RAIRO-Theor. Inf. Appl. 43 (2009) 165–178
DOI: 10.1051/ita:2008006

Available online at:
www.rairo-ita.org

LEAST PERIODS OF FACTORS OF INFINITE WORDS *, **

James D. Currie[1] and Kalle Saari[2]

Abstract. We show that any positive integer is the least period of a factor of the Thue-Morse word. We also characterize the set of least periods of factors of a Sturmian word. In particular, the corresponding set for the Fibonacci word is the set of Fibonacci numbers. As a by-product of our results, we give several new proofs and tightenings of well-known properties of Sturmian words.

Mathematics Subject Classification. 68R15.

1. Introduction

The combinatorial study of infinite words often entails considering periods of factors. For example, showing that an infinite word has a bounded critical exponent requires showing, perhaps implicitly, that the ratio between a factor and its least period is bounded. Therefore it seems natural to study directly the set of least periods of factors of an infinite word; we call this set the *period set* of an infinite word. To our knowledge, this is a novel area of inquiry into the periodicity of finite and infinite words [11], Chapter 8.

This paper initiates the study of the period set of infinite words. It is easy to see that the period set of an infinite word is finite if and only if the word is purely periodic. Therefore infinite aperiodic words give rise to infinite period sets, and it is natural to ask what kind of restrictions period sets have to obey. It is plain that the period set of an aperiodic infinite word must include periods 1, 2, and 3. But already 4 is avoidable, as is witnessed by the Fibonacci word, see Corollary 4.

Keywords and phrases. Periodicity, Fibonacci word, Thue-Morse word, Sturmian word.

* *Work of the first author supported by a Discovery Grant from NSERC.*

** *Work of the second author supported by the Finnish Academy under grant 8206039.*

[1] Department of Mathematics & Statistics, University of Winnipeg, Winnipeg, R3B2E9, Canada

[2] Department of Mathematics and Turku Centre for Computer Science, University of Turku, Turku, Finland; kasaar@utu.fi

Article published by EDP Sciences

In this paper we will characterize the period sets of the Thue-Morse word and of all Sturmian words. These much studied words have applications and connections to several fields, such as algebra, number theory, ergodic theory, crystallography, computer graphics, and text algorithms; see [1] and [11], Chapter 2 and the references therein. The characterizations of the period sets show that the gaps in the period set of the Fibonacci word grow exponentially, while the gaps in the period set of the Thue-Morse word have the lowest possible growth an aperiodic infinite word can have. As a by-product of our work, we give new proofs, tightenings, and generalizations of some known properties of Sturmian words.

An outline of this paper is as follows: in Section 2, we set the terminology used in the paper, and mention some basic results. In Section 3, we show that any positive integer is the least period of some factor of the Thue-Morse word. In Section 4, we characterize the set of least periods of a Sturmian word. Finally, in Section 5, we give four applications of our results, including a tightening of a result by de Luca and De Luca [7] and a characterization of the least periods of standard words.

2. Preliminaries

In this section we briefly define the terminology used in this work. For a statement without a citation in this section, we refer to [4,10,11].

We will be dealing with words over the alphabet $\{0,1\}$. The set of all such words, including the empty word, is denoted by $\{0,1\}^*$.

Let $w = a_1a_2\cdots a_n$ be a word with $a_i \in \{0,1\}$ and $n \geq 1$. The *length* of w is the integer n, and is denoted by $|w|$. We denote the number of occurrences of a letter $a \in \{0,1\}$ in w by $|w|_a$.

A *factor* of w is a word of the form $u = a_ia_{i+1}\cdots a_k$ with $1 \leq i \leq k \leq n$. It is a *prefix* if $i = 1$ and a *suffix* if $k = n$. In each case, we add the attribute *proper* if $w \neq u$.

Let $0 \leq i < |w|$. The word $a_{i+1}a_{i+2}a_{i+3}\cdots a_na_1a_2\cdots a_i$ is called a *conjugate* of w, and is denoted by $\sigma^i(w)$.

We write $w' = a_1a_2\cdots a_{n-1}$ and $\sim w = a_2a_3\cdots a_n$. The *reverse* of w is the word $a_na_{n-1}\cdots a_1$, and we denote it by w^R. We denote by $\overline{w}$ the word obtained from w by exchanging 0's and 1's; it is called the *complement* of w.

A *period* of the word w is an integer $p \geq 1$ such that, for all $i = 1, 2, \ldots, n-p$, we have $a_i = a_{i+p}$. The word w is said to be a *rational power* of the word $u = a_1a_2\cdots a_p$, and u is called a *word period* of w. If no period p divides the length of w, then w is termed *primitive*. Primitivity of a word implies that all conjugates of the word are distinct.

In this work, we are interested in the *least* period of a word w, which we denote by $p(w)$. The word w is called *unbordered* if $p(w) = |w|$. Finally, the prefix of w of length $p(w)$ is called the *fractional root* of w.

Let the words in $\{0,1\}^*$ be ordered by the lexicographic order induced by the relation $0 < 1$. If w is a primitive word, then its least conjugate with respect to

the lexicographic order is called a *Lyndon word*. If $|w| > 1$, we get a different Lyndon word by using the lexicographic order induced by the relation $1 < 0$. One of the basic properties of a Lyndon word is that it is unbordered.

Let $\mathbf{x}$ be an infinite word, that is, a mapping from the nonnegative integers to a finite alphabet. The notion of a factor is extended naturally to infinite words with the agreement that a factor is always a finite word. The set of finite factors of $\mathbf{x}$ is denoted by $F(\mathbf{x})$. We call the set of least periods of all factors of $\mathbf{x}$ the *period set* of $\mathbf{x}$.

