ELSEVIER

Discrete Mathematics 210 (2000) 137–149

DISCRETE MATHEMATICS



# Lyndon factorization of sturmian words

Guy Melançon *,[1]

*LaBRI, URA 1304 CNRS-Université Bordeaux I, France*

Received 19 July 1996

**Abstract**

We express any general characteristic sturmian word as a unique infinite non-increasing product of Lyndon words. Using this identity, we give a new $\omega$-division for characteristic sturmian words. We also give a short proof of a result by Berstel and de Luca (Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1997) 171–203.); more precisely, we show that the set of factors of sturmian words that qualify as Lyndon words is the set of primitive Christoffel words.


## 1. Introduction

Infinite sturmian words appear through many chapters of the litterature: number theory, combinatorics of dynamical systems, combinatorics on words, as well as theoretical computer science. These (right) infinite words have both geometrical and combinatorial characterizations (cf. [15,1,7] or [3]). The present paper proposes to look at characteristic sturmian words, using Lyndon factorization of infinite words.

Lyndon words are minimal representatives of primitive conjugacy classes (w.r.t. the lexicographical order); equivalently they are strictly smaller than their non-empty proper right factors. The Lyndon factorization theorem [5] asserts that any word can be expressed as a non-increasing product of Lyndon words. A beautiful algorithm by Duval [8], exploiting the combinatorics of Lyndon words, computes this factorization in linear time. Siromoney et al. [17] introduced infinite Lyndon words and gave a generalization of Lyndon's theorem: any right infinite word may be expressed as a non-increasing product of Lyndon words (finite or infinite) (cf. Theorem 2.5).

The results of the present paper confirm Siromoney's factorization theorem as a useful tool for studying right infinite words. The usefulness of this technique of

* LaBRI, Université de Bordeaux I, 351, Cours de la Libération, 33405 Talence Cedex, France.
*E-mail address:* melancon@labri.u-bordeaux.fr (G. Melançon)
[1] Work partially supported by EC grant CHRX-CT93-0400.

0012-365X/00/$ - see front matter 
PII: S0012-365X(99)00123-5

investigation was also mentionned in [12]. We first recall in Section 2 the basic facts we need about Lyndon words. Section 3 contains our central result: we give the explicit computation of the factorization of any characteristic sturmian word $s$ as a non-increasing product of finite Lyndon words (Theorem 3.3):

$$s = \prod_{n \geqslant 0} [(a\bar{s}_{2n+1})^{c_{2n}-1} a s_{2n} \bar{s}_{2n+1}]^{c_{2n+1}}$$

in terms of exponents $(c_n)_{n \geqslant 0}$ intimately linked to the word $s$.

In the last section, we show how one can use this information on $s$ and give two applications. First, we prove that the factorization of $s$ gives an $\omega$-division for it. Second, we give a short proof of a recent result by Berstel and de Luca [3]: we show that the set of factors of sturmian words that qualify as Lyndon words is equal to the set of primitive Christoffel words (cf. Definition 4.3).

## 2. Lyndon words, finite and infinite

The basic definitions and notations we use are those usual in theoretical computer science (see [10]). We denote by $A = \{a, b\}$ the two letter alphabet and suppose it is totally ordered by $a < b$. This order is naturally extended to the set of all words $A^*$ lexicographically.

### 2.1. Finite Lyndon words

Let us go through basic facts about Lyndon words; for details, the reader is referred to [10, Chapter. 5.1]. All the results concerning Lyndon words (finite or infinite) we state here hold true over an arbitrary alphabet. Recall that a word $w \in A^*$ is a *Lyndon word* if it is strictly smaller than any of its non-empty proper right factors (w.r.t. the lexicographical order $<$). An equivalent definition may be given in terms of conjugation of words (cf. [10]): any Lyndon word is primitive and minimal in its conjugacy class. Recall that $w \in A^+$ is primitive if it is not a proper power of another word $u$, that is, $w = u^n$ implies $n = 1$ and $w = u$. For example, with $A = \{a, b\}$, the word $aababb$ is a Lyndon word, with conjugacy class $\{aababb, ababba, babbaa, abbaab, bbaaba, baabab\}$. Remark that, in particular, letters are Lyndon words.

