Discrete Applied Mathematics 157 (2009) 1669–1675

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Communication

On intermediate factorial languages

Arseny M. Shur [*,1]

Ural State University, Ekaterinburg, Russia

ARTICLE INFO

Article history:
Received 18 February 2008
Accepted 2 September 2008
Available online 29 October 2008
Communicated by A. Salomaa

Keywords:
Factorial languages
Combinatorial complexity
Intermediate complexity
Web-like automata

ABSTRACT

We prove that factorial languages defined over non-trivial finite alphabets under some natural conditions have intermediate complexity functions, i.e., the number of words in such a language grows faster than any polynomial but slower than any exponential function.

The aim of this paper is to give new examples of factorial languages having intermediate complexity functions. Using very natural conditions, we introduce two families of such languages over an arbitrary non-trivial finite alphabet. Then we estimate their complexity, relying on finite automata of a special kind, introduced in [4].

1. Notation and definitions

1.1. Words and languages

An *alphabet* Σ is a non-empty set, elements of which are called *letters*. A *word* is a finite sequence of letters, say $W = a_1 \ldots a_n$. A word U is a *factor* (respectively *prefix*, *suffix*) of the word W if W can be written as PUQ (respectively UQ, PU) for some (possibly empty) words P and Q. The *power factorization* of a word W is its factorization $W = a_{i_1}^{n_1} \ldots a_{i_t}^{n_t}$ to the minimal number of factors with all factors being powers of a single letter. As usual, we write Σ^n for the set of all n-letter words and Σ^* for the set of all words over Σ. The subsets of Σ^* are called *languages*. A language is *factorial* if it is closed under taking factors of its words, and *antifactorial* if no one of its words is a factor of another one.

1.2. Automata

A *deterministic finite automaton* (DFA) is a 5-tuple $(\Sigma, Q, \delta, s, T)$ consisting of a finite input alphabet Σ, a finite set of states (vertices) Q, a partial transition function $\delta : Q \times \Sigma \to Q$, one initial state s, and a set of terminal states T. The underlying digraph of the automaton contains states as vertices and transitions as directed labeled edges. Then every path in this digraph is labeled by a word, and every cycle is labeled by a *cyclic word*. We make no distinction between a DFA and its underlying digraph. A *reading path* is any path from the initial to a terminal vertex. A DFA *recognizes* the language which is the set of all labels of the reading paths.

We use also deterministic *infinite* automata over the finite input alphabet. The definition of such an automaton is distinct from the one above in exactly one point, namely, the set Q is countably infinite.

* Tel.: +7 343 350 75 79.
E-mail addresses: Arseny.Shur@usu.ru, aimsure@mail.ru.

1 The paper was written when the author was staying at the University of Turku.

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.09.007

1.3. Complexity

For an arbitrary language L over a finite alphabet Σ the *complexity function* is defined by $C_L(n) = |L \cap \Sigma^n|$. For a factorial language the complexity is known to be either bounded by a constant or strictly increasing (cf. [3], and also [1] for the proof in the general case). We are interested in the growth rate rather than in the precise form of the complexity function. As usual, we call a complexity function *polynomial* if it is $O(n^p)$ for some $p \geq 0$ (bounded from above by a polynomial of degree p), and *exponential* if it is $\Omega(\alpha^n)$ for some $\alpha > 1$ (bounded from below by an exponential function at base α). We also write $\Theta(n^p)$ for the function which is bounded from above and from below by polynomials of degree p. A complexity function is said to be *intermediate* if it is bounded neither by a polynomial from above nor by an exponential function from below. Alternatively, it can be said that such a function is *superpolynomial* and *subexponential*.

