Discrete Mathematics and Theoretical Computer Science **1**, 1997, 43–52

Lyndon factorization of the Thue–Morse word and its relatives

Augustin Ido and Guy Melançon

LaBRI, URA 1304 CNRS – Université Bordeaux I, 351 Cours de la Libération, 33405 Talence Cedex, France
E-Mail: `melancon@labri.u-bordeaux.fr`

We compute the Lyndon factorization of the Thue–Morse word. We also compute the Lyndon factorization of two related sequences involving morphisms that give rise to new presentations of these sequences.

Keywords: Lyndon factorization, Thue–Morse word, morphisms

1 Introduction

Some attention has recently been given to the Lyndon factorization of infinite words [16], [10], [12]. These works are themselves related to the earlier works by Reutenauer [13] and Varricchio [17], concerned with unavoidable regularities and semigroup theory.

The results we present here reinforce those in [10] and [12], and give an additional application of the general Lyndon factorization theorem for infinite words ([16, Theorem 2.4]; see [11] for a generalization). In [10] we explicitely compute the Lyndon words appearing in the factorization of Sturmian words and identify them as Christoffel primitive words (a result obtained differently by Berstel and de Luca [3]). In this paper, we concentrate on the Thue–Morse word and give the computation of its Lyndon factorization (Theorem 3.1) and describe some of its properties (Corollary 3.2, Remark 3.3 and Corollary 3.4). Incidentally, we are able to compute the factorization for the 'dual' Thue–Morse word in which appears an infinite Lyndon word (cf Theorem 3.7). We also look at relatives (Equations (4) and (6)) of the Thue–Morse word from the same point of view; these were first studied in [7] and [4], and later in [1]. The factorizations given here for these infinite words (cf Theorems 4.6 and 4.7) use morphisms having special properties with respect to Lyndon words. Moreover, we give identities involving these morphisms for these infinite words.

2 Basic Results and Notations

The notations used are those usual in theoretical computer science (cf [8]). Throughout the paper, we use the alphabet $A = \{a, b\}$, totally ordered by $a < b$, and we denote by A^* the set of all words with the lexicographical order.

2.1 Lyndon words

Let L denote the set of Lyndon words over A: they are words strictly smaller than any of their proper non empty right factors. For instance, letters are Lyndon words, and ab, aab, abb are Lyndon words if $a, b \in A$ satisfy $a < b$. More generally, given two Lyndon words $u, v \in L$, we have $uv \in L$ iff $u < v$. The central result about Lyndon words is Lyndon's factorization theorem:

Theorem 2.1 ([5]) *Any non-empty word $w \in A^*$ is a unique non-increasing product of Lyndon words: $w = \ell_1 \cdots \ell_n$, with $\ell_i \in L$ ($i = 1, \ldots, n$) and $\ell_1 \geq \cdots \geq \ell_n$.*

For a proof, see [8]. The expression of a Lyndon word as an increasing product of two Lyndon words may not be unique. For example, we have $aababb = (a)(ababb) = (aab)(abb) = (aabab)(b)$. Given $w \in L$, define w'' to be the *longest* right factor of w qualifying as a Lyndon word. Denote by w' the unique left factor of w such that $w = w'w''$. Then we have $w', w'' \in L$ and $w' < w < w''$. Thus, we have e.g. $(aababb)' = a, (aababb)'' = ababb$.

Proposition 2.2 (cf [8, Prop. 5.1.4]) *Let $u = u'u'' \in L$ and $v \in L$ be Lyndon words such that $u < v$. Then the factorization uv is standard (i.e. $(uv)' = u, (uv)'' = v$) iff $u'' \geq v$.*

2.2 Infinite Lyndon Words

Siromoney *et al.* [16] have extended Lyndon's theorem to (right) infinite words. They define an infinite word $s = a_0 a_1 \cdots$ to be an *infinite Lyndon word* if an infinite number of its left factors qualify as Lyndon words. For instance, the infinite word $abbb \cdots = \lim_n ab^n$ is an infinite Lyndon word; more generally, given $u, v \in L$ with $u < v$ the infinite word $\lim_n uv^n$ is an infinite Lyndon word. The central result in [16] is:

Theorem 2.3 ([16, Theorem 2.4]) *Any infinite word s factorizes uniquely into one of the following forms:*