Denote by $L$ the set of Lyndon words over $A$. Any Lyndon word $w \in L$ of length $\geqslant 2$ may be expressed as a product of two Lyndon words, $w = uv$, with $u, v \in L$ and $u < v$. This factorization may not be unique; indeed, we have $aababb = (a)(ababb) = (aab)(abb) = (aabab)(b)$. Let $v$ be the longest right factor of $w$ that qualifies as a Lyndon word. Then $w = uv$, and we have $u \in L$ and $u < uv < v$. This factorization of $w$ is called its *right standard factorization*. Given a Lyndon word $w \in L \setminus A$, we will write $w = w'w''$ to denote the left and right factors of its right standard factorization. We may similarly define the left standard factorization of a Lyndon word, by taking $u$ of maximal length.

**Proposition 2.1.** *Let $u,v \in L$ be such that $u < v$ and suppose $u$ has right standard factorization $u = u'u''$. Then the factorization $uv$ is right standard if and only if $u'' \geqslant v$.*

*Similarly, suppose $v$ has left standard factorization $v = v'v''$. Then the factorization $uv$ is left standard if and only if $v' \leqslant u$.*

**Corollary 2.2.** *Let $u,v \in L$. We have $uv \in L$ iff $u < v$. Consequently, for all $p,q \geqslant 1$, the word $u^p v^q \in L$ is a Lyndon word. Moreover, suppose $uv$ is a right standard factorization (i.e. $u \in A$ or $u'' \geqslant v$). Then $u^p v^q$ has right standard factorization:*

$$\begin{aligned}(u^p v^q)'' &= u^{p-1}v^q \\ (u^p v^q)' &= u\end{aligned} \quad \text{if } p \geqslant 2,$$

$$\begin{aligned}(u^p v^q)'' &= v \\ (u^p v^q)' &= uv^{q-1}\end{aligned} \quad \text{if } p = 1.$$

**Corollary 2.3.** *Let $u,v \in L$ be such that $u < v$ and suppose that the factorization $uv$ is right standard. Then left and right standard factorizations of the Lyndon words $uv^q$ and $u^p v$ coincide, for any $p,q \geqslant 1$.*

The proposition is originally from [5]; for a proof, the reader is refered to [10].

The first corollary is from Duval [8]. The case $p \geqslant 2$ follows from $u'' \geqslant v > u^{p-1}v^q$. The case $p = 1$ follows by induction on $q$; indeed, suppose $(uv^{q-1})'' = v$ (if $q \geqslant 2$) or $u''$ (if $q = 1$), then $(uv^{q-1})'' \geqslant v$.

Let us prove the last corollary. It suffices to prove that the right and left standard factorizations of $uv^q$ coincide (the proof for $u^p v$ is similar). We momentarily denote by $(w)'_r, (w)''_r$ and $(w)'_l, (w)''_l$ the right and left standard factorizations of Lyndon words. According to Proposition 2.1, the right standard factorization of $uv^q$ is $(uv^q)''_r = v$ and $(uv^q)'_r = uv^{q-1}$. One may give a left version of the preceding proposition and find that $(u^p v^q)''_l = v, (u^p v^q)'_l = u^p v^{q-1}$ if $q \geqslant 2$ and $(u^p v^q)''_l = u^{p-1}v, (u^p v^q)'_l = u$ if $q = 1$. From this it follows that $(uv^q)''_l = v = (uv^q)''_r$, and $(uv^q)'_l = uv^{q-1} = (uv^q)'_r$.

Recall that a word $v \in A^*$ is a *factor* of a word $w$ if $w = uvt$, where $u,t \in A^*$.