2. Main result

Throughout the rest of the paper we suppose that the finite alphabet $\Sigma = \{a_1, \ldots, a_k\}$ is fixed, and $k \geq 2$. Consider the language L of all words $U \in \Sigma^*$ with the power factorization $U = a_{i_1}^{t_1} \ldots a_{i_n}^{t_n}$ satisfying $t_j \leq t_{j+1}$ for all $j = 1, \ldots, n-2$. Thus, the powers of letters in U are non-decreasing, with the last letter being the only possible exception. This exception is necessary to make L factorial. The language L is obviously exponential if $k \geq 3$. For example, it contains the language of all square-free words over Σ, which is known to be exponential (cf. [2]). However, there is no evidence about the complexity of L in the case of the binary alphabet. It appears to be intermediate, as a partial case of a more general result on sublanguages of L for arbitrary k.

Now introduce two subsets of L. For the first one fix a cyclic order on Σ, say, $a_1 \prec a_2 \prec \cdots \prec a_k \prec a_1$, and consider all words $U \in L$ with the power factorizations satisfying the additional condition $a_{i_j} \prec a_{i_{j+1}}$ for all $j = 1, \ldots, n-1$. We denote the language obtained by $L_\prec$. Actually, we defined a finite family, parametrized by $\prec$, of subsets of L. All languages of this family obviously have the same complexity.

As to the second subset of L mentioned, we can informally say that all its words *locally* satisfy the same additional condition. More precisely, this subset consists of all words $U \in L$ with the power factorizations satisfying the following condition. There exists an infinite sequence $\{\prec_m\}$ of cyclic orders on Σ such that for any $m \in \mathbb{N}$ the statement (*) below holds true. A *segment* of the power factorization is a product of several consecutive factors.

(*) Suppose that $a_{i_l}^m \ldots a_{i_r}^m$ is a segment of the power factorization such that $t_{l-1} < m$ or $l = 1$, and $t_{r+1} > m$ or $r + 1 = n$. Then

$$a_{i_l} \prec_m \cdots \prec_m a_{i_r} \prec_m a_{i_{r+1}}.$$

The language obtained is denoted by $\bar{L}$. We see that we defined here a countably infinite family of languages, indexed by sequences of cyclic orders. All these languages have the same complexity.

Two remarks should be added to the above definitions. First, $L_\prec \subseteq \bar{L}$ for any cyclic order $\prec$. Second, $L_\prec = \bar{L} = L$ for the binary alphabet.

Now we are able to formulate the main result of this paper.

Theorem 1. *The languages $L_\prec$ and $\bar{L}$ have intermediate complexity.*

3. Special automata and their properties

The proof of the theorem is based on the careful study of two families of finite automata introduced in [4]. First we recall the definitions and the properties from [4], and then establish other properties needed in the proof.

Recall that the *antidictionary* of a language L is the set of all minimal words that do not occur as factors of words of L. Any factorial language is determined by its antidictionary, and either they are both rational or neither is rational.

Suppose that Σ is cyclically ordered, say, $a_1 \prec a_2 \prec \cdots \prec a_k \prec a_1$. Let us write $\bar{a}$ for the successor of the letter a, and consider the family $\{AD_{m,\prec}\}$ of finite antidictionaries defined by

$$AD_{m,\prec} = \{ab \mid a, b \in \Sigma,\ b \neq a,\ b \neq \bar{a}\} \cup$$

$$\{a^2\bar{a}\bar{\bar{a}}, a^3\bar{a}^2\bar{\bar{a}}, \ldots, a^m\bar{a}^{m-1}\bar{\bar{a}} \mid a \in \Sigma\} \cup \{a^{m+1}\bar{a} \mid a \in \Sigma\}. \quad (1)$$

These antidictionaries generate a family of rational languages $\{L_{m,\prec}\}$, recognized by so-called *web-like* automata $\{\mathcal{W}_{m,\prec}\}$ (see Fig. 1). We associate each vertex of $\mathcal{W}_{m,\prec}$ with a word that labels the shortest reading path ending in this vertex. The automaton $\mathcal{W}_{m,\prec}$ contains one *level 1 cycle* labeled by the cyclic word $a_1 \ldots a_k$, one *level 2 cycle* labeled by the cyclic word $a_1^2 \ldots a_k^2$, and so on, up to the *level m cycle* labeled by $a_1^m \ldots a_k^m$, and k *level* $m + 1$ *cycles* that are loops. The initial vertex of $\mathcal{W}_{m,\prec}$ belongs to no cycles, while each other vertex belongs to a unique cycle, so we refer to them as to *level s vertices*. Each edge in the automaton either belongs to one of the level s cycles or leads from some level s vertex a^s to the level $s + 1$ vertex a^{s+1}. Hence, the levels of vertices along any reading path constitute a non-decreasing sequence. Therefore, any reading path encounters at most $m + 1$ different cycles. This gives us the bound $\Theta(n^m)$ for the complexity of $L_{m,\prec}$ (see [5], Theorem 3.1).

