Journal of Combinatorial Theory, Series A 121 (2014) 34–44

Contents lists available at ScienceDirect

Journal of Combinatorial Theory, Series A

www.elsevier.com/locate/jcta

Lyndon words and Fibonacci numbers

Kalle Saari

Department of Mathematics and Statistics, University of Winnipeg, 515 Portage Avenue, Winnipeg, MB R3B 2E9, Canada

ARTICLE INFO

Article history:
Received 16 November 2012
Available online 26 September 2013

Keywords:
Lyndon word
Fibonacci word
Central word
Golden ratio
Sturmian word
Periodicity

ABSTRACT

It is a fundamental property of non-letter Lyndon words that they can be expressed as a concatenation of two shorter Lyndon words. This leads to a naive lower bound $\lceil \log_2(n) \rceil + 1$ for the number of distinct Lyndon factors that a Lyndon word of length n must have, but this bound is not optimal. In this paper we show that a much more accurate lower bound is $\lceil \log_\phi(n) \rceil + 1$, where ϕ denotes the golden ratio $(1 + \sqrt{5})/2$. We show that this bound is optimal in that it is attained by the Fibonacci Lyndon words. We then introduce a mapping $\mathcal{L}_\mathbf{x}$ that counts the number of Lyndon factors of length at most n in an infinite word $\mathbf{x}$. We show that a recurrent infinite word $\mathbf{x}$ is aperiodic if and only if $\mathcal{L}_\mathbf{x} \geqslant \mathcal{L}_\mathbf{f}$, where $\mathbf{f}$ is the Fibonacci infinite word, with equality if and only if $\mathbf{x}$ is in the shift orbit closure of $\mathbf{f}$.

1. Introduction

Lyndon words are primitive words that are the lexicographically smallest words in their conjugacy classes [19]. Originally defined in the context of free Lie algebras [6], Lyndon words have shown to be a useful tool for a variety of problems in combinatorics ranging from the construction of de Bruijn sequences [16] to proving the optimal lower bound for the size of uniform unavoidable sets [5]. One of the fundamental properties of Lyndon words is their recursive nature: if w is a non-letter Lyndon word, then there exist two shorter Lyndon words u and v such that $w = uv$ [6]. This implies that the number of different Lyndon factors of w is bounded below by $\lceil \log_2 |w| \rceil + 1$, but a little experimentation shows that this is hardly optimal. One of the results of this paper, Corollary 1, is that a much better lower bound is $\lceil \log_\phi |w| \rceil + 1$, where ϕ denotes the golden ratio $(1 + \sqrt{5})/2$. Here the base of the logarithm is optimal, because the Fibonacci Lyndon words attain the lower bound.

E-mail address: kasaar2@gmail.com.

0097-3165/$ – see front matter
http://dx.doi.org/10.1016/j.jcta.2013.09.002

This follows from Theorem 1, in which we show that if w is a Lyndon word with $|w| \geqslant F_n$, where F_n is the nth Fibonacci number, then the number of distinct Lyndon factors in w is at least n with equality if and only if w equals one of the two Fibonacci Lyndon words of length F_n, up to renaming letters.

It also makes sense to count the number of Lyndon factors of infinite words, but here we have to use caution: if an infinite word is aperiodic, it will have infinitely many Lyndon factors, as we will show in Corollary 2. Thus we define a mapping $\mathcal{L}_{\mathbf{x}} : \mathbb{N} \to \mathbb{N}$ for which $\mathcal{L}_{\mathbf{x}}(n)$ is the number of distinct Lyndon words of length at most n occurring in a given infinite word $\mathbf{x}$. Of special importance is the Fibonacci infinite word $\mathbf{f}$. Our first main result in this setting, Theorem 3, is that if $\mathbf{x}$ is aperiodic, then $\mathcal{L}_{\mathbf{x}} \geqslant \mathcal{L}_{\mathbf{f}}$. As Lyndon words are unbordered, this is an improvement of a classic result by Ehrenfeucht and Silberger [13] stating only that an aperiodic infinite word must have arbitrarily long unbordered factors. If we confine our realm to recurrent infinite words, then the above result can be improved as follows. In Theorem 4 we show that a recurrent infinite word $\mathbf{x}$ is aperiodic if and only if $\mathcal{L}_{\mathbf{x}} \geqslant \mathcal{L}_{\mathbf{f}}$ with equality if and only if $\mathbf{x}$ is in the shift orbit closure of the Fibonacci word $\mathbf{f}$, up to renaming letters.

