Theoretical Computer Science 412 (2011) 6507–6516

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

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

On the number of Dejean words over alphabets of 5, 6, 7, 8, 9 and 10 letters

Roman Kolpakov [a,*], Michaël Rao [b]

[a] *Moscow State University, Leninskie Gory, 119992 Moscow, Russia*

[b] *LaBRI, Université Bordeaux 1, 351 cours de la libération, 33405 Talence, France*

ARTICLE INFO

Article history:
Received 24 January 2010
Received in revised form 10 January 2011
Accepted 2 August 2011
Communicated by J. Karhumaki

Keywords:
Combinatorics on words
Repetitions
Growth rates

ABSTRACT

We give lower bounds on the growth rate of Dejean words, *i.e.* minimally repetitive words, over a k-letter alphabet, for $5 \le k \le 10$. Put together with the known upper bounds, we estimate these growth rates with the precision of 0.005. As a consequence, we establish the exponential growth of the number of Dejean words over a k-letter alphabet, for $5 \le k \le 10$.

1. Introduction

Let $w = a_1 \cdots a_n$ be a word over an alphabet Σ. The number n is called the *length* of w and is denoted by $|w|$. The symbol a_i of w is denoted by $w[i]$. A word $a_i \cdots a_j$, where $1 \le i \le j \le n$, is called a *factor* of w and is denoted by $w[i : j]$. For any $i = 1, \ldots, n$ the factor $w[1 : i]$ ($w[i : n]$) is called a *prefix* (a *suffix*) of w. A positive integer p is called a *period* of w if $a_i = a_{i+p}$ for each $i = 1, \ldots, n-p$. If p is the minimal period of w, the ratio $e(w) = n/p$ is called the *exponent* of w. Two words w', w'' over Σ are called *isomorphic* if $|w'| = |w''|$ and there exists a bijection $\sigma : \Sigma \longrightarrow \Sigma$ such that $w''[i] = \sigma(w'[i])$, $i = 1, \ldots, |w'|$. By $\mathcal{K}(w)$, we will denote the set of all words over Σ which are isomorphic to the word w. We also denote by $|A|$ the number of elements of a finite set A. Let $|\Sigma| = k$. It is easy to note that $|\mathcal{K}(w)| = k!$ if w contains at least $k-1$ different symbols of Σ.

Let W be an arbitrary set of words. This set is called *factorial* if for any word w from W all factors of w are also contained in W. We denote by $W(n)$ the subset of W consisting of all words of length n. If W is a factorial then it is not difficult to show (see, e.g., [3,1]) that there exists the limit $\lim_{n\to\infty} \sqrt[n]{|W(n)|}$ which is called the *growth rate* of words from W. For any words u, v we denote by $W^{(v)}(n)$ the set of all words from $W(n)$ which contain v as a suffix, and by $W^{(u,v)}(n)$ the set of all words from $W(n)$ which contain v as a suffix and u as a prefix.

One can mean by a repetition any word of exponent greater than 1. The best known example of repetitions is a *square*; that is, a word of the form uu, where u is an arbitrary nonempty word. Avoiding ambiguity,[1] by the *period* of the square uu we mean the length of u. In an analogous way, a *cube* is a word of the form uuu for a nonempty word u, and the *period* of this cube is also the length of u. A word is called *square-free* (*cube-free*) if it contains no squares (cubes) as factors. It is easy to see that there are no binary square-free words of length larger than 3. On the other hand, by the classical results of Thue [20,21], there exist ternary square-free words of arbitrary length and binary cube-free words of arbitrary length. For ternary

* Corresponding author. Tel.: +7 495 939 42 68; fax: +7 495 932 89 52.
E-mail addresses: foroman@mail.ru (R. Kolpakov), rao@labri.fr (M. Rao).

[1] Note that the period of a square is not necessarily the minimal period of this word.

0304-3975/$ – see front matter
doi:10.1016/j.tcs.2011.08.006

square-free words this result was strengthened by Dejean in [9]. She found ternary words of arbitrary length which have no factors with exponents greater than 7/4. On the other hand, she showed that any long enough ternary word contains a factor with an exponent greater than or equal to 7/4. Thus, the number 7/4 is the minimal limit for exponents of avoidable factors which is universally called *the repetition threshold* in arbitrarily long ternary words. Dejean conjectured also that the repetition threshold in arbitrarily long words over a k-letter alphabet is equal to 7/5 for $k = 4$ and $k/(k-1)$ for $k \geq 5$. This conjecture is now proved for any k through the work of several authors [5–8,13,12,15,16].

Denote the repetition threshold in arbitrarily long words over a k-letter alphabet by φ_k. In the paper we will call the words having no factors with exponents greater than φ_k *minimally repetitive* words or *Dejean words*. By $S^{\langle k\rangle}(n)$ we denote the number of all minimally repetitive words of length n over a k-letter alphabet. Note that the set of all minimally repetitive words is obviously factorial. So for any k there exists the growth rate $\gamma^{\langle k\rangle} = \lim_{n\to\infty} \sqrt[n]{S^{\langle k\rangle}(n)}$.