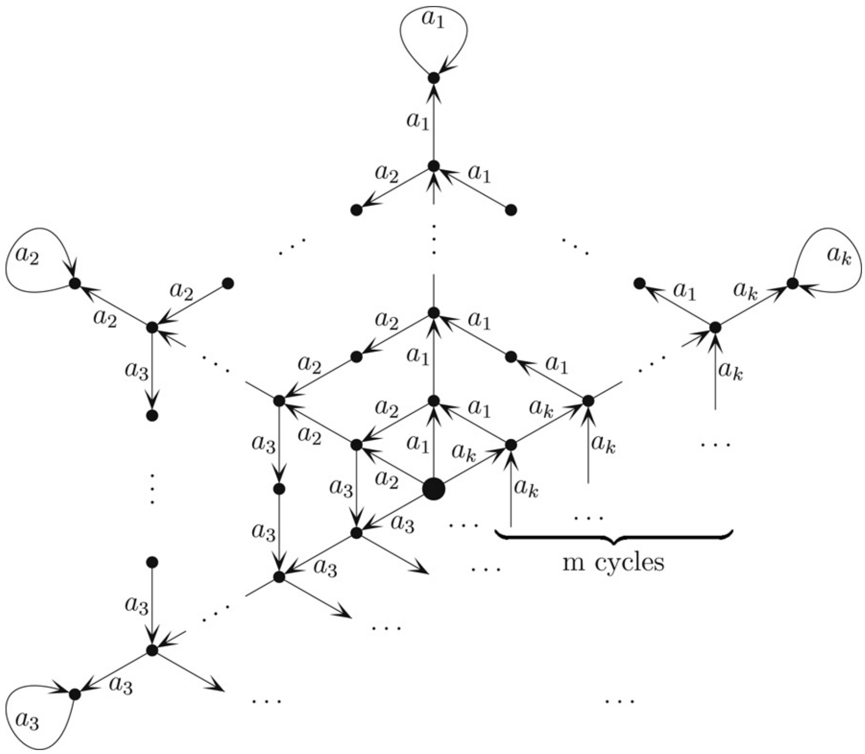


Fig. 1. The "web-like" automaton for the polynomial language of degree m. The bigger circle represents the initial vertex.

Lemma 1. *The language $L_{m,\prec}$ consists of all words with the power factorization $U = a_{i_1}^{t_1} \dots a_{i_n}^{t_n}$ such that*

(1) $a_{i_j} \prec a_{i_{j+1}}$ *for all* $j = 1, \dots, n-1$,
(2) $t_j \le t_{j+1}$ *for all* $j = 1, \dots, n-2$, *and*
(3) $t_j \le m$ *for all* $j = 1, \dots, n-1$.

Proof. The first, second and third sets of words in the antidictionary $AW_{m,\prec}$ (see (1)) provide exactly the conditions (1), (2), and (3) of the lemma, respectively.

The following example represents the structure of the reading path for U in $\mathcal{W}_{m,\prec}$.