- *either there exists an infinite non-increasing sequence of finite Lyndon words $(\ell_k)_{k\geq 0}$ such that:*

$$s = \ell_0 \ell_1 \cdots \tag{1}$$

- *or there exist finite Lyndon words $\ell_0, \ldots, \ell_{m-1}$ ($m \geq 0$) and an infinite Lyndon word t such that:*

$$s = \ell_0 \cdots \ell_{m-1} t, \quad \text{with } \ell_0 \geq \cdots \geq \ell_{m-1} > t \tag{2}$$

Remark 2.4 In [17], the author implicitely shows the existence of the Lyndon factorization of type (1) for certain infinite words. This work, as well as [13], is related to the study of unavoidable regularities in infinite words. This is echoed by results in [10, Sect. 4] and [11].

2.3 Morphisms and Lyndon Words

This last subsection contains a proposition we shall need in the sequel. It formulates a condition for a morphisms to preserve Lyndon words and lexicographical order.

Proposition 2.5 *Let $A = \{a < b\}$ and Z be finite alphabets. Suppose $\theta : A^* \to Z^*$ is a morphism given by $\theta(a) = a^m b^p$ and $\theta(b) = a^n b^q$ with $a^m b^p < a^n b^q$.*

Then θ is strictly increasing over A^*. Moreover, θ sends Lyndon words to Lyndon words and preserve their standard factorizations. That is, given $w \in L(A)$, we have $\theta(w) \in L(B)$ and $\theta(w)' = \theta(w')$, $\theta(w)'' = \theta(w'')$.

As a consequence, the sequence $(\theta^n(b))_{n\geq 0}$ forms a strictly decreasing sequence of Lyndon words.

Remark 2.6 The last statement of Proposition 2.5 has a geometrical interpretation. To each Lyndon word is associated a (planar rooted) binary tree having its leaves labelled by letters. Indeed, the tree associated with $w \in L$ is either a single vertex labelled by a if $w = a \in A$, or is formed of a left tree associated with w' and a right tree associated with w''. Hence, for a morphism θ to preserve standard factorization means that the tree structure of $\theta(w)$ is obtained from that of w by attaching to a leave labelled by a, the tree associated with $\theta(a)$ (see Fig. 1).

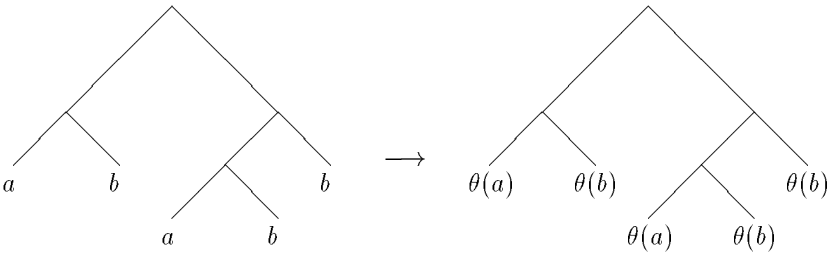

Fig. 1: Computing the image of $ababb$ under θ (preserving standard factorization).

Proof of Proposition 2.5. That θ is strictly increasing over A^* is easy. An induction then allows to show $\theta(L) \subset L$ since any Lyndon word is an increasing product of two Lyndon words of smaller length. The last part of the proposition is proved using Prop. 2.2. The last part of the statement is clear. □

3 Factorizing Thue–Morse's Word

In this section, we give the computation of the Lyndon factorization (1) for the Thue–Morse word. Let $A = \{a, b\}$ and set $u_0 = a$ and $v_0 = b$. Define for all $n \geq 1$, $u_n = u_{n-1}v_{n-1}$ and $v_n = v_{n-1}u_{n-1}$. Hence, $u_1 = ab$, $v_1 = ba$, $u_2 = abba$, $v_2 = baab$, and so on. The sequence $(u_n)_{n\geq 0}$ converges to a unique infinite word μ, called the Thue–Morse word (over $\{a, b\}$). This infinite word possess numerous interesting properties (cf [8, Chap. 2]), and has been studied by a large number of authors; the interested reader is refered to (the bibliography of) the survey by Berstel [2]. The words u_n may alternatively be obtained using a morphism we denote $\alpha : A^* \to A^*$, defined by $\alpha(a) = ab$ and $\alpha(b) = ba$. One then finds $u_n = \alpha(u_{n-1})$, for all $n \geq 1$. Iterating α to infinity leads to $\mu = \lim_{n\to\infty} \alpha^n(a)$; this is equivalent to the fact that μ is a fixed-point of α.