A *morphism* is a mapping $h\colon \{0,1\}^* \to \{0,1\}^*$ with the property that $h(uv) = h(u)h(v)$ for every $u, v \in \{0,1\}^*$. The domain of h extends to infinite words such that if

$$\mathbf{x} = a_1 a_2 \cdots a_n \cdots, \quad \text{then} \quad h(\mathbf{x}) = h(a_1)h(a_2)\cdots h(a_n)\cdots$$

The *Thue-Morse word*, denoted by $\mathbf{t}$, is the infinite word starting with the letter 0 that is a fixed point of the morphism $\mu\colon \{0,1\}^* \to \{0,1\}^*$ determined by $\mu\colon 0 \mapsto 01, 1 \mapsto 10$. If u is a factor of $\mathbf{t}$, then so are $\overline{u}$ and u^R. The Thue-Morse word is *overlap-free*, which means that $\mathbf{t}$ does not have a factor of the form uua, where u is a nonempty word and a is the first letter of u.

A *Sturmian word* is an infinite word $\mathbf{x}$ over $\{0,1\}$ such that, for every integer $n \geq 1$, the word $\mathbf{x}$ has precisely $n+1$ different factors of length n. The frequency of letters 0 and 1 in $\mathbf{x}$ exists; the frequency of 1 is called the *slope* of $\mathbf{x}$, and we denote it by θ. The slope θ is an irrational number, and therefore it has an infinite continued fraction expansion

$$\theta = [\,0, d_1 + 1, d_2, d_3, \dots\,], \tag{1}$$

where $d_1 \geq 0$ and $d_n \geq 1$ for $n \geq 2$.

Next we define words s_n corresponding to the expansion (1) as follows:

$$s_{-1} = 1, \qquad s_0 = 0, \qquad s_n = s_{n-1}^{d_n} s_{n-2} \qquad (n \geq 1).$$

Words that can be recursively defined as above are called *standard*. All standard words are primitive. Furthermore, consecutive standard words s_{n-1} and s_n are near-commutative in the following sense: if $n \geq 1$, then there exists a word p_n such that

$$s_n s_{n-1} = p_n a\bar{a} \qquad \text{and} \qquad s_{n-1}s_n = p_n \bar{a}a, \tag{2}$$

where $a \in \{0,1\}$. Therefore, for $n \geq 2$, we have

$$s_n s''_{n-1} = s_{n-1} s''_n. \tag{3}$$

Let us denote $q_n = |s_n|$ for all $n \geq -1$.

The standard words corresponding to the slope θ are related to $\mathbf{x}$ in the following way. Since s_n is a prefix of s_{n+1} for all $n \geq 1$, there is a unique infinite word, which we denote by $\mathbf{c}$, such that s_n is a prefix of $\mathbf{c}$ for all $n \geq 1$. The word $\mathbf{c}$ is called the *characteristic word* with slope θ. The sets of finite factors of $\mathbf{x}$ and $\mathbf{c}$ coincide, that is, we have $F(\mathbf{c}) = F(\mathbf{x})$.

Since $\mathbf{x}$ has $n+1$ factors of length n, it follows that there exist precisely one factor u of length n such that both $0u$ and $1u$ are factors of $\mathbf{x}$. Such a factor is called *left special*. A factor of $\mathbf{x}$ is left special if and only if it is a prefix of $\mathbf{c}$.

The set of factors of $\mathbf{x}$ is closed under reversal, that is to say, if $u \in F(\mathbf{x})$, then also $u^R \in F(\mathbf{x})$.

Now we will adopt a notation from [13]. For each integer $n \geq 1$, there exists a unique representation

$$n = d_1 + d_2 + \cdots + d_{i-1} + j, \qquad 1 \leq j \leq d_i.$$

With this representation, we denote

$$t_n = s_{i-1}^j s_{i-2}. \tag{4}$$

It is also useful to denote $t_{-1} = 1$ and $t_0 = 0$. Observe that $t_{d_1+\cdots+d_n} = s_n$ for all $n \geq 1$.

The following result by Berstel [2] is one of the key observations we need in characterizing the period set of a Sturmian word.

Theorem 1 (Berstel)**.** *For $n \geq 2$, the longest prefix of $\mathbf{c}$ that is a rational power of the word s_n is $s_n^{d_{n+1}+1} s''_{n-1}$.*

3. Periods of factors of the Thue-Morse word

In this section we will show that every positive integer is the least period of some factor of the Thue-Morse word. To do that, we need some auxiliary results. Recall that μ denotes the morphism given by $\mu\colon 0 \mapsto 01, 1 \mapsto 10$.

Lemma 1. *Let u be a factor of the Thue-Morse word $\mathbf{t}$. Then u does not have any odd period p such that $p < |u| - 3$.*

Proof. Suppose that u has an odd period p with $p < |u| - 3$. We may suppose that $p \geq 3$ because $\mathbf{t}$ does not contain 000 or 111. Then $|u| \geq 7$. Let v be the prefix of u of (odd) length $p + 4$.

Observe that, since v is a factor of $\mathbf{t}$, also v^R and $\overline{v}$ are factors of $\mathbf{t}$. Therefore, without loss of generality, replacing v by its reversal or complement or both if necessary, write $v = \mu(w)a = v_0 v_1 v_2 \cdots v_{p+3}$, where v_i, $a \in \{0, 1\}$, and $v_0 = 0$.

Since v has period p, we find that $v_p = v_0 = 0$, so that $v_{p-1}v_p = 10$. Similarly, $v_{p+1} = v_1 = 1$, so that $v_{p+1}v_{p+2} = 10$. Thus $v_2 = v_{p+2} = 0$, whence $v_2 v_3 = 01$. This implies that $v_{p+3} = v_3 = 1$, and v contains the overlap $v_{p-1}v_p v_{p+1}v_{p+2}v_{p+3} = 10101$, which is impossible because $\mathbf{t}$ is overlap-free. □

Recall that we denote by $\sim u$, u' and $\sim u'$ the words obtained from u by deleting respectively the first letter, the last letter, or the first and last letters.