Example 1. Let $\Sigma = \{a, b, c, d\}$, $a \prec b \prec c \prec d \prec a$. A word U of $L_{3,\prec}$ is shown below together with the levels of vertices in the reading path (since the letters correspond to edges of the path, we consider a vertex as a position "between" two letters of the word):

$$U = .\mathrm{c}\ \underbrace{.\mathrm{d.a.b.c.d.a.}}_{\text{level 1}}\ \mathrm{a}\ \underbrace{.\mathrm{b.b.c.c.d.d.a.a.b.b.}}_{\text{level 2}}\ \mathrm{b}\ \underbrace{.\mathrm{c.c.c.d.d.d.a.a.a.}}_{\text{level 3}}\ \mathrm{a}\ \underbrace{.\mathrm{a.a.a.a.a.}}_{\text{level 4}}$$

Now introduce the "limit" of the sequence $\{L_{m,\prec}\}$. Take the infinite antidictionary

$$AD_{\prec} = \{ab \mid a, b \in \Sigma,\ b \ne a,\ b \ne \bar{a}\} \cup \{a^2\bar{a}\bar{\bar{a}}, a^3\bar{a}^2\bar{\bar{a}}, \dots, a^{m+1}\bar{a}^m\bar{\bar{a}}, \dots \mid a \in \Sigma\}, \tag{2}$$

and denote the factorial language with this antidictionary by $\hat{L}_{\prec}$. The following straightforward lemma is similar to Lemma 1.

Lemma 2. *The language $\hat{L}_{\prec}$ consists of all words with the power factorization $U = a_{i_1}^{t_1} \dots a_{i_t}^{t_n}$ such that*

(1) $a_{i_j} \prec a_{i_{j+1}}$ *for all* $j = 1, \dots n-1$, *and*
(2) $t_j \le t_{j+1}$ *for all* $j = 1, \dots n-2$.

Thus, we have $\hat{L}_{\prec} = L_{\prec}$. Combining Lemmas 1 and 2 we obtain the following simple lemma.

Lemma 3. *The following formulas are true:*

(1) $L_{1,\prec} \subset L_{2,\prec} \subset \dots \subset L_{m,\prec} \subset \dots \subset L_{\prec}$;
(2) $\bigcup_{m=1}^{\infty} L_{m,\prec} = L_{\prec}$;

(3) $L_{\prec} \cap \Sigma^m = L_{m,\prec} \cap \Sigma^m$.

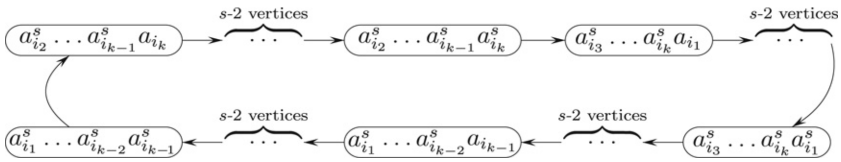
We also note that an infinite web-like automaton can be constructed as the limit of the sequence of finite web-like automata, and this automaton recognizes exactly the language $L_{\prec}$.

Now we proceed with the second family of finite antidictionaries and corresponding sequence of automata. Consider a family of symmetric finite antidictionaries $\{AD_m\}$, where AD_m is the minimal symmetric (that is, stable under all permutations of letters) language containing the following set of words:

$$\begin{array}{llll} \{ & a_1^{m+1}a_2, & & \\ & a_1^m a_2^m a_1, & a_1^m a_2^m a_3^m a_1, & \dots, & a_1^m a_2^m \dots a_{k-1}^m a_1, \\ & a_1^m a_2^{m-1} a_3, & & \\ & a_1^{m-1} a_2^{m-1} a_1, & a_1^{m-1} a_2^{m-1} a_3^{m-1} a_1, & \dots, & a_1^{m-1} a_2^{m-1} \dots a_{k-1}^{m-1} a_1, \\ & a_1^{m-1} a_2^{m-2} a_3, & & \\ & \dots & & \\ & a_1^2 a_2 a_3, & & \\ & a_1 a_2 a_1, & a_1 a_2 a_3 a_1, & \dots, & a_1 a_2 \dots a_{k-1} a_1 \qquad \}. \end{array} \tag{3}$$

These antidictionaries generate a family of rational languages $\{L_m\}$, recognized by finite automata $\{\mathcal{A}_m\}$. The automaton $\mathcal{A}_m$ is complicated enough even in the simplest particular cases, so we do not draw it here (the picture of $\{\mathcal{A}_2\}$ for the ternary alphabet can be found in [4]). However, this automaton can be mentioned as constructed from a finite, but large number of copies of $\mathcal{W}_{m,\prec}$. The necessary properties of $\mathcal{A}_m$ are collected below.