Recall that if $u \in A^*$ and $a \in A$ then the expression ua^{-1} consists in deleting the last a in u (if possible). Our main result concerning the Thue–Morse word is:

Theorem 3.1 *Let $w_1 = abb$, $w_2 = ab$, and for all $n \geq 2$, $w_{n+1} = a\alpha(w_n)a^{-1}$. The words $(w_n)_{n\geq 1}$ form a strictly decreasing sequence of Lyndon words, and we have:*

$$\mu = \prod_{n\geq 1} w_n \tag{3}$$

The following corollary is a straightforward consequence of a general result concerning the Lyndon factorization of infinite words. See [10, Proposition 15].

Corollary 3.2 *Equation 3 shows that the Thue–Morse word μ is ω-divided.*

Remark 3.3 For all $n \geq 2$, the word w_n is a conjugate of u_{n-1} (and of v_{n-1}). This is straightforward from the definition for w_n in terms of u_{n-1}, since $\alpha(u_{n-1}) = u_n$.

As a consequence of Theorem 3.1, we obtain a second recursive construction for the words w_n that does not use the morphism α. This result was announced in [10] (without proof). We prove it here for sake of completeness.

Corollary 3.4 *For all $n \geq 2$, we have $w_n = (w_{n-1}b^{-1})w_1 \cdots w_{n-2}$.*

We must first observe that it makes sense to compute $w_n b^{-1}$ since every word w_n ends with b, as follows from their definition given in Theorem 3.1.

We proceed by induction and compute, for $n \geq 1$:

$$
\begin{aligned}
w_{n+1} &= a\alpha(w_n)a^{-1} \\
&= a\alpha((w_{n-1}b^{-1})w_1 \cdots w_{n-2})a^{-1} \\
&= a\alpha(w_{n-1})(ba)^{-1}\alpha(w_1)\alpha(w_2)\cdots\alpha(w_{n-2})a^{-1} \\
&= (a\alpha(w_{n-1})a^{-1})b^{-1}\alpha(w_1)\alpha(w_2)\cdots\alpha(w_{n-2})a^{-1} \\
&= (a\alpha(w_{n-1})a^{-1})b^{-1}(abbaba)(abba)\cdots\alpha(w_{n-2})a^{-1} \\
&= (a\alpha(w_{n-1})a^{-1})b^{-1}(abb)(ab)(aabba)\cdots\alpha(w_{n-2})a^{-1} \\
&= (a\alpha(w_{n-1})a^{-1})b^{-1}(abb)(ab)(w_3a)\alpha(w_3)a^{-1}a\cdots a\alpha(w_{n-2})a^{-1} \\
&= (a\alpha(w_{n-1})a^{-1})b^{-1}(abb)(ab)w_3(a\alpha(w_3)a^{-1})\cdots(a\alpha(w_{n-2})a^{-1}) \\
&= (w_nb^{-1})w_1w_2\cdots w_{n-1}
\end{aligned}
$$

Proof of Theorem 3.1. First observe that if w_n ends with a b, then the last letter of $\alpha(w_n)$ is equal to a. So we may compute $\alpha(w_n)a^{-1}$, showing that w_n is well defined for all $n \geq 1$. To show that μ is obtained by the infinite product expansion (3) we only have to verify that $\prod_{n\geq 1} w_n$ is kept fixed by α. We have:

$$
\begin{aligned}
\alpha(\prod_{n\geq 1} w_n) &= \alpha(abb)\prod_{n\geq 2}\alpha(w_n) \\
&= (abb)(ab)a\prod_{n\geq 2}\alpha(w_n) \\
&= w_1w_2\prod_{n\geq 2} a\alpha(w_n)a^{-1} \\
&= w_1w_2\prod_{n\geq 3} w_n = \prod_{n\geq 1} w_n
\end{aligned}
$$

Now, we need to show that $a\alpha(w_n)a^{-1}$ form a decreasing sequence of Lyndon words. Observe first that α is increasing, since $\alpha(a)$ and $\alpha(b)$ have the same length and $\alpha(a) < \alpha(b)$. Hence, assuming $w_n > w_{n+1}$

for $n \geq 1$, we find $\alpha(w_n) > \alpha(w_{n+1})$ from which follows $w_{n+1} = a\alpha(w_n)a^{-1} > a\alpha(w_{n+1})a^{-1} = w_{n+2}$. Hence the sequence $(w_n)_{n\geq 1}$ is decreasing.