Lemma 2. *Let $u = \mu(w)$, some $w \in \{0, 1\}^+$. Let $v = \sim u'$. Suppose that v has an even period $2r < |v|$. Then w has period r.*

Proof. Write $w = w_0w_1w_2\cdots w_{s-1}w_s$ and $v = v_0v_1v_2\cdots v_{2s-1}$, so that $r < s$. We see that

$$v = \overline{w}_0w_1\overline{w}_1w_2\overline{w}_2\cdots w_{s-1}\overline{w}_{s-1}w_s.$$

Since v has period $2r$, we have $\overline{w}_i = v_{2i} = v_{2i+2r} = \overline{w}_{i+r}$ whenever $0 \le 2i + 2r \le |v| - 2$, that is, $0 \le i \le s - 1 - r$. Therefore,

$$w_i = w_{i+r} \quad \text{for all } 0 \le i \le s - 1 - r.$$

Similarly, since v has period $2r$, we have $w_i = v_{2i-1} = v_{2i-1+2r} = w_{i+r}$ whenever $0 \le 2i - 1 + 2r \le |v| - 1$, that is, $1 \le i \le s - r$. In total,

$$w_i = w_{i+r} \quad \text{for all } 0 \le i \le s - r. \qquad \square$$

The claim in the lemma above does not hold if we allow $|v| = 2r$. Indeed, if $w = 01$, then $v = 11$. Even though 2 is plainly a period of v, the word w certainly does not have period 1.

Corollary 1. *Let $u = \mu(w)$, some $w \in \{0,1\}^+$. Let v be obtained from u by possibly deleting first or last or both letters; that is, let v be one of u, u', $\sim u$, $\sim u'$. Then v has period $2r < |u| - 2$ if and only if w has period r.*

Proof. Suppose that v has period $2r$. Then $\sim u'$ is a factor of v and has period $2r$, so that, by Lemma 2, w has period r.

If w has period r, then $\mu(w)$ has period $2r$ since $|\mu(0)| = |\mu(1)| = 2$. It follows that the factor v of $\mu(w)$ has period $2r$. $\square$

Lemma 3. *Let $r \ge 4$ be a positive integer. Then the following statements hold:*

(i) *if $r \equiv 4 \pmod 6$, then $\mathbf{t}$ has a factor u of the form $u = 00y11$ with $|u| = r$ and $p(u) = r$;*
(ii) *if $r \equiv 0, 2, 3$, or $5 \pmod 6$, then $\mathbf{t}$ has a factor u of the form $u = 00y101$ with $|u| = r$ and $p(u) = r$;*
(iii) *if $r \equiv 0, 1$, or $3 \pmod 6$, then $\mathbf{t}$ has a factor u of the form $u = 00y010$ with $|u| = r + 1$ and $p(u) = r$.*

Proof. We prove this by induction. The item (i) with $r = 4$ is witnessed by the factor 0011. The item (ii) with $r = 5, 6, 8$, or 9 is witnessed by factors

$$00101, \quad 001101, \quad 00101101, \quad \text{and} \quad 001100101.$$

The item (iii) with $r = 6, 7, 9$ is witnessed by factors

$$0011010, \quad 00110010, \quad \text{and} \quad 0011010010.$$

Let us now assume that $r \ge 10$, and that the lemma is satisfied for all smaller values of r.

Case 1. $r \equiv 0 \pmod 6$. First, let $s = r/2$. Then either $s \equiv 0$, or $3 \pmod 6$, and $s < r$. By the minimality of r and the item (iii), there is a factor w of $\mathbf{t}$ of the form

$00z010$ having length $s+1$ and least period s. Let $u = \sim\mu(\overline{w}^R)$. Then u is a factor of $\mathbf{t}$, it is of length $r+1$, and it has the form $u = 00y010$, where $y = 110\mu(\overline{z}^R)1$.

Evidently, u has period r. Corollary 1 implies that u has no even period shorter than $r = 2s$. Writing $u = u_0u_1u_2\cdots u_r$, we see that $u_0 \neq u_{r-1}$, $u_2 \neq u_r$, $u_3 \neq u_r$, showing that u does not have period $r-1$, $r-2$, or $r-3$. By Lemma 1, u can have no odd period, and therefore the least period of u is r, witnessing the item (iii).

Next, let $v = \sim\mu(\overline{w}^R)'$. Then v is of length r, and thus has period r. Furthermore, $v = \sim\mu(101\overline{z}^R11)'$ has the form $00y101$, where $y = 110\mu(\overline{z}^R)$. It has no even period shorter than $r = 2s$ by Corollary 1. Writing $v = v_0v_1v_2\cdots v_{r-1}$, we see that $v_0 \neq v_{r-1}$, $v_1 \neq v_{r-1}$, $v_0 \neq v_{r-3}$, showing that v does not have period $r-1$, $r-2$, or $r-3$. By Lemma 1, v can have no odd period. Thus the least period of v is r, witnessing the item (ii).

Case 2. $r \equiv 3 \pmod 6$. First, let $s = (r+3)/2$. Then either $s \equiv 0$, or $3 \pmod 6$, and $s < r$. Thus there is a factor w of $\mathbf{t}$ of the form $00z101$ having length s and least period s. Let $u = \sim\mu(w^R)'$. Then u is a factor of $\mathbf{t}$, it is of length $r+1$, and it is of the form $u = 00y010$, where $y = 110\mu(z^R)$.