All cycles in $\mathcal{A}_m$ are disjoint. For any $s = 1, \dots, m$ $\mathcal{A}_m$ contains $(k-1)!$ *level s cycles* labeled by all possible cyclic words of the form $a_{i_1}^s \dots a_{i_k}^s$, where all letters $a_{i_1}, \dots, a_{i_k}$ are distinct. Such a cycle connects a cyclic sequence of vertices



All k level $m+1$ cycles are loops, similar to the case of $\mathcal{W}_{m,\prec}$. In the present case we retain the notion of level s vertex. But in $\mathcal{A}_m$ there are vertices which belong to no cycle, provided that $k > 2$. These are vertices of the form $a_{i_1}^s \dots a_{i_{r-1}}^s a_{i_r}^t$, where $r < k-1$. Such vertices are called *prefix level s vertices*, because they are prefixes of some level s vertices, but not of vertices of lesser level. Each level s vertex is prefix level s also.

Each edge of $\mathcal{A}_m$ either preserves the prefix level or increases it by 1; in the last case the destination vertex is a power of a single letter. Hence, the prefix levels of vertices of a reading path constitute a non-decreasing sequence. Like for the case of $\mathcal{W}_{m,\prec}$, we see that the maximum number of different cycles encountered by a reading path is $m+1$, whence we get the complexity bound $\Theta(n^m)$ for L_m.

Now we prove the analogues of Lemmas 1–3.

Lemma 1′. *The language L_m consists of all words with the power factorization $U = a_{i_1}^{t_1} \dots a_{i_n}^{t_n}$ such that:*

(1) *there exist cyclic orders $\prec_1, \dots, \prec_m$ on Σ such that if $t_j = s$ then $a_{i_j} \prec_s a_{i_{j+1}}$ for all $j = 1, \dots n-1$,*
(2) *$t_j \le t_{j+1}$ for all $j = 1, \dots n-2$, and*
(3) *$t_j \le m$ for all $j = 1, \dots n-1$.*

Proof. The words of the antidictionary that are situated in the odd rows of (3) clearly provide the conditions (2) and (3). To prove (1) we construct the required cyclic order $\prec_s$ for any $s = 1, \dots, m$. Fix the number s and consider the fragment $U_s = a_{i_l}^{t_l} \dots a_{i_r}^{t_r} a_{i_{r+1}}^{t_{r+1}}$ of the power factorization of U such that

$$t_l = t_r = s,$$
$$t_{l-1} < s \quad \text{or} \quad l = 1,$$
$$t_{r+1} > s \quad \text{or} \quad r+1 = n.$$

Using that the antidictionary contains the words

$$a_1^s a_2^s a_1,\ a_1^s a_2^s a_3^s a_1,\ \dots,\ a_1^s a_2^s \dots a_{k-1}^s a_1,$$

together with all their cyclic permutations, we obtain the following. If U_s consists of at most k powers of letters, then all these letters are different. Hence we can order the existing letters as $a_{i_l} \prec \cdots \prec a_{i_r} \prec a_{i_{r+1}}$, and complete this partial order to a cyclic order on Σ in an arbitrary way. Suppose that U_s consists of more than k powers of letters. We have that any k successive powers are those of different letters. Then $a_{i_{l+k}} = a_{i_l}$, $a_{i_{l+k+1}} = a_{i_{l+1}}$, and so on. Therefore we can define the required cyclic order as

$$a_{i_l} \prec_s a_{i_{l+1}} \prec_s \cdots \prec_s a_{i_{l+k-1}} \prec_s a_{i_l}.$$

It can be directly verified that any word with the power factorization satisfying the conditions (1)–(3) has no factors from AD_m and hence belongs to L_m.

The following example represents the structure of the reading path for U in $\mathcal{A}_m$.

