RAIRO-Theor. Inf. Appl. **42** (2008) 647–655
DOI: 10.1051/ita:2008021

Available online at:
www.rairo-ita.org

COMPARING COMPLEXITY FUNCTIONS OF A LANGUAGE AND ITS EXTENDABLE PART *

Arseny M. Shur[1]

Abstract. Right (left, two-sided) extendable part of a language consists of all words having infinitely many right (resp. left, two-sided) extensions within the language. We prove that for an arbitrary factorial language each of these parts has the same growth rate of complexity as the language itself. On the other hand, we exhibit a factorial language which grows superpolynomially, while its two-sided extendable part grows only linearly.

Mathematics Subject Classification. 68Q70, 68R15.

Introduction

The combinatorial complexity of a language (simply *complexity* throughout the paper) is a function defined for an arbitrary language L over a finite alphabet Σ by the rule $C_L(n) = |L \cap \Sigma^n|$. This is the most natural counting function associated with the language. The complexity was intensively studied for many languages and particular classes of languages. Probably the first results in this direction were obtained by Morse and Hedlund [10]. A systematic study of combinatorial complexity was initiated by Ehrenfeucht and Rosenberg in [4]; they focused mostly on an important, but narrow class of D0L-languages. A representative selection of results on complexity can be found in Section 9 of [1]. Besides this, some general results on complexity of arbitrary regular (rational) languages can be found in [6,11]; the sets of possible *polynomial* complexity functions coincide for regular and context-free languages, and explicit formulas for such complexity functions can be effectively constructed [3].

On the other hand, there exist only a few results on "general" complexity properties, which can be applied to arbitrary languages, or at least to wide classes

* *Supported by Federal Science and Innovation Agency of Russia under the grant 02.442.11.7521.*

[1] Ural State University, Ekaterinburg, Russia; Arseny.Shur@usu.ru

Article published by EDP Sciences

of languages, apart from the top two levels of the Chomsky hierarchy. Here we consider one of such "general" questions.

To introduce the question considered, we need two more notions. The language is *factorial*, if it is closed under taking factors of its elements. In most (but not all) cases complexity functions are studied for factorial languages. The right (left, two-sided) *extendable* part of a language consists of all words having infinitely many right (resp. left, two-sided) extensions within the language.

Here we study the following question: *what is the connection between complexity of a language and complexities of its extendable parts?* This question was explicitly stated by Karhumäki at the Automata theory seminar in the University of Turku. In this paper we show that the ratio of complexity of a factorial language and complexity of any of its extendable parts *always* grows subexponentially, and *sometimes* grows superpolynomially. To complete this result, we give a simple example showing that for non-factorial languages this ratio can grow exponentially.

1. Preliminaries

Recall some notions on words, languages, automata, complexity, and graphs.

We consider a finite alphabet Σ and finite words over it. The length of the word W is denoted by $|W|$. A word U is a *factor* of the word W if W can be written as PUQ for some (possibly empty) words P and Q. The *reversal* of a word is obtained by writing its letters in the reversed order. We write Σ^n ($\Sigma^{\leq n}$) for the set of all words of length n (resp. of length at most n) over Σ. As usual, Σ^* denotes 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. The *reversal* of a language consists of the reversals of all its elements.

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 its fastest growing infinite subsequence is $\Omega(\alpha^n)$ for some $\alpha > 1$ (bounded from below by an exponential function at base α). A complexity function which is superpolynomial and subexponential is called *intermediate*. We write $\Theta(n^p)$ for the function which is bounded from above and from below by two polynomials of degree p.

The complexity of a language can be coarsely described by the *growth rate* $\alpha(L) = \limsup_{n\to\infty} C_L(n)^{1/n}$. For factorial languages, the following theorem holds.