Evidently, the word u has period r. Corollary 1 implies that it has no even period strictly shorter than $|u| = r+1 = 2s-2$. Writing $u = u_0u_1u_2\cdots u_r$, we see that $u_0 \neq u_{r-1}$, $u_1 \neq u_{r-1}$, $u_3 \neq u_r$, showing that u does not have period $r-1$, $r-2$, or $r-3$. By Lemma 1, u can have no odd period less than r. Thus the least period of u is r, witnessing the item (iii).

Next, let $s = (r+1)/2$. Then either $s \equiv 2$, or $5 \pmod 6$. There is a factor v of $\mathbf{t}$ of the form $00z101$ having length s and least period s. Let $u = \sim\mu(v^R)$. Then u is a factor of $\mathbf{t}$, it is of length r, and it has the form $u = 00y101$, where $y = 110\mu(z^R)0$.

Evidently, the word u has period $|u| = r$. Corollary 1 implies that it has no even period strictly shorter than $r+1 = 2s$. Writing $u = u_0u_1u_2\ldots u_{r-1}$, we see that $u_0 \neq u_{r-1}$, $u_1 \neq u_{r-1}$, $u_0 \neq u_{r-3}$, showing that u does not have period $r-1$, $r-2$, or $r-3$. By Lemma 1, u can have no odd period less than r. Thus the least period of u is r, witnessing the item (ii).

Case 3. $r \equiv 1 \pmod 6$. Let $s = (r+3)/2$. Then either $s \equiv 2$, or $5 \pmod 6$, and $s < r$. Thus there is a factor v of $\mathbf{t}$ of the form $00z101$ having length s and least period s. Let $u = \sim\mu(v^R)'$. As in the previous case, u is a factor of $\mathbf{t}$, it is of length $r+1$, has the form $u = 00y010$, and its least period equals $r = |u|-1$, witnessing the item (iii).

Case 4. $r \equiv 4 \pmod 6$. Let $s = r/2$. Then either $s \equiv 2$, or $5 \pmod 6$. There is a factor v of $\mathbf{t}$ of the form $00z101$ having length s and least period s. The word $v = 00z101$ must be obtained by deleting the first and possibly last letter of some word $\mu(t)$, where t is some factor of $\mathbf{t}$. Let u denote a word of the form $u = 1001x101$ that is obtained from $\mu(t)$ by possibly deleting the last letter.

Next we will show that u has no even period less than s. To derive a contradiction, suppose that u has period $2k < s$. Then by Corollary 1, the word t has period k. But then, again by Corollary 1, the word v has a period $2k < s$, a contradiction.

Writing $u = u_0u_1 \cdots u_s$, we see that $u_1 \neq u_s$, $u_2 \neq u_s$, $u_1 \neq u_{s-2}$, so that u does not have period $s-1$, $s-2$, or $s-3$. Therefore u has no odd period less than s, and it follows that its least period is s.

We now let $w = \sim\mu(u)' = 00y11$, where $y = 10110\mu(x)100$. Then w is of length r. The same argument used before shows that w has no even period less than r. Writing $w = w_0w_1 \cdots w_{r-1}$, we see that $w_0 \neq w_{r-1}$, $w_0 \neq w_{r-2}$, $w_1 \neq w_{r-2}$, and so w has no odd period less than r either. Therefore the least period of w is r, witnessing the item (i), as desired.

Case 5. $r \equiv 2 \pmod 6$. Let $s = r/2$. Then we have two possibilities.

If $s \equiv 1 \pmod 6$, then $\mathbf{t}$ has a factor $w = 00z010$ of length $s+1$, minimum period s. Let $v = \sim\mu(\overline{w}^R)'$. Then v has form $00y101$ with length r and least period r, as can be seen as above.

If $s \equiv 4 \pmod 6$, then $\mathbf{t}$ has a factor of the form $00z11$ having length s and least period s. It follows that $u = 100z11$ is a factor of $\mathbf{t}$ having length $s+1$ and its least period is s. Let $w = \sim\mu(u)'$. Then $w = 00y101$, where $y = 101\mu(z)$. As in previous cases, the word w is of length r, and its least period is r, witnessing the item (ii).

Case 6. $r \equiv 5 \pmod 6$. Let $s = (r+1)/2$. Then either $s \equiv 0$, or $3 \pmod 6$, and $s < r$. Therefore, $\mathbf{t}$ has a factor of the form $v = 00z101$ having length s and least period s. It follows that $u = \sim\mu(v^R)$ has the form $u = 00y101$ where $y = 110\mu(z^R)0$. As in previous cases, u is of the length r, and the least period of u is r, witnessing the item (ii). □

Remark 1. The previous lemma shows that the Thue-Morse word has an unbordered factor for each length $r \not\equiv 1 \pmod 6$. It is readily verified that all factors of length 7 are bordered. Since the factors of length 1 are trivially unbordered, it is natural to ask, for which lengths $r \equiv 1 \pmod 6$ are all factors of length r bordered. This question remains open.

We are ready for the main theorem of this section.

Theorem 2. *For each integer $n \geq 1$, the Thue-Morse word has a factor of least period n.*

Proof. The least periods 1,2,3 are displayed by factors 0, 01, and 001. For integers $n \geq 4$, appropriate factors exist according to Lemma 3. □

4. Periods of factors of Sturmian words

In this section we will characterize the period sets of all Sturmian words, and by doing so, we obtain a few older results on Sturmian words as a by-product in the next section.