Example 2. Let $\Sigma = \{a, b, c, d\}$. A word U of L_3 is shown below together with the levels of vertices in the reading path:

$$U = \underbrace{.c.d.a\ \underbrace{.b.c.d.a.}_{\text{level 1}}}_{\text{prefix level 1}}\ \underbrace{a\ .c.c.b\ \underbrace{.b.d.d.a.a.c.c.}_{\text{level 2}}}_{\text{prefix level 2}}\ \underbrace{c\ .a.a.a.d\ \underbrace{.d.d.b.b.b.}_{\text{level 3}}}_{\text{prefix level 3}}\ \underbrace{b\ \underbrace{.b.b.b.b.b.b.}_{\text{level 4}}}_{\text{prefix level 4}}$$

The letters in this word follow each other according to three different cyclic orders.

In order to define the "limit" of the sequence $\{L_m\}$ consider an infinite antidictionary AD which is the minimal symmetric language containing the infinite set of words

$$\begin{array}{lllll} \{ & \ldots, & & & \\ & a_1^{m+1}a_2^m a_3, & & & \\ & a_1^m a_2^m a_1, & a_1^m a_2^m a_3^m a_1, & \ldots, & a_1^m a_2^m \ldots a_{k-1}^m a_1, \\ & a_1^m a_2^{m-1} a_3, & & & \\ & \ldots & & & \\ & a_1^2 a_2 a_3, & & & \\ & a_1 a_2 a_1, & a_1 a_2 a_3 a_1, & \ldots, & a_1 a_2 \ldots a_{k-1} a_1 \quad \}, \end{array} \tag{4}$$

and denote the factorial language with this antidictionary by $\hat{L}$. The following two lemmas are easy analogues of Lemmas 2 and 3.

Lemma 2′. *The language $\hat{L}$ consists of all words with the power factorization $U = a_{i_1}^{t_1} \ldots a_{i_n}^{t_n}$ such that:*

(1) *there exists an infinite sequence $\{\prec_m\}$ of cyclic orders on Σ such that if $t_j = s$ then $a_{i_j} \prec_s a_{i_{j+1}}$ for all $j = 1, \ldots n-1$, and*
(2) $t_j \le t_{j+1}$ for all $j = 1, \ldots n-2$.

Thus, we have $\hat{L} = \bar{L}$.

Lemma 3′. *The following formulas are true:*

(1) $L_1 \subset L_2 \subset \cdots \subset L_m \subset \cdots \subset \bar{L}$;
(2) $\bigcup_{m=1}^{\infty} L_m = \bar{L}$;
(3) $\bar{L} \cap \Sigma^m = L_m \cap \Sigma^m$.

4. Proof of the main result

The proof of superpolynomiality is straightforward. Suppose that the complexity function of $L_\prec$ is bounded from above by a polynomial of degree m. The language $L_{m+1,\prec}$ with the complexity $\Theta(n^{m+1})$ is a subset of $L_\prec$ by Lemma 3. This complexity is not bounded from above by a polynomial of degree m, so we get a contradiction. The proof for $\bar{L}$ is the same.

To prove that these languages are subexponential we need some auxiliary construction. We first get a subexponential upper bound for the complexity function of $L_\prec$, and then use the same idea in a slightly more complicated form for the case of $\bar{L}$.

The value $C_{L_\prec}(n)$ equals the number of reading paths of length n in the web-like automaton $\mathcal{W}_{n,\prec}$. We associate a *characteristic word* of length n over $\{0, 1\}$ with each such path as follows. The i-th letter of the characteristic word is 0 (respectively 1) if the i-th edge of the path belongs to a cycle (respectively belongs to no cycle). Since a vertex of $\mathcal{W}_{n,\prec}$ has neither two outgoing cyclic edges nor two outgoing noncyclic edges, a characteristic word determines a word of $L_\prec$ up to the first letter. Hence, $C_{L_\prec}(n)$ equals the number of characteristic words of length n multiplied by the size of Σ. So, it is sufficient to prove that the language CH of all characteristic words is subexponential.

The properties of characteristic words are collected in the following lemma.