**Proposition 2.4.** *The left and right standard factorizations of a Lyndon word $w$ coincide if and only if $w$ is a unique increasing product $w = uv$ $(u < v)$ of two Lyndon words. In that case, if $r$ is a factor of $w$ that qualifies as a Lyndon word, then either it is equal to $w$ itself, or it is a factor of $u$ or a factor of $v$.*

The first statement of the proposition is trivial. The proof of the proposition relies on a special property of Lyndon words. Suppose $xy \in L$ (with $x, y$ non-empty) and $r \in L$ are such that $y < r$ then $(xy)^n xr \in L$, for any $n \geqslant 0$. Suppose that $r$ is a Lyndon factor of $w$ which overlaps both $u$ and $v$. Then $r = yz$ with $u = xy$ and $v = zt$ and we have $y < r$ since $y$ is a left factor of $r$. Then, by virtue of the result we just described, we have $xr = uz \in L$. But that contradicts the fact that $u$ is the longest left factor of $w$ qualifying as a Lyndon word.

### 2.2. Infinite Lyndon words

A right *infinite word* $s$ is a sequence of letters $(a_i)_{i\geqslant 0}$, written as $s = a_0a_1a_2\ldots$ Siromoney et al. [17] introduced *infinite Lyndon words* as limits of sequences of finite Lyndon words. Recall the Lyndon factorization theorem [5]: any non-empty word can be expressed as a non-increasing product of Lyndon words (cf. [10, Theorem 5.1.5]). In [17], the authors showed how this theorem may be extended to (right) infinite words. Let us state their result:

**Theorem 2.5** (Siromoney et al. [17, Theorem 2.4]). *Any right infinite word $s$ may be uniquely expressed as a non-increasing product of Lyndon words, finite or infinite, in one of the two following forms: either there exists an infinite non-increasing sequence of finite Lyndon words $(\ell_k)_{k\geqslant 0}$ such that*

$$s = \prod_{n\geqslant 0} \ell_n = \ell_0\ell_1\ldots, \tag{1}$$

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

$$s = \ell_0\ldots\ell_{m-1}\ell_m, \quad \text{with } \ell_0\geqslant\cdots\geqslant\ell_{m-1} > \ell_m. \tag{2}$$

In [18, Theorem 3.7], Varricchio implicitely shows that certain infinite words admit a factorization of type (1) over Viennot factorizations (for a definition, see [10, Theorem 5.4.4]; see also Remark 3.6).

In this paper, we compute the explicit Lyndon factorization of characteristic sturmian words. They all have a factorization of type (1). Since infinite Lyndon words do not appear in these factorizations, we do not take time here to define them, and refer the interested reader to [17]. We look at characteristic sturmian words and show how one can get information about them out of their factorization (1). In particular, we give a proof of a recent result by Berstel and de Luca [3].

We close this section by stating a result we will need in Section 4.

**Proposition 2.6.** *Let $s$ be an infinite word with unique non-increasing factorization (finite or infinite):*

$$s = \ell_0\ell_1\ell_2\ldots\ .$$

*A word $u\in L$ is a factor of $s$ if and only if it is a factor of one of the $\ell_i$'s.*

This is a consequence of a general result on factorizations of the free monoid according to which a factor of the form $v\ell_{i+1}\ldots\ell_{j-1}w$ (where $v$ and $w$ are right and left factors of $\ell_i$ and $\ell_j$, respectively) factorizes into a non-increasing product of at least two Lyndon words (see [11]).

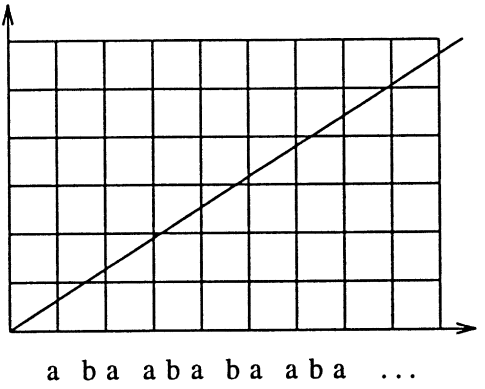


Fig. 1. The first letters of the Fibonacci word, with associated slope $(-1+\sqrt{5})/2$.