Theorem 1.1 [7]**.** *For an arbitrary factorial language L,*

$$\alpha(L) = \lim_{n\to\infty} C_L(n)^{1/n} = \inf_{n\in\mathbb{N}} C_L(n)^{1/n}.$$

Furthermore, $\alpha(L) = 0$ iff L is finite, $\alpha(L) > 1$ iff L is infinite and has the exponential complexity, and $\alpha(L) = 1$ otherwise.

We consider *deterministic finite automata* (dfa's) with a *partial* transition function, and identify a dfa with a digraph, which contains states as vertices and transitions as directed labeled edges. A dfa is *consistent* if each its vertex is contained in some accepting path.

The adjacency matrix of a digraph is nonnegative, whence its eigenvalue of the maximum absolute value is a nonnegative real number, called the *Frobenius root*. This number is usually referred to as the *index* of a digraph [2]. We denote the index of a dfa $\mathcal{A}$ by $r(\mathcal{A})$.

A *strongly connected component* (scc) of a digraph G is a maximal with respect to inclusion subgraph G' such that there exists a (directed) path from any vertex of G' to any other vertex of G'. A well-known result (see [2]) states that the index of a digraph equals maximum of the indices of its scc's. The scc's of index 0 (singletons) and of index 1 (simple cycles) are called *trivial*. The index of any nontrivial scc is strictly greater than 1.

The connection between growth rates of *regular* languages and Frobenius roots of some matrices is well-known. In the most general form, this connection is expressed in the following theorem.

Theorem 1.2. *Let a language L be recognized by a consistent dfa $\mathcal{A}$. Then the growth rate of L coincides with the index of $\mathcal{A}$.*

As far as we know, this theorem was not yet published in this form (a restricted version is proved, for example, in [7]). Since we need such a general form to prove further results, we give the proof here.

Proof. Let $A = (a_{ij})$ be the adjacency matrix of $\mathcal{A}$, m be its size, and $|A|$ be the sum of all elements of A. For any n, consider the matrix $A^n = (a^n_{ij})$. One of the properties of the Frobenius root (see [5]) is the equality $\lim\limits_{n\to\infty} |A^n|^{1/n} = r(\mathcal{A})$.

Note that a^n_{ij} is the number of paths of length n in $\mathcal{A}$ from the state q_i to q_j. Hence, $|A^n|$ is the total number of paths of length n in $\mathcal{A}$, and $P_i = \sum\limits_{j=1}^{m} a^n_{ij}$ is the number of paths of length n in $\mathcal{A}$, starting at q_i. Suppose that the vertex q_1 is initial, and denote $R_j(n) = a^n_{1j}$. Then the complexity $C_L(n)$ equals the sum of these *reading* functions R_j over the set of terminal states. Thus, the maximum growth rate of the functions R_j over the set of terminal states is $\alpha(L)$.

Suppose that $\mathcal{A}$ contains an edge (q_i, q_j). Then $R_j(n{+}1) \geq R_i(n)$, yielding that the growth rate of the function R_j is greater than or equal to the one of R_i. Therefore, the growth rates of the reading functions can only increase along a path in the automaton. Since $\mathcal{A}$ is consistent, for every state there exists a path from it to some terminal state. Thus, the overall maximum of the growth rates of the reading functions is achieved on a terminal state. We obtain that the function $P_1 = \sum\limits_{j=1}^{m} R_j$ has the growth rate $\alpha(L)$.

There exists a path from the initial vertex to any vertex q_i, because $\mathcal{A}$ is consistent. Dually to the above argument on reading functions, we conclude that P_1

has at least the same growth rate as P_i. Since $|A^n| = \sum_{i=1}^{m} P_i(n)$, the growth rate of $|A^n|$ is equal to the maximum of the growth rates of P_i, that is, to the growth rate of P_1. This gives us the required equality $r(\mathcal{A}) = \alpha(L)$. □