Lemma 4. (1) *Each non-empty characteristic word begins with* 1.
(2) *The language CH is closed under taking prefixes.*
(3) *The number of zeros between i-th and $(i+1)$-th* 1*'s in any characteristic word is either zero or a multiple of i.*

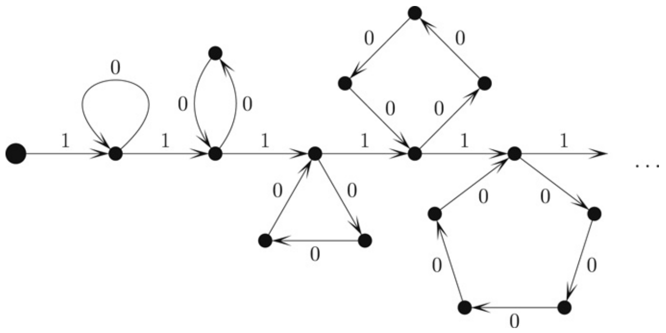


Fig. 2. An infinite automaton recognizing a binary language of complexity $C_{L_{\prec}}(n)/k$. The bigger circle represents the initial vertex.

Proof. (1) The initial vertex of $\mathcal{W}_{n,\prec}$ belongs to no cycles.

(2) An initial segment of a reading path is a reading path also.

(3) The i-th 1 in a characteristic word codes the edge going from the level $i-1$ cycle to the level i cycle of the automaton. After this edge the path can either immediately leave the level i cycle or go with it. The first alternative means that the $(i+1)$-th 1 immediately follows the i-th 1. As regards the second alternative, note that among the level i vertices only the vertices of the form $a^i,\ a \in \Sigma$, have outgoing edges to the level $i+1$ cycle. The nearest such vertices are connected by a fragment of the level i cycle of length exactly i. Thus, a path that exits the level i cycle contains an integer number of such fragments. □

An infinite automaton recognizing CH is shown in Fig. 2.

Count the maximum possible number of occurrences of the factor 10 in a characteristic word. If the i-th 1 in a characteristic word is followed by 0, then it is followed by at least i 0's by statement (3) of Lemma 4. Hence, the difference between the initial positions of i-th and $(i+1)$-th occurrences of 10 is at least $i+1$. The first and the last factors 10 can begin in the first and the penultimate positions of the word respectively. Thus, the minimum length of a characteristic word with t factors 10 is

$$1+2+\cdots+t+1=\frac{t(t+1)}{2}+1,$$

and this yields that the number of (10)'s in a characteristic word of length n is $O(\sqrt{n})$. Note that the number of (10)'s and the number of (01)'s in any binary word differs by at most one. In our case the number of (01)'s is equal to the number of (10)'s or is less by one due to statement (1) of Lemma 4.

It is clear that any word over $\{0, 1\}$ is determined by its first letter (it is 1 in our case) and the set of positions of all its 10 and 01 factors. To determine such a word of length n with i such factors, we choose i positions for them among $n-1$ possible. This means that the total number of such words is $\binom{n-1}{i}$. In our case $i \le B\sqrt{n}$ for some constant B. Now we can evaluate the complexity of the language CH ($\lfloor x \rfloor$ stands for the integer part of x):

$$C_{CH}(n) \le \sum_{i=1}^{\lfloor B\sqrt{n}\rfloor} \binom{n-1}{i} < \lfloor B\sqrt{n}\rfloor \binom{n-1}{\lfloor B\sqrt{n}\rfloor} = \lfloor B\sqrt{n}\rfloor \frac{(n-1)\dots(n-\lfloor B\sqrt{n}\rfloor)}{\lfloor B\sqrt{n}\rfloor !}$$

$$< (n-1)\dots(n-\lfloor B\sqrt{n}\rfloor) < n^{B\sqrt{n}} = 2^{B\sqrt{n}\log n}. \tag{5}$$

We have proved that the language CH is subexponential, and so is $L_{\prec}$.