## 3. Factorization of characteristic sturmian words

As Berstel and de Luca [3] point out, infinite sturmian words may be defined either geometrically or combinatorially. Combinatorially, they may be defined as infinite non-ultimately periodic words having a minimal number $p(n)$ of factors of length $n$. It can be shown that an infinite word $s$ is ultimately periodic if there exist an integer $n_0$ such that $p(n_0)=n_0$. Hence, the complexity functions of sturmian words satisfy $p(n)=n+1$. Hence, sturmian words are two letter words. We refer the reader to (the bibliography of) [2] for a proof of these elementary facts. Geometrically they correspond to lines in the planes. More precisely, we draw a half-line having a given irrationnal slope $\theta>0$ (and a $y$-axis ordinate). Writing a letter $a$ for an intersection with a vertical segment, and a letter $b$ for an intersection with a horizontal segment, we get an infinite word with minimal complexity.

The Fibonacci word is with no doubt the world's most famous sturmian word. It is defined as the limit of a sequence of finite words $f_0=b$, $f_1=a$ and for $n\geqslant 1$, $f_{n+1}=f_nf_{n-1}$. Hence, $f=abaababaabaababaabab\ldots$ . Note that the length of the words $f_n$ correspond to the Fibonacci sequence of integers; the associated slope of the Fibonacci word is the golden ratio (see Fig. 1).

### 3.1. Characteristic sturmian words

Let $(c_n)_{n\geqslant 0}$ be any sequence of integers satisfying $c_0\geqslant 0$ and $c_n\geqslant 1$ for $n\geqslant 1$. Define finite words $s_0=b$ and $s_1=a$, and $s_{n+1}=s_n^{c_{n-1}}s_{n-1}$ for $n\geqslant 1$. A result by Rauzy [15] asserts that the word $s=\lim s_n$ is infinite sturmian. For example, if $c_n=1$ for all $n\geqslant 0$, the word we get is the Fibonacci word. Note that $c_0=0$ corresponds to exchanging $a$'s with $b$'s in $s$. The infinite word $t$ obtained from $s$ by exchanging $a$'s and $b$'s is a sturmian word with directive sequence $(d_n)_{n\geqslant 0}$, where $d_n=c_{n+1}$.

Not all sturmian words are obtained this way. An infinite sturmian word $s$ obtained with Rauzy's process is called a *characteristic sturmian word* with *directive sequence*

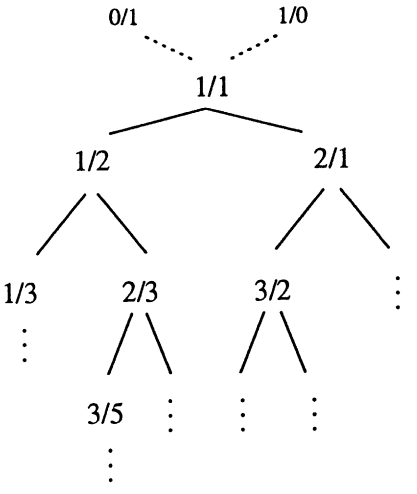


Fig. 2. The Stern–Brocot tree.

$(c_n)_{n\geqslant 0}$. Characteristic sturmian words correspond to half-lines starting at the origin. They form an important subclass among all sturmian words; more precisely, one can show that the set of finite factors of a given sturmian word only depends on the slope of the line associated with it. Thus, as far as we are concerned with finite factors of sturmian words, we may restrict ourselves to characteristic sturmian words.

**Remark 3.1.** In their paper, Berstel and de Luca [3] describe the Stern–Brocot tree and the numerous and beautiful results they give show (among many things) how its structure links the directive sequence $(c_n)_{n\geqslant 0}$ of $s$, and the slope $\theta$ associated with $s$. We collect here some observations borrowed from their paper, that may help us to get a better understanding of the results in the last section. The Stern–Brocot tree is an infinite rooted planar binary tree, with the reduced fraction $\frac{1}{1}$ at its root (Fig. 2). The reduced fraction sitting at a node is $(p'+p'')/(q'+q'')$, where $p'/q'$ and $p''/q''$ are the reduced fraction sitting at the nearest right and left ancestor of the node, respectively. This tree contains every positive rational in its reduced form exactly once. The left–right path one has to follow to go from the root to a given rational $p/q$ may be coded as a unique word of the form $R^{a_0}L^{a_1}\ldots R^{a_n}$ where each exponent is strictly positive with the exception that $a_0\geqslant 0$. These exponents $a_0, a_1, \ldots, a_n$ correspond to the (finite) continued fraction of $p/q$. This extends to irrational numbers as well. Indeed, an irrationnal number $\theta$ viewed as the limit of a sequence of rational numbers corresponds to an infinite left–right path in the tree, thus to an infinite word $R^{a_0}L^{a_1}\ldots$, the exponents corresponding this time to the (infinite) continued fraction for $\theta$. The link with characteristic sturmian word is natural: the sequence of exponents $(a_n)_{n\geqslant 0}$ coincides exactly with the directive sequence for the characteristic sturmian word with associated slope $\theta$.