2. Extendable parts of a language

For a language L over Σ we consider three subsets of *extendable* words:

right	$re(L) = \{W \in L \mid \forall n{\in}\mathbb{N}\ \exists V{\in}\Sigma^+ :\ |V| \geq n, WV \in L\};$
left	$le(L) = \{W \in L \mid \forall n{\in}\mathbb{N}\ \exists U{\in}\Sigma^+ :\ |U| \geq n, UW \in L\};$
two-sided	$e(L) = \{W \in L \mid \forall n{\in}\mathbb{N}\ \exists U,V{\in}\Sigma^+ :\ |U|, |V| \geq n, UWV \in L\}.$

Obviously, $re(L) \cap le(L) \supseteq e(L)$. Actually, this inclusion is often strict; the following example involves well-known combinatorial objects.

Example 2.1. Recall that a word is *overlap-free*, if it contains no factors of the form XXc, where c is the first letter of the word X. Let $OF \subset \{a, b\}^*$ denote the language of all binary overlap-free words. An infinite *Thue-Morse* word

$$T = abba\,baab\,baab\,abba\,b\ldots$$

over the same alphabet is a fixed point of the morphism ϕ, defined by $\phi(a) = ab$, $\phi(b) = ba$. We write $\bar{T}$ for the reversal of T. It is well known that the word T is overlap-free; hence, so is $\bar{T}$.

From the definition of the Thue-Morse word it is easy to see that T (and also $\bar{T}$) contains no factor $bbabb$. Then, the infinite words $bbabbaT$ and $\bar{T}abbabb$ are overlap-free, and $bbabb \in re(OF) \cap le(OF)$. On the other hand, any word $PbbabbQ$ with nonempty P, Q contains either b^3 or $abbabba$, whence it is not overlap-free. Thus, $bbabb \notin e(OF)$.

The following observation is simple but very useful.

Observation 2.2. $e(L) = le(re(L))$. (By symmetry, $e(L) = re(le(L))$ as well.)

Example 2.1 (continued)**.** The word $bbabb \in re(OF)$ has no left extensions in $re(OF)$, since $bbbabb \notin OF$, and $abbabb \in OF \backslash re(OF)$. This is another way to show that $bbabb \notin e(OF)$.

3. Comparing growth rates

In this section we study how the growth rate of a language relates to the growth rates of its extendable subsets. The answers are quite different in the case of factorial languages and in the case of arbitrary ones.

Theorem 3.1. *For an arbitrary factorial language L, $\alpha(e(L)) = \alpha(le(L)) = \alpha(re(L)) = \alpha(L)$.*

Proof. By Observation 2.2, it is sufficient to prove the statement for $re(L)$, because the result for $le(L)$ can be obtained in the same way, considering the reversal of L instead of L.

Since $re(L) \subseteq L$, the inequalities $C_{re(L)}(n) \leq C_L(n)$ and $\alpha(re(L)) \leq \alpha(L)$ are straightforward. Hence, $\alpha(L) = 0$ implies $\alpha(re(L)) = 0$. Since an infinite language must contain an infinite number of extendable words, $\alpha(L) = 1$ implies $\alpha(re(L)) = 1$ by Theorem 1.1. From now on, we assume that $\alpha(L) > 1$. First we consider the case of a regular language.

Consider a consistent dfa $\mathcal{A}$ recognizing L. Then $\alpha(L) = r(\mathcal{A})$ by Theorem 1.2. Since $r(\mathcal{A}) > 1$, the automaton contains a non-trivial scc; thus, removing trivial scc's from it does not affect the index.

Since L is factorial, and $\mathcal{A}$ is consistent, we see that all vertices of $\mathcal{A}$ are terminal. Let us partition these vertices into two groups, Q_1 and Q_2. A vertex q belongs to Q_1 iff $\mathcal{A}$ contains a cycle, which is attainable from q. It is easy to see that a word $W \in L$ belongs to $re(L)$ iff the reading of W by the automaton $\mathcal{A}$ terminates in some vertex of Q_1. Now we remove all vertices of Q_2 to obtain a consistent dfa $\mathcal{A}'$, recognizing $re(L)$. Note that each vertex of Q_2 forms a singleton scc in $\mathcal{A}$. Hence, $r(\mathcal{A}') = r(\mathcal{A})$, and $\alpha(L) = \alpha(re(L))$, as desired.

