been deeply studied in formal language theory. Simon [7] proved that a language is piecewise testable if and only if its syntactic monoid is $\mathcal{J}$-trivial. Simon found necessary and sufficient conditions for an automaton to be piecewise testable. Stern [6] modified these conditions and described a polynomial time algorithm for piecewise testability problem of order $O(n^5)$.

In this paper we analyze the word problem for the syntactic monoids of the varieties of k-piecewise testable languages. We present a normal form of the words for $k = 2$ and 3, and give an asymptotic formula for the logarithm of the number of words for arbitrary k.

2. Preliminaries

At first, some basic notions and definitions are going to be introduced. The word w is a *subword* of u if w is a subsequence of not necessarily consecutive variables taken from u. Given an integer $k > 0$, let $u \sim_k v$ if and only if the words u, v over the alphabet X have the same set of subwords of length at most k. A language L over an alphabet X is *k-piecewise testable* if and only if L is a union of classes of the equivalence relation $\sim_k$. Another characterization says that a language L over an alphabet X is k-piecewise testable if and only if it is a finite boolean combination of languages of the form

$$X^*x_1X^*x_2X^* \dots X^*x_lX^*, \text{ where } x_1, \dots, x_l \in X, \ 0 \leq l \leq k.$$

A language is piecewise testable if there exists a natural number k such that the language is k-piecewise testable.

Simon [7] found a basis of identities for k-piecewise testable languages if $k = 1, 2$. Moreover, Blanchet-Sadri [1, 2] gave a basis of identities for k=3, and proved that there is no finite basis of identities for $k > 3$.

For an integer k the classes of $u \sim_k v$ form the free syntactic monoid in the variety corresponding to k-piecewise testable languages. We denote this variety by V_k. The free monoid (generated by countably many elements) in V_k will be denoted by $F(V_k)$ and the n-generated free monid by $F_{V_k}(n)$. In the paper we often refer to the $\sim_k$ classes as the elements of $F(V_k)$. The number of $\sim_k$ classes of words on at most n letters is equal to $|F_{V_k}(n)|$, the size of the n-generated free monoid in V_k. The sequence $|F_{V_k}(n)|$, $n = 1, 2, \dots$ is called the free spectra of the variety V_k. For a word w let us denote the set of its subwords of length at most k by $C_k(w)$. This way $v \sim_k w$ if and only if $C_k(v) = C_k(w)$.

3. Normal form for $k = 2$

In this section a normal form is given for the terms of $F(V_2)$. Let w be a word. For $a \in c(w)$ let $I_w(a) = \{b \mid ab \text{ is a subword of } w\}$ and for $a, b \in c(w)$ let $a \sim b$ if $I_w(a) = I_w(b)$. Clearly, $\sim$ is an equivalence relation. For an $a \in c(w)$ the $\sim$ class of a is called the block of a and denoted by $B_w(a)$. Two letters are in the same block if they are followed by the same set of letters. A letter a occurs exactly once in w if and only if aa is not a subword of w. There is a natural ordering $\prec$ on the set of blocks. We say that $B_w(a) \prec B_w(b)$ if $I_w(b) \subsetneq I_w(a)$. Clearly, $\prec$ is a well-defined complete ordering. Let $B_1 \prec B_2 \prec \cdots \prec B_t$ be the blocks. For an $a \in c(w)$ let $L_w(a) = \{b \mid ba \text{ is a subword of } w\}$. If $b \in L_w(a)$ for some $b \in B_i$, then $a \in I_w(b)$. Let $c \in B_j$ for some $j \leq i$. By the definition of $\prec$ we obtain $a \in I_w(b) \subseteq I_w(c)$, hence $c \in L_w(a)$. Therefore, $L_w(a)$ is the union of the first few blocks. Let i_a be the index such that $a \in B_{i_a}$ and j_a be the index such that $B_{j_a} \subset L_w(a)$ and $B_{j_a+1} \not\subset L_w(a)$. Note that if the first letter of w occurs only once in w, then $L_w(a)$ is empty. In this case let $j_a = 0$. If a occurs once in w, then a is the first letter of w from B_{i_a}, and $j_a = i_a - 1$. In every other case $j_a \geq i_a$ holds.