For more details on this, and on infinite sturmian words in general, the reader may consult [15,1,7,3] and a recent survey by Berstel [2].

### 3.2. The factorization

We first compute the factorization of the Fibonacci word. The computation of the factorization in the more general case of a characteristic sturmian word follows the same line.

**Proposition 3.2** (cf. [12, Proposition 11]). *The factorization of the Fibonacci word $f$ is of type* (1) *and is given by the sequence of words $(\ell_k)_{k\geqslant 0}$ with $\ell_0 = ab$ and $\ell_{k+1} = \varphi(\ell_k)$, where $\varphi : \{a,b\}^* \to \{a,b\}^*$ is the homomorphism defined by $\varphi(a) = aab$ and $\varphi(b) = ab$. Moreover, we have $|\ell_k| = F_{2k+2}$ (where $F_k$ denotes the kth Fibonacci number, with $F_0 = F_1 = 1$).*

Thus the factorization of $f$ is

$$f = (ab)(aabab)(aabaababaabab)\dots .$$

**Proof.** Every word $f_{2n+1}$ ends with the letter $a$; denote by $\bar{w}$ the word obtained from $w$ by deleting the $a$ at its end (if possible). One shows by induction that the words $\ell_n = a f_{2n} \bar{f}_{2n+1}$ are Lyndon words. Corollary 2.2 then implies $\ell'_n = a\bar{f}_{2n+1}$ and it follows that the sequence $(\ell_n)_{n\geqslant 0}$ is strictly decreasing. It is straightforward to compute $f = \prod_{n\geqslant 0} \ell_n$, after observing that

$$f = f_1 f_0 f_1 f_2 f_3 \dots . \tag{3}$$

The second part of the statement is a consequence of the fact that we have $\ell_0 = ab$ and $\ell_{k+1} = \ell'_k \ell''_k$ (as shows the preceding induction). The result then follows from the fact that the homomorphism $\varphi$ respects standard factorization, i.e. $\varphi(\ell'_k \ell''_k) = \varphi(\ell'_k)\varphi(\ell''_k)$. The equality $|\ell_k| = F_{2k+2}$ is easy.

In [14], it is proved that the words $(\ell_n)_{n\geqslant 0}$ form a prefix code. This follows easily from $\ell_{k+1} = \ell'_k \ell''_k$; indeed, this implies that $\ell_i$ is a suffix of $\ell_j$ if $i < j$. The property then follows from the fact that no word may be both a prefix and a suffix of a given Lyndon word. Eq. (3) was also observed in [14].

We now turn to the general case of a characteristic sturmian word with directive sequence $(c_n)_{n\geqslant 0}$. Note that $s_{2n+1}$ ends with an $a$. Our central result is:

**Theorem 3.3.** *Let $s$ be a characteristic sturmian word with directive sequence $(c_n)_{n\geqslant 0}$. Set $\ell_n = (a\bar{s}_{2n+1})^{c_{2n}-1} a s_{2n} \bar{s}_{2n+1}$, where it is understood that $\ell_0 = b$ if $c_0 = 0$.*

*Then the words $(\ell_n)_{n\geqslant 0}$ form a strictly decreasing sequence of Lyndon words and the unique factorization of $s$ as a non-increasing product of Lyndon words is*

$$s = \prod_{n\geqslant 0} \ell_n^{c_{2n+1}}. \tag{4}$$

**Lemma 3.4.** *The words* $a\bar{s}_{2n+1}$, $as_{2n}\bar{s}_{2n+1}$ *are Lyndon words. Furthermore, one has* $(as_{2n}\bar{s}_{2n+1})' = a\bar{s}_{2n+1}$. *Consequently, the words* $(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}$ *form a strictly decreasing sequence of Lyndon words.*