Again, we use the fact that α is increasing to show by induction that w_n $(n \geq 1)$ is a Lyndon word. This holds true for w_1, w_2. By virtue of Remark 3.3, we know that w_n is a conjugate of u_{n-1}. Since Lyndon words are minimal representatives of their conjugacy classes, assume inductively that the least element of the conjugacy class of u_{n-1} is w_n. Now, observe that the elements of the conjugacy class for $u_n = \alpha(u_{n-1})$ are of the form $\alpha(v)$, $a\alpha(v)a^{-1}$, $b\alpha(v)b^{-1}$ where v is a conjugate of u_n, since $|\alpha(a)| = |\alpha(b)| = 2$. So we deduce that the least element among these is $a\alpha(v)a^{-1}$ where v is the least element of the conjugacy class for u_n. This shows that $w_{n+1} = a\alpha(w_n)a^{-1}$ is a Lyndon word. That concludes the proof of Theorem 3.1. □

Remark 3.5 The idea of using the pattern $w_{n+1} = a\alpha(w_n)a^{-1}$ for proving Theorem 3.1 was suggested to us by G. Sénizergues, who happened to read a first version of the manuscript. This idea may be exploited to obtain the factorization for the 'dual' Thue–Morse word, namely $lim_n\, v_n$ (see the next remark).

Remark 3.6 Note that we could have set $a > b$; this would amount to imposing on A^* the *inverse* lexicographical order. Note that, for all $n \geq 0$, v_n is obtained from u_n by exchanging a's and b's. Hence, the factorization of u_n using the total order $a > b$ is directly obtained from that of v_n with $a < b$. The next theorem fully answers the question just raised.

Theorem 3.7 *Let $w_1 = aab$ and and for all $n \geq 1$, $w_{n+1} = a\alpha(w_n)a^{-1}$. The words $(w_n)_{n\geq 1}$ form a stricly increasing sequence of Lyndon words such that w_n is a left factor of w_{n+1}. Thus, $\ell = \lim_n w_n$ is an infinite Lyndon word, and we have $a\ell = \alpha(\ell)$. Moreover, the factorization of the 'dual' Thue–Morse word is of type (2) and is $lim_n\, v_n = b\ell$.*

Proof. That $(w_n)_{n\geq 1}$ is an increasing sequence of Lyndon words is proved as in Theorem 3.1. That w_n is a left factor of w_{n+1} is a property gained from the morphism α. So, we may indeed define the limit $\ell = \lim_n w_n$ which is by definition an infinite Lyndon word (cf. Sect. 2.2). The identity $a\ell = \alpha(\ell)$ is equivalent to $\ell = a^{-1}\alpha(\ell)$, which comes at once from the definition for ℓ. To show that we have $\lim_n v_n = b\ell$, we verify that the latter is kept fixed by α. We have:

$$\begin{aligned}
\alpha(b\ell) &= \alpha(b \lim_{n\to\infty} w_n) \\
&= ba \lim_{n\to\infty} \alpha(w_n) \\
&= b \lim_{n\to\infty} a\alpha(w_n) \\
&= b \lim_{n\to\infty} a\alpha(w_n)a^{-1} \\
&= b \lim_{n\to\infty} w_{n+1} = b\ell
\end{aligned}$$

Remark 3.8 Another proof of Theorem 3.1 proceeds by induction and first computes the Lyndon factorization of all u_n (and all v_n). It then exploits the fact that these factorizations stabilize, i.e. they form a converging sequence of finite decreasing sequence of Lyndon words. This proof we first developed enabled us to obtain the exact number of factors occuring in the factorization for u_n (and v_n).

More precisely, it is possible to show that the words u_n factorize as a decreasing product of $p(n)$ Lyndon words, $u_n = w_1^n \cdots w_{p(n)}^n$ where $p(n) = 3k - 1$ if $n = 2k$ and $p(n) = 3k$ if $n = 2k + 1$, and that w_i^{n-1} and w_i^n coincide for $i = 1, \ldots, n - 2$. For more details, the reader is referred to [6].