Now turn to the general case. A factorial language L over Σ possesses an *antidictionary*, which is the set of minimal *forbidden* words, defined by the formula

$$M = L\Sigma \cap \Sigma L \cap (\Sigma^* \backslash L).$$

It is easy to see that $L = \Sigma^* \backslash \Sigma^* M \Sigma^*$, whence L is regular iff M is. In particular, we may assume for the rest of the proof that M is infinite. Consider the sequence

$$M_1 \subseteq M_2 \subseteq \ldots \subseteq M_i \subseteq \ldots \subseteq M$$

of finite antidictionaries, where $M_i = M \cap \Sigma^{\leq i}$. Let L_i be the factorial language over Σ with the antidictionary M_i. One has

$$L \subseteq \ldots \subseteq L_i \subseteq \ldots \subseteq L_1, \tag{1}$$

and

$$\forall n \exists i_n \ (L \cap \Sigma^n = L_{i_n} \cap \Sigma^n). \tag{2}$$

Then for any n

$$C_L(n) = \ldots = C_{L_{i_n}}(n) \leq \ldots \leq C_{L_1}(n). \tag{3}$$

Hence the sequence $\{C_{L_i}(n)\}$ converges to $C_L(n)$ from above, implying that $\{(C_{L_i}(n))^{1/n}\}$ converges to $(C_L(n))^{1/n}$ from above. By Theorem 1.1, $\alpha(L) = \lim\limits_{n\to\infty} (C_L(n))^{1/n}$ and $\alpha(L_i) \leq (C_{L_i}(n))^{1/n}$. Thus, $\{\alpha(L_i)\}$ converges to $\alpha(L)$.

Now repeat this argument for right extendable parts of the languages L_i. From (1) and (2) we have

$$re(L) \subseteq \ldots \subseteq re(L_i) \subseteq \ldots \subseteq re(L_1), \tag{4}$$

$$\forall n \exists i_n \ (re(L) \cap \Sigma^n = re(L_{i_n}) \cap \Sigma^n), \tag{5}$$

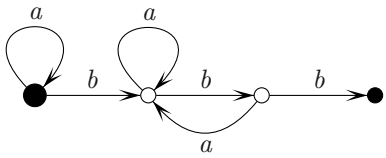


FIGURE 1. This dfa recognizes an exponential language with only constant number of right extendable words. The bigger circle denotes the initial vertex, the terminal vertices are filled.

and finally

$$C_{re(L)}(n) = \ldots = C_{re(L_{i_n})}(n) \leq \ldots \leq C_{re(L_1)}(n). \tag{6}$$

Arguing as above, we obtain that $\{\alpha(re(L_i))\}$ converges to $\alpha(re(L))$. Now it remains to note that each language L_i is regular, whence $\alpha(re(L_i)) = \alpha(L_i)$. The result now follows. □

Remark 3.2. If we consider arbitrary languages instead of factorial ones, the statement of Theorem 3.1 fails even for regular languages. The dfa in Figure 1 recognizes the language $a^*+a^*b(a+ba)^*bb$ having exponential complexity but only one right extendable word of each length. Indeed, the two "middle" vertices of the automaton constitute a nontrivial scc, whose index is equal to the golden ratio. At the same time, only the words from a^* are right extendable in this language.