We proceed by induction. We have $a\bar{s}_1 = a$ and $as_0\bar{s}_1 = ab$. Now compute, for $n \geqslant 0$:

$$\begin{aligned}
&as_{2n+2}\bar{s}_{2n+3}\\
&\quad = a(s_{2n+1}^{c_{2n}}s_{2n})(s_{2n+2}^{c_{2n+1}}\bar{s}_{2n+1})\\
&\quad = a(\bar{s}_{2n+1}(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n})(\bar{s}_{2n+1}(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n})^{c_{2n+1}}\bar{s}_{2n+1}\\
&\quad = a\bar{s}_{2n+1}(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}[(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}]^{c_{2n+1}}\\
&\quad = (a\bar{s}_{2n+1})\ [(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}]^{c_{2n+1}+1} \qquad (5)
\end{aligned}$$

and

$$\begin{aligned}
&a\bar{s}_{2n+3}\\
&\quad = a(s_{2n+2}^{c_{2n+1}}\bar{s}_{2n+1})\\
&\quad = a(\bar{s}_{2n+1}(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n})^{c_{2n+1}}\bar{s}_{2n+1}\\
&\quad = a\bar{s}_{2n+1}[(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1})^{c_{2n+1}-1}](a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}\\
&\quad = (a\bar{s}_{2n+1})[(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}]^{c_{2n+1}}. \qquad (6)
\end{aligned}$$

We conclude at once, that $as_{2n+2}\bar{s}_{2n+3}$ and $a\bar{s}_{2n+3}$ are Lyndon words, and that $a\bar{s}_{2n+3} = (as_{2n+2}\bar{s}_{2n+3})'$ by virtue of Corollary 2.2. The sequence of Lyndon words $(\ell_n)_{n\geqslant 0}$ is strictly decreasing since $(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}$ is a right factor of $as_{2n+2}\bar{s}_{2n+3}$.

**Proof of Theorem 3.3.** First we write an identity, analog to (3):

$$\begin{aligned}
s &= s_k^{c_{k-2}}s_{k-1}s_{k+1}^{c_{k-1}-1}s_ks_{k+2}^{c_k-1}s_{k+1}\cdots\\
&= s_{k+1}s_{k+1}^{c_{k-1}-1}s_ks_{k+2}^{c_k-1}s_{k+1}\cdots\\
&= s_{k+1}^{c_{k-1}}s_ks_{k+2}^{c_k-1}s_{k+1}\cdots\ .
\end{aligned}$$

Second, we compute

$$\begin{aligned}
as_{2n+1}^{c_{2n}-1}s_{2n}s_{2n+2}^{c_{2n+1}-1}s_{2n+1} &= [(a\bar{s}_{2n+1})^{c_{2n}-1}as_{2n}\bar{s}_{2n+1}]^{c_{2n+1}}a\\
&= \ell_n^{c_{2n+1}}\ a.
\end{aligned}$$

From this it follows, when $c_0 > 0$, that

$$\begin{aligned}
s &= s_1^{c_0}s_0s_2^{c_1-1}s_1s_3^{c_2-1}s_2s_4^{c_3-1}s_3\cdots\\
&= (as_1^{c_0-1}s_0s_2^{c_1-1}s_1)(s_3^{c_2-1}s_2s_4^{c_3-1}s_3)\cdots\\
&= (\ell_0^{c_1}a)(s_3^{c_2-1}s_2s_4^{c_3-1}s_3)\cdots = \ell_0^{c_1}(as_3^{c_2-1}s_2s_4^{c_3-1}s_3)\cdots\\
&= \prod_{n\geqslant 0}\ell_n^{c_{2n+1}}.
\end{aligned}$$