Let $\mathbf{x}$ be a Sturmian word with slope θ. Denote the continued fraction expansion of θ by

$$\theta = [\,0, d_1 + 1, d_2, d_3, \dots\,]. \tag{5}$$

Let $(s_n)_{n\geq -1}$ be the corresponding sequence of standard words, and let $(t_m)_{m\geq -1}$ denote the corresponding auxiliary words defined in (4). Further, let $\mathbf{c}$ denote the characteristic sequence with slope θ. Observe that $d_1 \geq 0$ and $d_n \geq 1$ for all $n \geq 2$. Since the period set of a sequence does not depend on the naming of letters, we may assume that $\mathbf{c}$ begins with 0. Therefore, *we assume in the rest of this section that* $d_1 \geq 1$.

Lemma 4. *For $n \geq 0$, the word s_n^2 is a factor of $\mathbf{x}$. For $m \geq d_1$, the word t_m^2 is a factor of $\mathbf{x}$.*

Proof. The word $s_{n+1}^{d_{n+2}} s_n s_{n+1}$ is a prefix of s_{n+3}, and therefore a factor of $\mathbf{x}$. Since $n \geq 0$ (and $d_1 \geq 1$), the word s_n is a prefix of s_{n+1}. Consequently, s_n^2 is a factor of $\mathbf{x}$.

If $m = d_1$, then $t_m^2 = s_1^2$ occurs in $\mathbf{x}$. So, we may suppose that $m > d_1$. Then we have $t_m = s_n^i s_{n-1}$ for some integers $n \geq 1$ and $1 \leq i \leq d_{n+1}$. Since s_{n+1}^2 occurs in $\mathbf{x}$, we see that the word $s_n^i s_{n-1} s_n^{d_{n+1}} s_{n-1}$ occurs in $\mathbf{x}$. Since $n \geq 1$, the word s_{n-1} is a prefix of s_n, and hence it follows that the square of the word $t_m = s_n^i s_{n-1}$ occurs in $\mathbf{x}$. □

Corollary 2. *For $m \geq -1$, all conjugates of t_m are factors of $\mathbf{x}$.*

Proof. The claim is trivial if m equals -1 or 0. When $1 \leq m < d_1$, the claim is witnessed by $s_1^2 = 0^{d_1}10^{d_1}1$. When $m \geq d_1$, the word t_m^2 occurs in $\mathbf{x}$, and so the claim obviously holds then as well. □

The words t_m clearly are standard, and hence primitive. Therefore all the conjugates of t_m are distinct. Since all conjugates of t_m are factors of $\mathbf{x}$, and $\mathbf{x}$ has $|t_m| + 1$ factors of length $|t_m|$, it follows that $\mathbf{x}$ has precisely one factor of length $|t_m|$ that is not a conjugate of t_m. We call this factor the *singular factor of $\mathbf{x}$ corresponding to t_m*[1]. With this definition, $t_{-1} = 1$ is the singular factor corresponding to $t_0 = 0$, and vice versa. We give the other singular factors in the next lemma.

Lemma 5. *Let $m \geq 1$, and let a denote the last letter of $t_m = s_n^i s_{n-1}$. The singular factor corresponding to t_m equals $\overline{a}t'_m$, and it is bordered with period q_n.*

Proof. First, observe that $n \geq 0$ and $1 \leq i \leq d_{n+1}$. It is clear that $s_{n+2}s_{n+1}$ is a prefix of $\mathbf{c}$, and hence a factor of $\mathbf{x}$. Since

$$s_{n+2}s_{n+1} = s_{n+1}^{d_{n+2}} s_n^{d_{n+1}+1} s_{n-1}, \tag{6}$$

we see that the word $s_n^{i+1} s_{n-1}$ is a factor of $\mathbf{x}$.

First, suppose that $n = 0$. Then the word 0^{i+1} occurs in $\mathbf{x}$, and it clearly is the singular factor corresponding to $s_n^i s_{n-1} = 0^i 1$. The claim holds for 0^{i+1}.

[1]Singular factors for Sturmian words seem to have been introduced by Cao and Wen [3], but only in cases that correspond to the words s_n.

Next, suppose that $n = 1$. Then

$$s_1^{i+1}s_0 = 0^{d_1}1\big(0^{d_1}1\big)^i 0,$$

and hence the word $1\big(0^{d_1}1\big)^i$ occurs in $\mathbf{x}$, and it clearly is the singular factor corresponding to $s_1^i s_0$, satisfying the claim.

Finally, suppose that $n \geq 2$. Let us denote $s_n = s_n'' ab$ and $s_{n-1} = s_{n-1}'' ba$, where $ab \in \{01, 10\}$. Equation (6) shows that the word $w = bs_n^i s_{n-1}'' b$ is a factor of $\mathbf{x}$. Also, w is not a conjugate of $s_n^i s_{n-1}$ because $|w|_b = |s_n^i s_{n-1}|_b + 1$. Hence w is the corresponding singular factor of t_m. Since $b = \overline{a}$, we have $w = \overline{a}t_m'$. Furthermore, the word w is bordered with period q_n because

$$w = bs_n^i s_{n-1}'' b = \big(bs_n'' a\big)^i bs_{n-1}'' b,$$

and $bs_{n-1}'' b$ is a prefix of $bs_n'' a$. □

Lemma 6. *Let $n \geq 0$ and $i \geq 1$. Denote $w_j = \sigma^j(s_n^i s_{n-1})$. Then w_j has a period*

$$\begin{cases} q_n & \text{if } 0 \leq j \leq q_n - 2; \\ (i-1)q_n + q_{n-1} & \text{if } q_n \leq j \leq iq_n + q_{n-1} - 2. \end{cases}$$

Furthermore, w_j is unbordered if and only if $j = q_n - 1$ or $j = iq_n + q_{n-1} - 1$, and then w_j is a Lyndon word.

Proof. The claim is readily verified for $n = 0$, so we may assume that $n \geq 1$.