Construction 3.1. Let w be a word in $F(V_2)$. Then the *normal form* of w can be obtained in the following way.

Step 1. Identify the blocks of the equivalence relation $\sim$; say $B_1, \ldots, B_t$, in their standard ordering.

Step 2. With each block B_i we will associate a word w_i in the following way. The variables will be the variables appearing in B_i and each letter of B_i appears once in w_i. The order of appearance of these letters is alphabetical, except that if there is a letter in B_i appearing just once in w, then this is placed first (with the remaining variables alphabetical after it).

Step 3. For each $i = 1, \ldots, t$, let v_i denote the product of all variables a with $i_a \leq j_a = i$, ordered alphabetically.

Step 4. The normal form $\overline{w} := w_1 v_1 w_2 \ldots w_t v_t$.

Proposition 3.2. *Let v and w be two words in X^*, then $w \sim_2 \bar{w}$. If $v \sim_2 w$, then $\bar{v} = \bar{w}$. Hence, $\bar{w}$ is a normal form of w.*

Proof. Recall that $v \sim_2 w$ if they have the same subwords of length at most 2. By definition ab is a subword of w if and only if $b \in I_w(a)$ if and only if $a \in L_w(b)$. Therefore, if $v \sim_2 w$, then $I_w(a) = I_v(a)$ and $L_w(a) = L_v(a)$ for any $a \in X$. In Construction 3.1 the normal form depends only on these subsets, so $\bar{v} = \bar{w}$. Observe that by the construction $I_w(a) = I_{\bar{w}}(a)$ and $L_w(a) = L_{\bar{w}}(a)$, hence $w \sim_2 \bar{w}$, as we claimed.

□

For example, consider the word $w = bbaababcbb$. Now, $C_2(w) = \{aa, ab, ac, ba, bb, bc, cb\}$ the blocks are $\{a, b\}$ and $\{c\}$, and the normal form is $abacb$.

For a more detailed example we give a visual description of normal forms. The normal form can be considered as a pair of indices, i_a and j_a, assigned to each variable $a \in c(w)$. The number i_a shows in which block a is and j_a shows after which block a is written. For example, ab is a subword of w if and only if $i_a \leq j_b$. Every variable occurs at most twice in the normal form. Obviously, $j_a \geq i_a - 1$ and the equality holds if and only if a occurs exactly once in w. As we have seen before, in this case a is the first letter of its block B_a. Thus in each block there is at most one letter occuring once in w (and $\bar{w}$), and all other letters occur twice in $\bar{w}$. Two consecutive blocks, B_l and B_{l+1}, can be separated in two ways.

I. There is a letter d such that $i_d \leq l$ and $j_d = l$,
II. The first letter of B_{l+1} denoted by h occurs once in w, and $i_h = j_h + 1 = l + 1$.

This suggests the following description of the normal forms in $F(V_2)$.

Proposition 3.3. *Let w be a word. Assume that there are m blocks in w. Then the normal form of w can be described by a map φ assigning a pair of indices, (i_a, j_a), to each variable $a \in c(w)$, where*

$$\begin{aligned} \varphi : c(w) \to & \{(i,j) \mid 0 \leq i, j \leq t\} \\ a \mapsto & (i_a, j_a) \end{aligned}$$

and the following conditions hold:

1. *each $1 \leq i \leq t$ occurs at least once as a first coordinate,*
2. *each $1 \leq j \leq t-1$ occurs at least once as a second coordinate,*
3. *$i_a - 1 \leq j_a$,*
4. *the pairs $(i, i-1)$ are assigned to at most one variable.*

Moreover, every map satisfying (1)-(4) corresponds to a word with exactly t blocks.