Remark 3.9 There exists a generalization of the Thue–Morse word over an arbitrary finite alphabet A. We define it here over the three letters and refer the interested reader to [4]. Define three sequence of words by setting $u_0 = a$, $v_0 = b$, $w_0 = c$ and $u_{n+1} = u_n v_n w_n$, $v_{n+1} = v_n w_n u_n$, $w_{n+1} = w_n u_n v_n$. Then the word $\mu = \lim_n u_n$ is the Thue–Morse word over $\{a, b, c\}$. It may also be obtained as the limit $\lim_n \alpha^n(a)$, where α is the morphism sending $a \mapsto abc$, $b \mapsto bca$ and $c \mapsto cab$.

It is natural to look at the Lyndon factorization for this general Thue–Morse word. However, the problem of describing this factorization is still open. Indeed, our techniques did not enable us to obtain any result as for the two letter case.

4 Factorization and Properties of Thue–Morse's Relatives

In this section, we give a complete description of the Lyndon factorization of two infinite words d and χ obtained from infinite bi-valued sequences $(d_n)_{n\geq 0}$ and $(\chi_n)_{n\geq 0}$ related to the Thue–Morse word. These were first studied in [7] and [4], and later in [1].

Definition 4.1 ([7]) *Let $c = (c_n)_{n\geq 0}$, $c_n \in \mathbb{N}$, be defined inductively by $c_0 = 1$ and:*

$$c_{n+1} = \begin{cases} c_n + 1 \text{ if } c_n + 1/2 \notin c \\ c_n + 2 \text{ otherwise} \end{cases}$$

Thus, c = 1, 3, 4, 5, 7, 9, 11, 12, 13, ... Equivalently, c is the lexicographically least sequence of positive integers satisfying $n \in c$ implies $2n \notin c$ (cf [1]). Note that the difference between two consecutive terms in the sequence is $c_{n+1} - c_n = 1$ or 2. Hence, we may define:

Definition 4.2 *Let $d = d_0 d_1 \cdots$ denote the infinite word defined by*

$$d_n = c_{n+1} - c_n \tag{4}$$

Hence, we have $d = 21122211211211222\cdots$. The link between this sequence and μ is given by the following result:

Theorem 4.3 ([1, Theorem 4]) *The Thue–Morse word has a coding:*

$$\mu = a^{d_0} b^{d_1} a^{d_2} b^{d_3} \cdots \tag{5}$$

The sequence $(d_n)_{n\geq 0}$ and the coding given in Equation 5 appeared for the first time in [4]. In [1], it is proved that $d_n = c_{n+1} - c_n$. The sequence c may also be studied by means of its characteristic function (or sequence) we now define.

Definition 4.4 *Let $(\chi_n)_{n\geq 1}$ denote the* characteristic *sequence of c (over $\mathbb{N}^*$). That is, we define*

$$\chi_n = \begin{cases} 0 \text{ if } n \notin c \\ 1 \text{ if } n \in c \end{cases} \tag{6}$$

for all $n \geq 1$. We then define the infinite word χ by setting $\chi = \chi_0 \chi_1 \chi_2 \cdots$

Hence, we have $\chi = 1011101010111\cdots$. We will make use of a result borrowed from [1].

Lemma 4.5 ([1, Lemma 2]) *The infinite word χ is completely determined by the following conditions:*

$$\begin{aligned} \chi_{2n+1} &= 1 \\ \chi_{4n+2} &= 0 \\ \chi_{4n} &= \chi_n \end{aligned} \tag{7}$$

We define two morphisms $\beta : \{1,2\}^* \to \{0,1\}^*$ and $\gamma : \{1,2\}^* \to \{1,2\}^*$ by setting $\beta(1) = 01$, $\beta(2) = 0111$, and $\gamma(1) = 112$ and $\gamma(2) = 11222$. Note that, by virtue of Proposition 2.5, both β and γ preserve the lexicographical order on A^* and send Lyndon words to Lyndon words. As a consequence, we are able to show that χ is, except for its first letter, a morphic image of d. As we will see, that is in fact a consequence of [1, Lemma 2] (Lemma 4.5 above). We have the following theorems:

Theorem 4.6 *Consider the sequence of words $(s_n)_{n\geq 0}$ with $s_0 = 2$, $s_{n+1} = \gamma(s_n)$ ($n \geq 0$). The words $(s_n)_{n\geq 0}$ form a strictly decreasing sequence of Lyndon words and we have:*

$$d = \prod_{n\geq 0} s_n \tag{8}$$

Moreover, this infinite product expansion for d implies $d = 2\gamma(d)$.