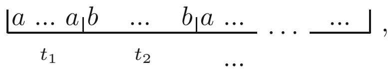
4. Subexponential gaps

According to Theorem 3.1, the gaps between complexities of L, $re(L)$ and $e(L)$ are always subexponential. From known results it follows that all these complexity functions can be polynomials of different degrees. For example, the already mentioned language OF of binary overlap-free words has complexity $\Omega(n^{1.217})$ [9], while $C_{re(OF)}(n) = \Theta(n^\alpha)$ with $\alpha \approx 1.155$ [8]. The language $e(OF)$ coincides with the set of all finite factors of the Thue-Morse word T, and $C_{e(OF)}(n) = \Theta(n)$ (folklore). We now show that such gaps can be much bigger. The language defined below has a superpolynomial gap.

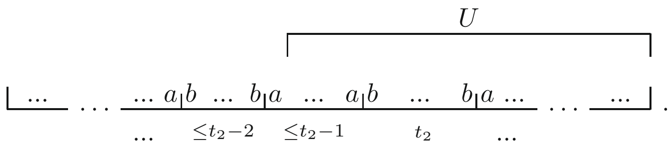
Let $K \subset \{a,b\}^*$ be the language consisting of all words of the form $U = c_1^{t_1} \ldots c_m^{t_m}$ such that $m \in \mathbb{N}$, $c_i \in \{a,b\}$, $c_i \neq c_{i+1}$ for all i, and $t_1 < \ldots < t_{m-1}$. Thus, the powers of letters in U are strictly increasing, with the last letter being the only possible exception. This exception is necessary to make K factorial. We note that K is very close to the family of languages of intermediate complexity, introduced in [12]. The binary language of that family is defined in the same way, as K, with the only difference: the inequalities for t's are not strict. So, we adapt some ideas of [12] to estimate the complexity of K.

Theorem 4.1. *The language K has intermediate complexity, while the complexity of $e(K)$ is linear.*

Proof. First we note that $re(K) = K$, whence $e(K) = le(K)$ by Observation 2.2. If $m \leq 2$, then the word U can be infinitely extended to the left (by the letters c_1, for example). However, for $m \geq 3$ the situation changes drastically. The word U, up to renaming the letters, looks like



while its left extension looks like



It is clear that the length of a left extension of U does not exceed the value

$$(t_2 - 1 - t_1) + (t_2 - 2) + (t_2 - 3) + \ldots + 1 = \frac{t_2(t_2-1)}{2} - t_1,$$

implying $U \notin le(K)$. Therefore, $U \in le(K)$ iff $m \leq 2$. We have

$$le(K) \cap \Sigma^n = \{a^n, a^{n-1}b, \ldots, ab^{n-1}, b^n, b^{n-1}a, \ldots, ba^{n-1}\},$$

whence $C_{le(K)}(n) = 2n$. By Theorems 1.1 and 3.1, the language K has subexponential complexity also. Thus, it remains to show that the complexity of K is superpolynomial.

Let K_m denote the subset of K consisting of all words $U = c_1^{t_1} \ldots c_m^{t_m}$. We show that for n large enough the function $C_{K_m}(n)$ is $\Theta(n^{m-1})$. Since $K = \bigcup_{m=1}^{\infty} K_m$, we then obtain that $C_K(n)$ is not bounded from above by any polynomial.

Note that $C_{K_m}(n)$ equals the number of positive solutions of the diophantine equation

$$(m-1)x_1 + (m-2)x_2 + \ldots + x_{m-1} + x_m = n\,, \tag{7}$$

multiplied by two. Indeed, taking

$$t_1 = x_1,\ t_2 = x_1 + x_2,\ \ldots,\ t_{m-1} = x_1 + \ldots + x_{m-1},\ t_m = x_m,$$

we get the required correspondence between words of K_m and positive solutions of (7) (the multiplication by two is due to the choice of the letter c_1). There are $(m-1)$ free variables in (7), and their values are bounded from above by n. Hence, the number of positive solutions of (8) is $O(n^{m-1})$. To get the lower bound on the number of these solutions, consider an auxiliary equation

$$x_1 + \ldots + x_{m-1} + x_m = n_1. \tag{8}$$

There is a one-to-one correspondence between $(m-1)$-element subsets of the set $\{1,\dots,n_1-1\}$ and sequences of partial sums

$$x_1,\ x_1+x_2,\ \dots,\ x_1+\dots+x_{m-1}.$$

Each of these sequences uniquely determines a positive solution of (8), whence (8) has $\binom{n_1-1}{m-1}$ positive solutions.