Proof. By Construction 3.1 and Proposition 3.2 item (1) guarantees that the blocks are not empty in a normal form. By item (2) the blocks are separated, and by item (4) in each block there is at most one variable occuring once in w. Note that if there is no pair with $j_a = t$, then by (1) and (3) the pair $(t, t-1)$ has to be assigned to a variable. In this exceptional case B_t contains a single variable. □

FIGURE 1

On Figure 1 the words $a_2a_5a_6a_7a_8a_5a_9a_4a_4a_8a_3a_{10}a_{10}a_1a_2a_3a_7a_{11}a_6$ and $a_5a_4a_7a_8a_3a_9a_3a_9a_2a_6a_6a_1a_2a_4a_7a_{10}$ are depicted. The columns represent the blocks and the rows the places after the blocks. The letter a_k is written into the square (i_{a_k}, j_{a_k}). For example, in the first table $i_{a_8} = 1$ and $j_{a_8} = 2$. Hence, a_8 is written into the 1st block and after the 2nd block. To sum up, to obtain a normal form the letters have to be written into the table satisfying the following conditions:

- every letter is put into exactly one square of the table;
- there has to be at least one letter in each column;
- there has to be at least one letter in each row, except possibly the 0th and the last ones;
- there is at most one letter in the shaded squares.

Moreover, every arrangement satisfying the above conditions corresponds to a unique normal form.

4. Normal form for $k = 3$

In this section a normal form is given for the elements of $F(V_3)$. Let w be a word with $c(w) \subseteq \{x_1, x_2, \ldots, x_n\}$. For $a \in c(w)$ let

$$I_w^2(a) = \{bc \,|\, abc \text{ is a subword of } w\} \cup \{b \,|\, ab \text{ is a subword of } w\}$$

and

$$L_w^2(a) = \{bc \,|\, bca \text{ is a subword of } w\} \cup \{b \,|\, ba \text{ is a subword of } w\}.$$

The sets $I_w^2(a)$ and $L_w^2(a)$ are determined by $C_3(w)$. Let $c_2(w) = \{a | aa \text{ is a subword of } w\}$ be the set of the letters that occur at least twice in w. Similarly to the case $k = 2$, let $a \sim b$ if $I_w^2(a) = I_w^2(b)$ for $a, b \in c_2(w)$. Clearly, $\sim$ is an equivalence relation. For an $a \in c_2(w)$ the $\sim$ class of a is called the I-block of a. The L-blocks are defined dually. They are the equivalence classes of the relation $\approx$, where $a \approx b$ if and only if $L_w^2(a) = L_w^2(b)$ and both a and b occur in $c_2(w)$. The I-blocks, the L-blocks and the one element sets $\{a\}$, where a is a variable occuring once, are called blocks. The set of the blocks of w depend only on $C_3(w)$.

For an I-block B let z_B be the subword containing the left-most appearances of the elements of B. If B and B' are different I-blocks, and for $b \in B$ and $b' \in B'$ we have $I_w^2(b') \subsetneq I_w^2(b)$, then each letter of z_B is left to each letter of z_B'. This obseration provides an ordering on the set of the I-blocks. An ordering of the L-blocks can be defined dually.

Now, we give a normal form for the elements of the blocks. This normal form will be obtained in several steps.

Step 1. Let $B = \{b_1, \dots, b_t\}$ be an I-block and $w = ub_1 \dots b_t v$, where b_1 and b_t are the first occurences of the first- and last occuring letters of B in w, and u and v are words. Let $\bar{v}$ be the normal form of v in $F(V_2)$, and v_B' the subword of $\bar{v}$, where we keep only the first occurences of the elements of B and cancel all other letters. Let $v_B' = v_1 x$, where x is the last letter of v_B'. Finally, let $v_B = xv_1$ and $w_B = uv_B v$. Note that in this step we eliminate a subword and replace it by a word containig only the letters of the block B. The order of the letters does not depend on the eliminated subword.