Theorem 4.7 *Consider the sequence of words $(t_n)_{n\geq 0}$ with $t_0 = 1$ and $t_{n+1} = \beta(s_n)$ ($n \geq 0$). The words $(t_n)_{n\geq 0}$ form a strictly decreasing sequence of Lyndon words and we have:*

$$\chi = \prod_{n\geq 0} t_n \tag{9}$$

Moreover, this infinite product expansion for χ implies $\chi = 1\beta(d)$.

Remark 4.8 Theorems 4.7 and 4.6 should be looked at from a point of view developed in [15], where the author answers a question asking for conditions for the characteristic word of a sequence to be the image of a fixed point of a morphism.

Define the sequence of integers $m = (m_i)_{i\geq 0}$ with $m_0 = 1$ and $m_{n+1} = 4m_n + 1$; hence we have $m = 1, 5, 21, 85, \ldots$ Let $(w_n)_{n\geq 0}$ be the unique consecutive factors of d, starting with $w_0 = d_0 = 2$ defined by $w_{n+1} = d_{m_0+\cdots+m_n+1} \cdots d_{m_0+\cdots+m_n+m_{n+1}}$, satisfying $|w_n| = m_n$. Hence, we have $w_0 = 2$, $w_1 = 11222$, $w_2 = 112112112221122211222, \ldots$

Proposition 4.9 *We have, for any $n \geq 0$, $w_{n+1} = \gamma(w_n)$.*

As we will see, this proposition is a consequence of Equation (7). First, observe that by definition of γ, we have for any $w \in A^*$,

$$\begin{aligned} |\gamma(w)|_1 &= 2(|w|_1 + |w|_2) \\ |\gamma(w)|_2 &= |w|_1 + 3|w|_2 \end{aligned} \tag{10}$$

Then observe that, since $w_0 = 2$, we may show by induction that $|\gamma^n(w_0)|_2 = |\gamma^n(w_0)|_1 + 1$ and $|\gamma^n(w_0)| = m_n$. Recall from Equation (4) that for any n, the letter d_n is determined by the difference $c_{n+1} - c_n$. Moreover, we have $c_{n+1} = (\sum_{i=0}^n d_i) + 1 = (\sum_{i=0}^n c_{i+1} - c_i) + 1$. Hence, it is natural to think of the letter d_n as corresponding to the integer c_{n+1}. Any integer $m \in c$ is of the form $m = c_k$ for a given $k \geq 0$. We denote this unique integer by $c^{-1}(m)$. So for instance, $c^{-1}(3) = 1, c^{-1}(4) = 2$ and $c^{-1}(5) = 3$; hence $d_{c^{-1}(3)-1} = d_0 = 2$, and $d_{c^{-1}(4)-1} = d_1 = d_{c^{-1}(5)-1} = d_2 = 1$.

Lemma 4.10 *Let $n \geq 0$. Suppose first that $d_n = c_{n+1} - c_n = 1$. Then $4c_n \in c$, but $4c_n + 2 \notin c$. Consequently, $d_{c^{-1}(4c_n)-1} = d_{c^{-1}(4c_n+1)-1} = 1$ and $d_{c^{-1}(4c_n+3)-1} = 2$.*

Suppose now that $d_n = c_{n+1} - c_n = 2$. Then $4c_n \in c$, but $4c_n+2, 4c_n+4, 4c_n+6 \notin c$. Consequently, $d_{c^{-1}(4c_n)-1} = d_{c^{-1}(4c_n+1)-1} = 1$ and $d_{c^{-1}(4c_n+3)-1} = d_{c^{-1}(4c_n+5)} = d_{c^{-1}(4c_n+7)} = 2$.

Suppose $d_n = 1$. That $4c_n \in c$ follows from the fact that $2c_n \notin c$, since $c_n \in c$. That $4c_n + 2 \notin c$ is given by Equation (7). That $4c_n - 1, 4c_n + 1, 4n + 3 \in c$ follows from the fact that c contains every odd integer. Hence the integers $4c_n - 1, 4c_n, 4c_n + 1, 4c_n + 3$ are consecutive terms in c. Hence, we have $d_{c^{-1}(4c_n)-1} = d_{c^{-1}(4c_n+1)-1} = 1$ and $d_{c^{-1}(4c_n+3)-1} = 2$.