Now represent n in the form $n=(m-1)!\cdot n_1+n_2$, where $0\le n_2<(m-1)!$. Any positive solution $(\bar\xi_1,\dots,\bar\xi_m)$ of (8) generates some positive solution $(\xi_1,\dots,\xi_m)$ of (7) by the rule

$$\xi_i=\frac{(m-1)!}{m-i}\cdot\bar\xi_i,\ i=1,\dots,m-1;\ \xi_m=(m-1)!\cdot\bar\xi_m+n_2.$$

Indeed, one has

$$(m-1)\xi_1+\dots+\xi_{m-1}+\xi_m=(m-1)!\cdot(\bar\xi_1+\dots+\cdot\bar\xi_m)+n_2=n.$$

Thus, the number of solutions of (7) exceeds $\binom{n_1-1}{m-1}$. Since n_1 is obtained by dividing n by a constant, we obtain that (7) has $\Omega(n^{m-1})$ positive solutions for any n satisfying the condition $n_1\ge m$. Thus, we get $C_{K_m}(n)=\Theta(n^{m-1})$ for all $n\ge m!$, whence the result. □

We conclude the paper with a few notes on possible applications of the given results. The extendable parts of a language can have more clear structure, than the language itself. For example, the structure of the language OF of binary overlap-free words is rather complicated, while the extendable words in this language are just the factors of a single infinite word T which has a simple and regular form. Thus, if we want to estimate the complexity of some language, we may study its extendable part instead. Theorem 3.1 provides that we can find the growth rate of the target language in this way (and, in particular, decide whether the target language is exponential or subexponential). On the other hand, Theorem 4.1 shows that this method does not allow to distinguish different low (*i.e.* subexponential) complexities.

References

[1] C. Choffrut and J. Karhumäki, Combinatorics of words, in *Handbook of formal languages* **1**, edited by G. Rosenberg, A. Salomaa. Springer, Berlin (1997) 329–438.
[2] D.M. Cvetković, M. Doob and H. Sachs, *Spectra of graphs. Theory and applications*, 3rd edn. Johann Ambrosius Barth, Heidelberg (1995).
[3] F. D'Alessandro, B. Intrigila and S. Varricchio, On the structure of counting function of sparse context-free languages. *Theor. Comput. Sci.* **356** (2006) 104–117.
[4] A. Ehrenfeucht and G. Rozenberg, A limit theorem for sets of subwords in deterministic TOL languages. *Inform. Process. Lett.* **2** (1973) 70–73.
[5] F.R. Gantmacher, *Application of the theory of matrices*. Interscience, New York (1959).
[6] O. Ibarra and B. Ravikumar, On sparseness, ambiguity and other decision problems for acceptors and transducers. *Lect. Notes Comput. Sci.* **210** (1986) 171–179.

[7] Y. Kobayashi, Repetition-free words. *Theor. Comput. Sci.* **44** (1986) 175–197.
[8] Y. Kobayashi, Enumeration of irreducible binary words. *Discrete Appl. Math.* **20** (1988) 221–232.
[9] A. Lepistö, *A characterization of* 2^+*-free words over a binary alphabet*. Turku Centre for Computer Science, TUCS Tech. Report **74** (1996).
[10] M. Morse and G.A. Hedlund, Symbolic dynamics. *Amer. J. Math.* **60** (1938) 815–866.
[11] A.M. Shur, Combinatorial complexity of rational languages. *Discrete Anal. Oper. Res. 1* **12** (2005) 78–99 (Russian).
[12] A.M. Shur, *On intermediate factorial languages*. Turku Centre for Computer Science, TUCS Tech. Report **723** (2005).