Lemma 4.1. *Let w be a word and w_B the word defined in Step 1. Then $w_B \sim_3 w$.*

Proof. Let B be an I-block, $w = ub_1 \dots b_t v$ as in Step 1 and $z = b_1 \dots b_t$ denote the word between u and v in w that is $w = uzv$. We have to show that $C_3(w_B) = C_3(w)$. By the definition of an I-block, the following hold:

- (i) $B \cap c(u) = \emptyset$,
- (ii) $c(z) \setminus B \subseteq c(u)$,
- (iii) $c(z) \subseteq c(v)$, in particular
- (iv) $B \subseteq c(v)$,
- (v) $B \subseteq c(z)$,
- (vi) $C_3(uv) \subseteq C_3(w) \cap C_3(w_B)$.

Note that item (iii) holds, because no variable in z is a last occurence of a variable.

Case 1.$C_3(w_B) \subseteq C_3(w)$. Let $abc \in C_3(w_B)$. We may assume that a is a first and c is a last appearance of some variables. By the definition of a block, v_B does not contain a last appearance. If $a \in v$ or $c \in u$, then $abc \in C_3(w)$ by (vi). Hence, we may assume that c is from v

and a is from uv_B. At first let $a \in c(u)$. Then either $ab \in C_2(u)$ or b is in v_B or $bc \in C_2(v)$. In the first and third cases by (vi), in the second case by (v) we obtain that $abc \in C_3(w)$ holds. Secondly, let $a \in B$. Then either $bc \in C_2(v)$ or $ab \in C_2(v_B)$. In the first case $abc \in C_3(w)$ is obvious by (v). Finally, let $ab \in C_2(v_B)$. Note that $b_2c, b_3c, \ldots, b_lc \in I^2_w(b_1) = I^2_w(b_t)$, and $I^2_w(b_t) = C_2(v) = C_2(\overline{v})$. Thus if $b \neq b_1$, then $bc \in C_2(\overline{v})$, so $abc \in C_3(w)$, again by (v). The remaining case is $b = b_1$. As $ab = ab_1$ is a subword of $v_B = xv_1$, we have $x = b_i$ for some $i \neq 1$. Since $x = b_i$ is the variable in B which occurs last in v, the letter b_1 preceeds b_i in $\overline{v}$. Thus $b_1c \in C_2(\overline{v})$. Then $a \in c(z)$ implies that $abc \in C_3(w)$.

Case 2.$C_3(w) \subseteq C_3(w_B)$. As in the previous case, we may assume that a is a first appearance, and c is a last appearance of some variables. Again, by the definition of a block no variable in z is a last appearance. If $a \in v$ or $c \in u$, then $abc \in C_3(w_B)$ by (vi). So, we may assume that c is from v and a is from uz. If a is from u and b is from u or v, then, again by (vi) $abc \in C_3(w_B)$ holds. If b is from z and b is not the first occurence of b_1, then $bc \in I^2(b_1) = I^2(b_t) = C_2(v)$, hence $abc \in C_3(w_B)$. Finally, if b is the first occurence of b_1, then a is from u and b is from B, and the inclusion is implied by (v).

□

Step 2 Let $B_1, B_2, \ldots, B_t$ be the blocks. Proceed with Step 1 for each I-block and the dual of Step 1 for each L-block (in arbitrary order). Then we obtain the word w_1.

Lemma 4.2. *Let w be a word and w_1 the word obtained from w in Step 2. Then $w_1 \sim_3 w$, and w_1 does not depend on the order of implementation of Step 1 on the blocks.*

Proof. By Lemma 4.1 for any block B we have $w_B \sim_3 w$, and by the ordering of the blocks and the order of the blocks are determined by $C_3(w) = C_3(w_B)$. Hence, w_1 is detremined by $C_3(w)$ and $w_1 \sim_3 w$, as we wanted.