First, suppose that $0 \leq j \leq q_n - 2$. Then w_j is a factor of the word $z = s_n^i s_{n-1} s_n''$. If $n = 1$, then z clearly has a period q_n. If $n \geq 2$, then Equation (3) implies $z = s_n^{i+1} s_{n-1}''$, and we see that z has a period q_n. Therefore also w_j has a period q_n.

Next, suppose that $kq_n \leq j < (k+1)q_n$, where $1 \leq k \leq i-1$. This implies that $i \geq 2$. Then w_j is a factor of the word

$$z = s_n^{i-k} s_{n-1} s_n^k s_n'.$$

We claim that z is a prefix of the word $\big(s_n^{i-k} s_{n-1} s_n^{k-1}\big)^3$. Indeed, if $n = 1$, verifying this is a straightforward computation. And if $n \geq 2$, the claim follows by an application of Equation (3). Hence z, and consequently also w_j, has a period $(i-1)q_n + q_{n-1}$.

Finally, suppose that $iq_n \leq j \leq iq_n + q_{n-1} - 2$. This implies $n \geq 2$. Then the word w_j is a factor of $z = s_{n-1} s_n^i s_{n-1}''$. By Equation (3), we can write $z = s_{n-1} s_n^{i-1} s_{n-1} s_n''$, and hence z, and also w_j, have a period $(i-1)q_n + q_{n-1}$.

Since $s_n^i s_{n-1}$ is a primitive word over a two-letter alphabet, it has at least two conjugates that are Lyndon words, and therefore unbordered. We have seen that w_j is bordered in all other cases except possibly when $j = q_n - 1$ and $j = iq_n + q_{n-1} - 1$, so that the last claim of the lemma holds. □

Lemma 7. *A word w is an unbordered factor of $\mathbf{x}$ if and only if $w = t_{-1}$, $w = t_0$, or w is one of the two Lyndon words that are conjugates of t_m for some $m \geq 1$.*

Proof. According to Lemmas 5 and 6, the claim holds if $|w| = |t_m|$ for some $m \geq -1$. Hence we may suppose that $|w| \neq |t_m|$ for all $m \geq -1$. We will show that w is bordered.

First, observe that we have $|w| > |t_{d_1}| = d_1 + 1$ because $|t_i| = i + 1$ for $i = 0, 1, \ldots, d_1$. Furthermore, there exists an integer $n \geq 1$ such that either

$$q_n < |w| < q_n + q_{n-1} \qquad \text{or} \qquad iq_n + q_{n-1} < |w| < (i+1)q_n + q_{n-1}$$

for some $1 \leq i < d_{n+1}$. It follows that w is a proper prefix of some factor of $\mathbf{x}$ of length $iq_n + q_{n-1}$ with $1 \leq i \leq d_{n+1}$ such that

$$|w| > \max\{\, q_n,\, (i-1)q_n + q_{n-1} \,\}. \tag{7}$$

Denote this factor by z. Then z is either a conjugate of $s_n^i s_{n-1}$, or the singular factor corresponding to $s_n^i s_{n-1}$. If z is the singular factor, then w is bordered because z has a period q_n and $|w| > q_n$. Hence we may suppose that z is a conjugate of $t_m = s_n^i s_{n-1}$.

If z is bordered, then according to Lemma 6, z has a period q_n or a period $(i-1)q_n + q_{n-1}$. In either case, z has a period strictly less than $|w|$, and so w is bordered.

If z is unbordered, Lemma 6 implies that either

$$z = \sigma^{-1}(s_n^i s_{n-1}) \qquad \text{or} \qquad z = \sigma^{q_n - 1}(s_n^i s_{n-1}). \tag{8}$$

Now we have two possibilities regarding as to whether $n = 1$ or $n \geq 2$.

Suppose first that $n = 1$. Then either $z = 0(0^{d_1}1)^i$ or $z = (10^{d_1})^i 0$. In the first case, the inequality in (7) implies that $w = 0(0^{d_1})^{i-1}0^j$ for some $j \geq 1$, so that w is bordered. Similarly, in the second case we have $w = (10^{d_1})^{i-1}10^j$, where $1 \leq j \leq d_1$. If $i = 1$, the word w is a conjugate of t_j, a contradiction. Therefore, $i \geq 2$, and w is bordered.

Suppose then that $n \geq 2$. Now, the word w is a factor of either

$$\sigma^{-2}(s_n^i s_{n-1}) \qquad \text{or} \qquad \sigma^{q_n - 2}(s_n^i s_{n-1}). \tag{9}$$

Since we have already proved that w is bordered if w is a factor of a bordered word of length $iq_n + q_{n-1}$, we only have to show that both words in (9) are bordered. To do that, we only have to show that they are distinct from the words in (8). There are four cases to consider; one of them is

$$\sigma^{-2}(s_n^i s_{n-1}) = \sigma^{q_n - 1}(s_n^i s_{n-1}). \tag{10}$$

Since $s_n^i s_{n-1}$ is primitive, we get $iq_n + q_{n-1} - 2 = q_n - 1$, which implies that $n \leq 1$, a contradiction. The remaining three cases are proved similarly; we omit details here. □

The next result by de Luca and De Luca appears in the proof of [7], Theorem 10. The original proof was obtained with a clever use of Duval extensions and a result of Mignosi and Zamboni [12]. Here we give a different, more constructive, albeit longer, proof.

Lemma 8 (de Luca, De Luca)**.** *The least period of a factor w of $\mathbf{x}$ equals the length of a longest unbordered factor of w.*

Proof. Let u denote a longest unbordered factor of w. The claim clearly holds if u is a letter, so we may assume that $|u| \geq 2$. Clearly, $p(w) \geq |u|$. To show that w has a period $|u|$, it suffices to show that all factors of length $|u|$ of w are conjugates of u.