Now we are able to prove that $\bar{L}$ is also subexponential. Any length n word of $\bar{L}$ is recognized by the automaton $\mathcal{A}_n$. As above, a reading path in $\mathcal{A}_n$ will generate a unique characteristic word of $\{0, 1\}^*$; we slightly change the rule for generating such words. In the above case 0 in a characteristic word corresponds to an edge that retains the level of the vertex, while 1 corresponds to the edge that increases this level. We replace the *level* of the vertex by its *prefix level* in this condition to obtain a definition that suits the symmetric case. One can prove that the new definition leads to the same set CH of characteristic words. But we do not need this statement to prove the theorem, so we leave the proof to the reader. The only result that we need is the statement (3) of Lemma 4. The proof of it for the new definition is nearly the same, but we give it here.

Recall that all edges increasing the prefix level of a vertex from i to $i+1$ lead to some vertex a^{i+1} of $\mathcal{A}_m$. The i-th 1 in a characteristic word U codes the edge connecting some prefix level $i-1$ vertex to a prefix level i vertex, say a^i. After this edge the path can either immediately go to the prefix level $i+1$ or stay on the current prefix level. The first alternative means that the $(i+1)$-th 1 of U immediately follows the i-th 1. If the path stays on prefix level i, it goes to some vertex a^ib. The next $(i-1)$ edges of the path are uniquely determined and lead to the vertex a^ib^i, so we have i 0's following the i-th 1 in U. Here the path again has the possibility of leaving prefix level i, and if it does not, then the situation will repeat after each i edges. The required property follows from this.

Like in the above, using this property we obtain that a length n characteristic word has $O(\sqrt{n})$ factors 10 and 01. Thus, we have the upper bound (5) for the number of characteristic words of length n.

Now take a characteristic word U of length n and estimate the number of paths in $\mathcal{A}_n$ with this characteristic word. Any vertex V of $\mathcal{A}_n$ has at most one outgoing edge to a vertex of bigger prefix level. (This edge exists if V has prefix level i and ends with some a^i; the label of this edge is a.) On the other hand, V may have several outgoing edges to the vertices of the same prefix level. So we can conclude that a given 1 in a characteristic word corresponds to a unique edge, while a given 0 can correspond to several edges. If the i-th 1 in U is followed by 0, this means that any path with the characteristic word U tries to reach one of level i cycles. There are $(k-1)!$ such cycles; hence the number of different fragments of paths coded by the sequence of 0's after the i-th 1 in U does not exceed $(k-1)!$. As we already know, the number of 1's in U that are followed by 0's is $O(\sqrt{n})$. Thus, U codes less than $(k-1)!^{B\sqrt{n}}$ words for some constant B. By (5), we have

$$C_{\bar{L}}(n) \le C_{CH}(n) \cdot (k-1)!^{B\sqrt{n}} < (2(k-1)!)^{B\sqrt{n}\log n}.$$

The theorem is proved.

Acknowledgements

The author is grateful to J. Karhumäki for valuable remarks on the paper. The author was supported by the Federal Science and Innovation Agency of Russia under the grants RI-111.0/002/075 and 2227.2003.01, by the Russian Foundation for Basic Research under the grant 05-01-00540, and by the Federal Education Agency of Russia under the grant 49123.

References

[1] J. Balogh, B. Bollobas, Hereditary properties of words, RAIRO – Inform. Theoret. Appl. 39 (2005) 49–65.
[2] F.-J. Brandenburg, Uniformly growing k-th power free homomorphisms, Theoret. Comput. Sci. 23 (1983) 69–82.
[3] C. Choffrut, J. Karhumäki, Combinatorics of Words, in: G. Rosenberg, A. Salomaa (Eds.), in: Handbook of Formal Languages, vol. 1, Springer, Berlin, 1997, pp. 329–438 (Chapter 6).
[4] A.M. Shur, Polynomial languages with finite antidictionaries, TUCS Tech. Report, No. 721, November 2005.
[5] A.M. Shur, Combinatorial complexity of rational languages, Discrete Anal. Oper. Res. Ser. A 12 (2) (2005) 78–99 (in Russian).