□

Construction 4.3. Let w be a word in $F(V_3)$ and $\hat{w}$ be defined in the following way.

Step 1-2 Let $B_1, B_2, \ldots, B_h$ be the blocks of w. Proceed with Step 1 for each block to obtain the word $w_2 = v_{B_1}u_1v_{B_2}u_2 \ldots v_{B_h}$. where $u_1, \ldots u_h$ are words.

Step 3 For any subword $v_{B_l}u_1v_{B_{l+1}}u_2 \ldots v_{B_m}$ where
- B_l, B_{l+1}, …, B_m are all I-blocks and not L-blocks,
- B_{l-1} is not an I-block or $l = 1$,

– B_{m+1} is not an I-block

define

$$U_i = c(u_i) \setminus \bigcup_{h=i+1}^{m} c(u_h) \text{ for } l-1 \le i \le m-1 \text{ and } U_m = c(u_m)$$

In case B_l, B_{l+1}, ..., B_m are all L-blocks and not I-blocks and B_{l-1} and B_{m+1} are not L-blocks, then U_i is defined dually. The remaining case is when we have two consecutive 1-element blocks consisting of variables occuring once in w. Then let $U_i = c(u_i)$. For every i write the elements of U_i in alphabetical order to get the word u_i^* and replace each u_i by u_i^* in w_1 to obtain the normal form

$$\hat{w} = v_{B_1} u_1^* v_{B_{i+1}} u_2^* \dots v_{B_h}$$

The last step of the construction can be interpreted in the following way. We considered the subwords of the form $v_{B_l} u_1 v_{B_{l+1}} u_2 \dots v_{B_m}$ where B_l, B_{l+1}, ..., B_m are all I-blocks and B_{l-1} and B_{m+1} are not I-blocks. Then we kept only the last appearances of the variables from $\bigcup_{j=l}^{m} c(u_j)$, and put them into alphabetical order between two neighbouring blocks. The word $\hat{w}$ doesn't depend on the order of these procedures.

Proposition 4.4. *Let v and w be two words in $F(V_3)$. Then $w \sim_3 \hat{w}$. If $v \sim_3 w$, then $\hat{v} = \hat{w}$. Hence $\hat{w}$ is a normal form of w.*

Proof. The construction and Lemmas 4.1 and 4.2 imply that $w \sim_3 \hat{w}$. The order of the blocks is determined by $C_3(w)$, and by Lemma 4.1 for any block B the word w_B is determined by $C_3(w)$, as well.

In order to prove $\hat{v} = \hat{w}$, it remains to show that u_i^* is determined by $C_3(w)$. Let $v_{B_l} u_1 v_{B_{l+1}} u_2 \dots v_{B_m}$ be a subword, where B_l, B_{l+1}, ..., B_m are all I-blocks and not L-blocks and B_{l-1} and B_{m+1} are not I-blocks. Then for $i > 1$

$$U_i = \{x : y_i x y_{m+1} \in C_3(w), y_{i+1} x y_{m+1} \notin C_3(w)\}$$

for arbitrary $y_i \in B_i$, $y_{i+1} \in B_{i+1}$ and $y_{m+1} \in B_{m+1}$. If $i = 1$, then $l \ge 2$, and

$$U_1 = \{x : y x y_{m+1} \in C_3(w), y_l x y_{m+1} \notin C_3(w)\}$$

for arbitrary $y_1 \in B_1, y_l \in B_l$ and $y_{m+1} \in B_{m+1}$. For L-blocks the arguments are the dual ones. Finally, if $B_l = \{y\}$ and $B_{l+1} = \{z\}$ are one element blocks, then

$$U = \{x : yxz \in C_3(w).$$

□

Remark 4.5. Let w be a word. In each step of the construction the number of letters in w is not increasing. Hence, $\hat{w}$ is a shortest possible representant in its $\sim_3$ class. Listing the triples occuring in $C_3(w)$ takes $O(|w|^3)$ time and all other steps are linear in $|C_3(w)|$. Hence, finding and multiplying normal forms takes $O(|w|^3)$ time, as expected. Moreover, we can reduce the number of steps for long inputs. Let $\underline{w}$ be the word obtained from w by keeping w' and only one occurence of each variable between suuccessive positions in w'. Then $\underline{w} \sim_3 w$. As $\underline{w}$ has length $O(n^2)$, the normal form of $\underline{w}$ can be obtained in $O(n^6)$ time. Hence, the normal form of a word w can be determined in $O(\max\{|w|^3, |w| + n^6\})$ time.