To do that, suppose, contrary to what we want to show, that w has a factor z of length $|u|$ that is not a conjugate of u. Since the reversal of w is also a factor of $\mathbf{x}$, we may, possibly by replacing w by w^R, assume that u occurs on the left of z in w. Let v denote a prefix of w such that z is a suffix of v and u is a factor of v.

Since u is unbordered and $|u| \geq 2$, Lemma 7 implies that u is a conjugate of $t_m = s_n^i s_{n-1}$ for some $n \geq 0$ and $1 \leq i \leq d_{n+1}$. Therefore, z is the singular factor corresponding to $s_n^i s_{n-1}$. Hence, if a denotes the last letter of s_{n-1}, then it follows from Lemma 5 that $z = \overline{a} s_n^i s'_{n-1}$.

Next, denote $p = s_n s_{n+1} = s_n^{d_{n+1}+1} s_{n-1}$. Observe that, as a suffix of $s_{n+2}s_{n+1}$, the word ap is a factor of $\mathbf{x}$.

Let us denote the longest common suffix of ap and va by y. Since za is a suffix of both p and va, it is a suffix of y, as well. By Lemma 6, the word

$$ap' = \sigma^{-1}(s_n^{d_{n+1}+1} s_{n-1})$$

is unbordered. Since $|ap'| = |s_n^{d_{n+1}+1} s_{n-1}| > |u|$, it then follows that y is a proper suffix of ap because otherwise ap' is a factor of w, contradicting the maximality of $|u|$.

Since p, and hence also y, has a period q_n, the word u cannot be a factor of y because u is unbordered and $|u| = iq_n + q_{n-1}$. Consequently, y is also a proper suffix of va. This implies that y is a left special factor of $\mathbf{x}$, and as such, a prefix of $\mathbf{c}$. In particular, s_n is a prefix of y. Now the primitivity of s_n and the fact that y is a suffix of $p = s_n^{d_{n+1}+1} s_{n-1}$ imply that we have $y = s_n^j s_{n-1}$ for some $i + 1 \leq j \leq d_{n+1} + 1$ (for the left inequality, note that $|y| > |u| = |s_n^i s_{n-1}|$.

We can rule out the possibility that $j = d_{n+1} + 1$ because the word $s_n^{d_{n+1}+1} s_{n-1}$ is not a prefix of $\mathbf{c}$. Indeed, this is straightforward to verify for $n = 0$ and $n = 1$, and Theorem 1 handles the case when $n \geq 2$.

Now, we see that $\overline{a}y$ is a suffix of p. Since y is a proper suffix of va, the maximality of y implies that ay is a factor of va. Therefore ay' is a factor of v, and hence of w. But $ay' = as_n^j s'_{n-1}$ is unbordered by Lemma 6, and $|ay'| \geq |za| > |u|$, contradicting the maximality of u. The proof is complete. □

The next result is the strongest result in this section, and it gives our desired formula for the period set of $\mathbf{x}$ as a corollary.

Theorem 3. *The fractional root of a factor of* $\mathbf{x}$ *is a conjugate of* t_m *for some* $m \geq -1$.

Proof. Let w be a factor of $\mathbf{x}$. If w is unbordered, then according to Lemma 7, it is a conjugate of some t_m, where $m \geq -1$. If w is bordered, Lemmas 8 and 7 imply that $p(w) = |t_m|$ for some $m \geq 0$. Consequently, the fractional root of w is either a conjugate of t_m, or the singular factor $\overline{a}t'_m$, where a is the last letter of t_m. In the first case the claim holds, so may suppose that $\overline{a}t'_m$ is the fractional root of w.

Since $p(w) < |w|$, it follows that $t'_m\overline{a}$ is a factor of $\mathbf{x}$. By the definition of a singular factor, no other conjugates of $\overline{a}t'_m$ except $\overline{a}t'_m$ itself are factors of $\mathbf{x}$. Therefore, $\overline{a}t'_m = t'_m\overline{a}$. This implies that the fractional root of w is actually the letter $\overline{a}$, and the claim follows. □

Theorem 3 implies the following characterization of the period set of $\mathbf{x}$.

Corollary 3. *The period set of* $\mathbf{x}$ *is the set*

$$\{\, |t_m| : m \geq -1 \,\} = \{1\} \cup \{\, iq_n + q_{n-1} : n \geq 0, i = 1, \ldots, d_{n+1}\}.$$

The famous Fibonacci word is the characteristic sequence with slope $1/\phi$, where $\phi = (1 + \sqrt{5})/2$ denotes the golden ratio. As a special case of Corollary 3, we obtain the next result, which was first proved in [14].

Corollary 4. *The least period of a factor of the Fibonacci word is a Fibonacci number.*

5. Applications

In this section we give four applications of our results in the previous section. The first application by Harju and Nowotka [9] is a direct corollary of Lemma 7.

Corollary 5. *Unbordered words that are factors of Sturmian words are Lyndon words.*

The next characterization of finite Sturmian words is by de Luca and De Luca [7], Theorem 10.

Corollary 6. *A finite word is a factor of a Sturmian word if and only if its fractional root is a conjugate of a standard word.*

Proof. Let a finite word w be a factor of a Sturmian word, say a factor of $\mathbf{x}$ using the notation from the last section. Then by Theorem 3 the fractional root of w is a conjugate of t_m for some $m \geq -1$, and t_m is a standard word.

Conversely, suppose $w = u^\tau$, where u is a conjugate of a standard word, say s_n, and $\tau \geq 1$ is rational. Then w is a factor of s_n^{a+2}, where $a = \lfloor \tau \rfloor$, which clearly is a prefix of a characteristic word. □

Our last two corollaries below use a well-known theorem by Fine and Wilf [8], which states that if two words x^n and y^m have a common prefix of length $|x| + |y| - \gcd(|x|, |y|)$, then both of them have a period $\gcd(|x|, |y|)$.