Suppose now that $d_n = 2$. Again, we have $4c_n \in c$. That $4c_n + 2, 4c_n + 6 \notin c$ follows from Equation (7). Observe that, $c_{n+1} - c_n = 2$ implies that $c_n + 1 \notin c$, hence $2c_n + 2 \in c$ so $4c_n + 4 \notin c$. That $4c_n + k \in c$ for $k = -1, 1, 3, 5, 7$ follows from the fact that they all are odd. Hence, the integer $4c_n - 1, 4c_n, 4c_n + 1, 4c_n + 3, 4c_n + 5, 4c_n + 7$ are consecutive terms in c. Hence, we have $d_{c^{-1}(4c_n)-1} = d_{c^{-1}(4c_n+1)-1} = 1$ and $d_{c^{-1}(4c_n+3)-1} = d_{c^{-1}(4c_n+5)-1} = d_{c^{-1}(4c_n+7)-1} = 2$.

Proof of Proposition 4.9. We first associate to any integer c_n a subsequence $\mathcal{S}(c_n)$ of c by setting:

$$c_n \mapsto \mathcal{S}(c_n) = \begin{cases} \{4c_n - 1, 4c_n, 4c_n + 1, 4c_n + 3\} \text{ if } c_{n+1} - c_n = 1 \\ \{4c_n - 1, 4c_n, 4c_n + 1, 4c_n + 3, 4c_n + 5, 4c_n + 7\} \text{ otherwise} \end{cases}$$

Observe that $\mathcal{S}(c_n)$ and $\mathcal{S}(c_{n+1})$ only have a single element in common, namely the greatest element of $\mathcal{S}(c_n)$ which is also the least element of $\mathcal{S}(c_{n+1})$. This element is equal to $4c_n+3$ if $c_{n+1}-c_n = 1$, and to $4c_n + 7$ otherwise, as is easily checked. This shows that $c_0, \mathcal{S}(c_0), \mathcal{S}(c_1), \ldots$ coincides with c. Moreover, associating to c_{n+1} ($n \geq 0$) the letter d_n, we see that the mapping $\mathcal{S}$ is nothing else but the morphism γ. Indeed, suppose $d_n = c_{n+1} - c_n = 1$ then we have $\mathcal{S}(c_n) = \{4c_n - 1, 4c_n, 4c_n + 1, 4c_n + 3\}$; the three letter word associated with this subsequence, which is a factor of d, is 112. The case $d_n = 2$ is similar.

We now define for all $n \geq 0$ a subsequence I_n of c. Put $I_0 = \{c_0\}$, and $I_{n+1} = \mathcal{S}(I_n)$. We have $c = I_0, I_1, \ldots$; this follows from $c = c_0, \mathcal{S}(c_0), \mathcal{S}(c_1), \ldots$ Now, we claim that the number of elements in I_n ($n \geq 0$) is equal to $m_n + 1$. Indeed, this follows from the observation that $\mathcal{S}$ coincides with γ when going from c to d. Hence, a simple induction counting the number of consecutive terms c_k, c_{k+1} of I_n according to the value $c_{k+1} - c_k = 1$ or $c_{k+1} - c_k = 2$ leads to a result identical with Equations (10). This implies that the factor of d associated with I_n is equal to w_n, since its length is $|I_n| - 1 = m_n$. This, together with the previous observation that $\mathcal{S}$ coincides with γ, concludes the proof of Proposition 4.9. □

Proof of Theorem 4.6. The first part of the statement follows directly from Proposition 2.5 applied to γ and from Lemma 4.9. The last part of the statement is clear. □

Proof of Theorem 4.7. The first part of the statement is also proved using Proposition 2.5. Next, we use a technique similar to the one developed for the proof of Proposition 4.9.