Problem. *Is there an in-place algorithm that outputs the $k = 3$ normal form, and can it be done in log-space?*

5. Counting words

In this section we give estimates on the number of equivalence classes of $\sim_k$, in other words we estimate the free spectra of the variety V_k.

Proposition 5.1. *Let $f_k(n)$ denote the number of $\sim_k$ equivalence classes on n letters. Then*

$$\log f_k(n) = \Theta(n^{(k+1)/2}), \text{ if } k \text{ is odd}$$

and

$$\log f_k(n) = \Theta(n^{k/2} \log n), \text{ if } k \text{ is even,}$$

where $\Theta(f) = g$ if there are constants d_1 and d_2 such that $d_1 \cdot g(n) \leq f(n) \leq d_2 \cdot g(n)$.

Proof. The theorem is proved by induction. We have $f_1(n) = 2^n$, so $\log f_1(n) = n \log 2 = \Theta(n)$, and the statement holds for $k = 1$. Given a word w let w' denote the word in which only the first and last occurences of the variables are kept and the others are deleted. For $k = 2$ the word w' determines w as $C_2(w) = C_2(w')$. The word w' has length at most $2n$. This yields an obvious upper bound for $k = 2$, namely that $f_2(n) \leq (n+1)^{2n}$. On the other hand, for each permutation π of the numbers $\{1, 2, \dots n\}$ we obtain a different set $C_2(x_{\pi(1)} \cdot \dots \cdot x_{\pi(n)})$. Hence, $n! \leq f_2(n) \leq (n+1)^{2n}$ holds, and Stirling's formula implies $\log f_2(n) = \Theta(n \log n)$.

Let $k \geq 3$. For arbitrary words $w_0, w_1, \ldots, w_{[n/2]+1}$ with $c(w_i) \subseteq \{x_{[n/2]+1}, \ldots, x_n\}$ for every $1 \leq i \leq [n/2]+1$ let

$$w = w_0 x_1 w_1 x_1 x_2 w_2 x_2 \ldots x_{[n/2]} w_{[n/2]} x_{[n/2]} w_{[n/2]+1}.$$

As x_i occurs only twice in w, $x_i u x_i \in C_k(w)$ if and only if $u \in C_{k-2}(w_i)$ for $1 \leq i \leq [n/2]$. Similarly, $u x_1 x_1 \in C_k(w)$ if and only if $u \in C_{k-2}(w_0)$, and $x_{[n/2]} x_{[n/2]} u \in C_k(w)$ if and only if $u \in C_{k-2}(w_{[n/2]+1})$. Hence, for different tuples $(C_{k-2}(w_0), C_{k-2}(w_1), \ldots, C_{k-2}(w_{[n/2]+1}))$ we get different sets $C_k(w)$, thus for different tuples of words from $F(V_{k-2})$ we obtain different words in $F(V_k)$. Therefore, $f_k(n) \geq f_{k-2}^{[n/2]+2}(\lceil n/2 \rceil)$ and

$$\log f_k(n) \geq \frac{n}{2} \log f_{k-2}(\lceil n/2 \rceil). \tag{1}$$

Let w be a word. The word w is separated into $t-1$ (possibly empty) parts by the letters of w':