The problem of estimating the number of repetition-free words has been investigated actively during the last decades (reviews of results on the estimations for the number of repetition-free words obtained before 2008 can be found in [2,10]). The most progress in this field has been made for the case of the binary alphabet. In this case Dejean words reduce to overlap-free words which are also a classical object for combinatorial investigations. It is proved in [17] that the growth of the number of binary overlap-free words is polynomial. Actually, binary overlap-free words of each length are counted by a 2-regular function [4].

In [11] we proposed a new approach for obtaining lower bounds on the number of repetition-free words. Using this approach, we obtained precise lower bounds for the growth rates of ternary square-free words, binary cube-free words, and ternary minimally repetitive words. This approach proved to be very effective. In particular, in [19] Shur proposed an interesting modification of our approach which allows to compute more effectively lower bounds for the growth rates of words which contain no repetitions of exponent greater than or equal to a given bound if this bound is not less than 2. The direction of our further investigations in this field is testing the proposed approach for "extreme" cases when the prohibitions imposed on words are maximal possible for the existence of words of arbitrary length avoiding these prohibitions. These cases are obviously the most difficult for obtaining lower bounds on the number of appropriate words. The case of minimally repetitive words is a natural example of such "extreme" cases. Moreover, the general case of minimally repetitive words over a k-letter alphabet for $k \geq 5$ when $\varphi_k = k/(k-1)$ is the most interesting for us. So this paper is devoted to obtaining lower bounds on $\gamma^{\langle k\rangle}$ for $k \geq 5$ by using the proposed approach. Note that the method proposed in [11] is not directly applicable to resolving this problem because of the huge size of required computer computations. In this paper we propose an improvement of this method which requires significantly fewer computer computations. Using this improvement, we obtain lower bounds on $\gamma^{\langle k\rangle}$ for $5 \leq k \leq 10$ which have the precision of 0.005. As an evident consequence of these results, we establish the exponential growth of the number of minimally repetitive words over a k-letter alphabet for $5 \leq k \leq 10$ (for $k = 3, 4$ this fact was proved by Ochem in [14]).

2. Estimation for the number of minimally repetitive words

2.1. General

For obtaining a lower bound on $\gamma^{\langle k\rangle}$ we will consider the alphabet $\Sigma_k = \{a_1, a_2, \ldots, a_k\}$ where $k \geq 5$. We denote the set of all minimally repetitive words over Σ_k by $\mathcal{F}$. By a prohibited factor we mean a factor with an exponent greater than $k/(k-1)$. Let m be a natural number, $m > k$, and w', w'' be two words from $\mathcal{F}(m)$. We call the word w'' a *descendant* of the word w' if $w'[2:m] = w''[1:m-1]$ and $w'w''[m] = w'[1]w'' \in \mathcal{F}(m+1)$. The word w' is called in this case an *ancestor* of the word w''. We introduce a notion of closed words in the following inductive way. A word w from $\mathcal{F}(m)$ is called *right closed* (*left closed*) if and only if this word satisfies one of the two following conditions:

(a) **Basis of induction.** w has no descendants (ancestors);

(b) **Inductive step.** All descendants (ancestors) of w are right closed (left closed).

A word is *closed* if it is either right closed or left closed. We denote by $\hat{\mathcal{F}}(m)$ the set of all words from $\mathcal{F}(m)$ which are not closed. By $\mathcal{L}_m$ we denote the set of all words over Σ_k such that the length of these words is not less than m and all factors of length m in these words belong to $\hat{\mathcal{F}}(m)$. We also denote by $\mathcal{F}_m$ the set of all minimally repetitive words from $\mathcal{L}_m$. Note that a word w is closed if and only if any word isomorphic to w is also closed. So we have the following obvious fact.

Proposition 1. *For any isomorphic words w', w'' and any $n \geq |w'|$ the equality $|\mathcal{F}_m^{(w')}(n)| = |\mathcal{F}_m^{(w'')}(n)|$ holds.*

A word will be called *rarefied* if the distance between any two different occurrences of the same symbol in this word is not less than $k-1$.

Proposition 2. *Any word from $\mathcal{L}_m$ is rarefied.*

Proof. Let w be an arbitrary word from $\mathcal{L}_m$. Assume that $w[i] = w[j]$ where $j < i \leq j + (k-2)$. Consider the factor $f = w[j:i]$. Since $|f| = i - j + 1 \leq k - 1 < m$, in w the factor f is contained in some factor f' of length m. By the definition of $\mathcal{L}_m$ we have $f' \in \mathcal{F}(m)$, so $f \in \mathcal{F}$. On the other hand, f has the period $|f| - 1$, so

$$e(f) \geq \frac{|f|}{|f|-1} = \frac{i-j+1}{i-j} \geq \frac{k-1}{k-2} > \frac{k}{k-1}$$

which contradicts the definition of $\mathcal{F}(m)$. $\square$