Fibonacci words are sort of a universal optimality prover in that they possess a wide range of extremal properties, see e.g. [3,7,11,21,17,15]. The problem of the enumeration of Lyndon factors in automatic and linearly recurrent sequences has recently been studied in [9].

2. Preliminaries

In this section we establish the notation of this paper and present some preliminary results. We assume the reader is familiar with the usual terminology of words and languages as given in [1] or [20].

Let $\mathcal{A}$ be a finite, nonsingular alphabet totally ordered by $<$; thus every pair of distinct letters $\mathtt{a}, \mathtt{b} \in \mathcal{A}$ satisfy either $\mathtt{a} < \mathtt{b}$ or $\mathtt{b} < \mathtt{a}$, but not both. We use the same symbol '$<$' to denote the usual order relation among the integers, but this should not cause problems as the context always tells which order is meant. In what follows, we sometimes assume that $\mathtt{0}, \mathtt{1} \in \mathcal{A}$, sometimes $\mathtt{a}, \mathtt{b} \in \mathcal{A}$, and then their mutual order is implicitly assumed to be their "natural order," so that $\mathtt{0} < \mathtt{1}$ and $\mathtt{a} < \mathtt{b}$.

The set of all finite words over $\mathcal{A}$ is denoted by $\mathcal{A}^*$ and the set of finite words excluding the empty word ε is denoted by $\mathcal{A}^+$.

Let $w = a_1 a_2 \ldots a_n$ be a nonempty finite word with $a_i \in \mathcal{A}$ and $n \geqslant 1$. The *length* of w is $|w| = n$; we denote the cardinality of a set X by $\#X$. The *reversal* of w is the word $w^R = a_n a_{n-1} \ldots a_1$. If $w^R = w$, then w is a *palindrome*. The word w has period $p \geqslant 1$ if $a_{i+p} = a_i$ for all $i = 1, 2, \ldots, n - p$. According to this definition, any integer $p \geqslant n$ is a period of w. If $p \leqslant n$, then p is a period of w if and only if there exist words $x, y, z \in \mathcal{A}^*$ such that $w = xy = zx$ and $|y| = |z| = p$. If w has no periods smaller than $|w|$, then it is called *unbordered*, otherwise w is *bordered*. Suppose that $w = pfs$ with $p, f, s \in \mathcal{A}^*$. Then p, f and s, any of which may be empty, are respectively called a *prefix*, *factor*, and *suffix* of w. In addition, p and s are *proper* prefix and suffix if they do not equal w. We say that a word $z \in \mathcal{A}^+$ is a *periodic extension* of w if z is a prefix of a word in w^+. We abuse the word "extension" here in that we allow an "extension" to be a prefix of w. The word we get from w by deleting its last letter is denoted by $w^\flat$; thus $w^\flat = a_1 a_2 \ldots a_{n-1}$. Also if $w = xy$ for some words x, y, we denote $x^{-1} w = y$ and $w y^{-1} = x$. The word w is *primitive* if it cannot be written in the form $w = u^k$ for a word $u \in \mathcal{A}^+$ and an integer $k \geqslant 2$. If $w = uv$, then the word vu is called a *conjugate* of w. The set of all conjugates of w is called the *conjugacy class* of w.

Lemma 1. *(See Castelli, Mignosi, and Restivo [4].) Let $w \in \mathcal{A}^+$ be a word with periods p, q.*

(i) *If $q < p \leqslant |w|$, then the prefix and suffix of w of length $|w| - q$ have periods q and $p - q$.*
(ii) *Let u and v be the prefix and suffix of w of length q, respectively. Then uw and wv have periods q and $p + q$.*

In property (ii) in the previous lemma, the indicated source [4] only mentions and proves the claim for the periods of uw, but the case for the periods of wv can be proved similarly.