Here is one more application of Theorem 3, see also Damanik and Lenz [6].

Corollary 7. *If a square uu is a factor of $\mathbf{x}$ and u is primitive, then u is a conjugate of t_m for some $m \geq 0$.*

Proof. Let v denote the fractional root of uu, which by Theorem 3 is a conjugate of t_m for some $m \geq -1$. The word 11 does not occur in $\mathbf{x}$, so that $m \geq 0$. Then $uu = v^\tau$ for some rational $\tau \geq 2$, and we have $|u| + |v| \leq |uu|$. By the theorem of Fine and Wilf, uu has a period $\gcd(|u|, |v|)$. Since v is the fractional root of uu, this implies that $|v| = \gcd(|u|, |v|)$, and hence $|v|$ divides $|u|$. Since u is primitive, it follows that $u = v$. □

Cummings *et al.* [5] gave two proofs showing that, for $n \geq 2$, the least period of the finite Fibonacci word f_n is ${f_{n-1}}^2$. As our last result of this chapter, we generalize the result of Cummings *et al.* to standard words. Let us use the notation from the previous section, that is, s_n is a standard word and $s_n = s_{n-1}^{d_n} s_{n-2}$.

Corollary 8. *If $n \geq 2$, then the least period of s_n equals q_{n-1}.*

Proof. Since $s_n = s_{n-1}^{d_n} s_{n-2}$ and s_{n-2} is a prefix of s_{n-1}, we see that q_{n-1} is a period of s_n. Hence we only need to show that q_{n-1} is the least period. To do that, suppose the contrary: we have $p(s_n) < q_{n-1}$.

First, suppose that $d_n \geq 2$. Since s_n has periods q_{n-1} and $p(s_n)$, and

$$q_{n-1} + p(s_n) < 2q_{n-1} < q_n,$$

it follows from the theorem of Fine and Wilf that $p(s_n)$ is a proper divisor of q_{n-1}. Since s_{n-1} is a prefix of s_n, this implies that s_{n-1} is not primitive, a contradiction.

Second, suppose that $d_n = 1$. If $p(s_n) \leq q_{n-2}$, we derive a contradiction as above. Therefore we may assume that $p(s_n) > q_{n-2}$. Now, Theorem 3 implies that $p(s_n) = iq_{n-2} + q_{n-3}$ with $1 \leq i < d_{n-1}$. Then the word $s_{n-2}^i s_{n-3} s_{n-2}$ is a prefix of s_n. But since $s_{n-1} = s_{n-2}^{d_{n-1}} s_{n-3}$ is also a prefix of s_n, we obtain $s_{n-3}s_{n-2} = s_{n-2}s_{n-3}$, which is absurd by Equation (2). This contradiction completes the proof. □

Acknowledgements. The second author is grateful to Alessandro De Luca for pointing out the result stated in Lemma 8. A sincere thank you also to Gwénaël Richomme for pointing out a deficiency in the first version of Lemma 2 and also for other useful comments that helped to clarify the presentation. Finally, thank you to the referees for their remarks.

[2]To be precise, Cummings *et al.* showed that the longest border of f_n is f_{n-2}, but these two claims are equivalent.

References

[1] J.-P. Allouche and J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, in *Sequences and Their Applications: Proceedings of SETA'98. Springer Series in Discrete Mathematics and Theoretical Computer Science*, C. Ding, T. Helleseth and H. Niederreiter, Eds., Springer-Verlag, London (1999) 1–16.

[2] J. Berstel, *On the index of Sturmian words.* In *Jewels are forever*. Springer, Berlin (1999) 287–294.

[3] W.-T. Cao and Z.-Y. Wen, Some properties of the factors of Sturmian sequences. *Theor. Comput. Sci.* **304** (2003) 365–385.

[4] C. Choffrut and J. Karhumäki, Combinatorics on words. In A. Salomaa and G. Rozenberg, Eds., *Handbook of Formal Languages*, volume 1. Springer, Berlin (1997) 329–438.

[5] L.J. Cummings, D.W. Moore and J. Karhumäki, Borders of Fibonacci strings. *J. Comb. Math. Comb. Comput.* **20** (1996) 81–87.

[6] D. Damanik and D. Lenz, Powers in Sturmian sequences. *Eur. J. Combin.* **24** (2003) 377–390.

[7] A. de Luca and A. De Luca, Some characterizations of finite Sturmian words. *Theor. Comput. Sci.* **356** (2006) 118–125.

[8] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** (1965) 109–114.

[9] T. Harju and D. Nowotka, Minimal Duval extensions. *Int. J. Found. Comput. Sci.* **15** (2004) 349–354.

[10] M. Lothaire, *Combinatorics on Words*. Cambridge University Press, Cambridge (1997).

[11] M. Lothaire, Algebraic Combinatorics on Words, *Encyclopedia of Mathematics and its Applications*, Vol. **90**. Cambridge University Press, Cambridge (2002).

[12] F. Mignosi and L.Q. Zamboni, A note on a conjecture of Duval and Sturmian words. *RAIRO-Theor. Inf. Appl.* **36** (2002) 1–3.

[13] M. Mohammad-Noori and J.D. Currie, Dejean's conjecture and Sturmian words. *Eur. J. Combin.* **28** (2007) 876–890.

[14] K. Saari, *Periods of factors of the Fibonacci word.* in *Proceedings of the Sixth International Conference on Words (WORDS'07).* Institut de Mathématiques de Luminy (2007) 273–279.

Communicated by J. Karhumäki.
Received November 28, 2007. Accepted February 6, 2008.