$$w = y_1 u_1 y_2 u_2 \ldots u_{t-1} y_t$$

We claim that w' and $C_{k-2}(u_1), \ldots, C_{k-2}(u_{t-1})$ determine $C_k(w)$. Let $a_1, a_2, \ldots, a_s \in c(w)$ and $a = a_1 a_2 \ldots a_s$ be a word of length $s \leq k$. Let y_p be the first occurence of a_1 in w and y_q the final occurence of a_s in w. If $p > q$, then $a \notin C_k(w)$. Now suppose that $p \leq q$. As $a_1 \notin c(y_1 u_1 y_2 \ldots y_{p-1} u_{p-1})$ and $a_s \notin c(u_q y_{q+1} u_{q+1} \ldots y_t)$, the word a is in $C_k(w)$ if and only if $a_2 \ldots a_{s-1} \in C_{k-2}(u_p y_{p+1} u_{p+1} \ldots y_{q-1} u_{q-1})$. This happens if and only if $a_2 \ldots a_{s-1}$ can be written as a product $v_1 v_2 \ldots v_{2q-2p-1}$, where $v_1, v_2, \ldots, v_{2q-2p-1}$ are the (possibly empty) subwords of u_p, y_{p+1}, u_{p+1}, $\ldots$, u_{q-1}, respectively. Here the y_i-s are considered as a 1-letter words. Since each of them has at most $k-2$ letters, the sets $C_{k-2}(u_1)$, $C_{k-2}(u_2)$, $\ldots$, $C_{k-2}(u_{t-1})$ determine $C_k(w)$. Therefore $f_k(n) \leq (n+1)^{2n} \cdot f_{k-2}^{2n-1}(n)$ and

$$\log f_k(n) \leq 2n \log(n+1) + (2n-1) \log f_{k-2}(n). \tag{2}$$

Using inequalities (1) and (2) and induction on k, the proposition holds.

□

Note that from the previous proof it follows that if k is even, then $\frac{1}{3^{k^2}} n^k \log n < \log f_k(n) < 3^k n^k \log n$ and if k is odd, then $\frac{1}{3^{k^2}} n^k < \log f_k(n) < 3^k n^k$.

6. Acknowledgements

The authors are thankful to M. Volkov for turning their attention to this topic and for his useful hints. The research of the authors was supported by the Hungarian National Foundation for Scientific Research Grants K67870, K83219, K100246 and TAMOP-4.2.1/B-09/1/KONV-2010-0005.

References

[1] F. Blanchet-Sadri, *Games equations and dot-depth hierarchy*, Comput. Math. Appl. **18** (1989), 809-822.
[2] F. Blanchet-Sadri, *Equations and monoids varieties of dot-depth one and two*, Theoret. Comput. Sci. **123** (1994), 239-258.
[3] M. Lothaire, *Combinatorics on words*, Addison-Wesley, Reading, MA, 1983.
[4] J. E. Pin, *Finite semigroups and recognizable languages: an introduction*, Semigroups, formal languages and groups (York, 1993), 1-32, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 466, Kluwer Acad. Publ., Dordrecht, 1995.
[5] J. E. Pin, *Varieties of Formal Languages*, North Oxford Academic, Plenum, 1986.
[6] J. Stern, *Complexity of some problems from the theory of automata*, Inform. and Control **66** (1985), 163-176.
[7] I. Simon *Piecewise testable events*, in Proc. 2nd GI Conf., Lect. Notes in Comput. Sci. Vol. 33 (1975) 214-222.

(1) Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, H-6720 Szeged, Hungary
E-mail address: katai@math.u-szeged.hu

(2,3,5) Eötvös Loránd University, Department of Algebra and Number Theory, 1117 Budapest, Pázmány Péter sétány 1/c, Hungary
E-mail address: ppp24@cs.elte.hu
E-mail address: plugab@cs.elte.hu
E-mail address: csaba@cs.elte.hu

(4) Central European University, Department of Mathematics and Applications, 1051 Budapest, Nádor utca 9. Hungary
E-mail address: pongeee@cs.elte.hu