We first associate to any integer c_n a subsequence $\mathcal{T}(c_n)$ of *consecutive* integers by setting:

$$c_n \mapsto \mathcal{T}(c_n) = \begin{cases} \{2c_n, 2c_n + 1\} \text{ if } c_{n+1} - c_n = 1 \\ \{2c_n, 2c_n + 1, 2c_n + 2, 2c_n + 3\} \text{ if } c_{n+1} - c_n = 2 \end{cases}$$

Observe that $\mathcal{T}(c_n)$ and $\mathcal{T}(c_{n+1})$ are disjoint and that the greatest element of $\mathcal{T}(c_n)$ is one less than the least element of $\mathcal{T}(c_{n+1})$. This shows that every integer except 1 appears in $\mathcal{T}(c_0), \mathcal{T}(c_1), \ldots$ Moreover, associating to c_{n+1} ($n \geq 0$) the letter d_n, we see that the mapping $\mathcal{T}$ is nothing else but the morphism β. Indeed, the only integer in $\mathcal{T}(c_n)$ not belonging to c is the least element of $\mathcal{T}(c_n)$, namely $2c_n$. Let us prove this claim. Suppose $d_n = 1$; then we have $c_{n+1} - c_n = 1$ and $\mathcal{T}(c_n) = \{2c_n, 2c_n + 1\}$ and $2c_n \notin c, 2c_n + 1 \in c$ is obviously true. Suppose now $d_n = 2$. We obviously have $2c_n \notin c$, $2c_n + 1, 2c_n + 3 \in c$. Moreover, we have $2c_n + 2 \in c$ since $c_n + 1 \notin c$ because $c_{n+1} - c_n = 2$. The equality $\chi = 1\beta(d)$ is straightforward. This concludes the proof of Theorem 4.7. □

Acknowledgement

We wish to thank the referees for their comments that helped improve the organization of the paper.

References

[1] Allouche, J.-P. *et al.* (1995). A relative of the Thue–Morse sequence. *Discrete Math.* **139**(1–3) 455–461.

[2] Berstel, J. (1995). *Axel Thue's papers on repetitions in words: a translation*. In: *Publications du LaCIM*, vol. 20. Université du Québec à Montréal.

[3] Berstel, J., de Luca, A. (1997). Sturmian Words, Lyndon Words and Trees. *Theor. Comput. Sci.*. To appear.

[4] Brlek, S. (1989). Enumeration of Factors in the Thue–Morse word. *Discrete Appl. Math.* **24**(1–3) 351–354.

[5] Chen, K. T., Fox, R. H., Lyndon, R. C. (1958). Free Differential Calculus, IV – The Quotient Groups of the Lower Central Series. *Ann. Math.* **68** 81–95.

[6] Ido, A. (1996). Factorisation du mot de Thue–Morse et de deux mots cousins. *Technical Report 1146–96*, LaBRI, U.R.A. CNRS # 1304, Université Bordeaux I.

[7] Kimberling, C. (1980). Problem E 2850. *Am. Math. Monthly* **87** 351–354.

[8] Lothaire, M. (1983) *Combinatorics on Words*. Addison-Wesley.

[9] Melançon, G. (1992). Combinatorics of Hall Trees and Hall Words. *J. Combin. Theory A* **59**(2) 285–308.

[10] Melançon, G. (1996). Lyndon Factorization of Infinite Words. In: Puech, C., Reischuk, R., editors, *STACS '96, 13th Annual Symposium on Theoretical Aspects of Computer Science*. Lecture Notes in Computer Science 1046, pp. 147–154. Springer-Verlag.

[11] Melançon, G. (1996). Viennot Factorizations of Infinite Words. *Infor. Process. Lette.* **60** 53–57.

[12] Melançon, G. (1997). Lyndon Factorization of Sturmian Words. *Discrete Math.* To appear.

[13] Reutenauer, C. (1986). Mots de Lyndon et un théorème de Shirshov. *Annales des Sciences Mathématiques du Québec* **10**(2) 237–245.

[14] Reutenauer, C. (1993). *Free Lie Algebras.London Mathematical Society Monographs New Series* . Oxford University Press.

[15] Shallit, J. (1988). A Generalization of Automatic Sequences. *Theor. Comput. Sci.* **61** 1–16.

[16] Siromoney, R., Matthew, L., Dare, V. R., Subramanian, K. G. (1994). Infinite Lyndon Words. *Infor. Process. Lett.* **50** 101–104.

[17] Varricchio, S. (1990). Factorizations of Free Monoids and Unavoidable Regularities. *Theor. Comput. Sci.* **73** 81–89.

[18] Viennot, X. (1978). *Algèbres de Lie libres et monoïdes libres*. Lecture Notes in Mathematics 691. Springer-Verlag.