In the case $c_0 = 0$, we must be careful. Note that $s_2 = s_1^{c_0} s_0 = s_0 = b$ and compute

$$\begin{aligned} s &= s_1^{c_0} s_0 s_2^{c_1-1} s_1 s_3^{c_2-1} s_2 s_4^{c_3-1} s_3 s_5^{c_4-1} s_4 s_6^{c_5-1} s_5 \cdots \\ &= s_2^{c_1} s_1 s_3^{c_2-1} s_2 s_4^{c_3-1} s_3 s_5^{c_4-1} s_4 s_6^{c_5-1} s_5 \cdots \\ &= s_2^{c_1} (a s_3^{c_2-1} s_2 s_4^{c_3-1} s_3)(s_5^{c_4-1} s_4 s_6^{c_5-1} s_5) \cdots \\ &= b^{c_1} (\ell_1^{c_1} a)(s_5^{c_4-1} s_4 s_6^{c_5-1} s_5) \cdots \\ &= b^{c_1} \ell_1^{c_3} (a s_5^{c_4-1} s_4 s_6^{c_5-1} s_5) \cdots \\ &= \prod_{n \geqslant 0} \ell_n^{c_{2n+1}}. \end{aligned}$$

Observe that our notation for $\ell_0$ is in accordance with the usual notation $a^{-1}v$ corresponding to the deletion of $a \in A$ at the beginning of $v \in A^*$ (if possible). Indeed, we then compute $(a\bar{s}_1)^{c_{2n}-1} a s_0 \bar{s}_1 = a^{-1}(ab) = b$.

**Remark 3.5.** Note that we assumed $a < b$. It is natural to ask whether we get a different factorization if we choose $a > b$ instead. In most cases, we would get a completely different result. In our case, the formula we get for $s$ with directive sequence $(c_n)_{n \geqslant 0}$ with $c_0 > 0$ (resp. $c_0 = 0$) when $a < b$ is the same as for $s'$ with directive sequence $(c'_n)_{n \geqslant 0}$ with $c'_0 = 0$ and $c'_{n+1} = c_n$ (resp. $c'_n = c_{n+1}$) when $a > b$ (except that $a$'s and $b$'s have to be exchanged in all Lyndon words $\ell_n$).

This is explained by an invariance property satisfied by the Lyndon factors of sturmian words: this set of words is invariant under the transformation exchanging $a$'s with $b$'s and then taking mirror images (cf. [3, Corollary 3.1]; see Section 4).

**Remark 3.6.** The Lyndon factorization is part of a larger family of factorization called Viennot factorizations (for a definition, see [10, Chapter 5.4]). Fortunately, it is possible to show a complete analog of Theorem 2.5 for Viennot factorizations [12,13]. It is remarkable that the factorization we computed in Theorem 3.3 gives the factorization of the characteristic sturmian word $s$ over *any* Viennot factorization. This is mainly due to the very special properties of Christoffel primitive words.

**Remark 3.7.** Proposition 3.2 raises a natural question. When is the sequence $(\ell_n)_{n \geqslant 0}$ morphic? More precisely, is it possible to give a morphism $\varphi : \{a, b\}^* \to \{a, b\}^*$ and a word $\ell_0 \in L$ such that $\ell_{n+1} = \varphi(\ell_n)$? This question has a positive answer in the case where the directive sequence is constant. For instance, if $c_n = 2$ for all $n \geqslant 0$, then we may set $\ell_0 = aabb$ and use the morphism mapping $a \mapsto aaabaab$ and $b \mapsto aab$.

A characteristic sturmian word may be itself morphic. That is, it may be the limit $\lim_n \varphi^n(a)$ of a (non-erasing) morphism (satisfying $\varphi(a) \in aA^*$). It is known that this is essentially equivalent to the fact that its directive sequence is periodic. Unfortunately, even when a characteristic sturmian word $s$ has a periodic directive sequence, it seems that the sequence $(\ell_n)_{n \geqslant 0}$ is not always morphic, although it is possible to describe patterns in the factorization.

## 4. Applications

### 4.1. ω-division of infinite words

Recall that a finite word is $m$-divided if it can be expressed as $w=x_1 \dots x_m$, such that for all permutation $\sigma \in \Sigma_m$ ($\sigma \neq id$), we have $w > x_{\sigma(1)} \dots x_{\sigma(m)}$ (for a given total order on $A^*$). This definition can be extended to infinite words by asking for a factorization $s=x_1x_2\dots$ into finite words $x_i \in A^*$, to give rise to $m$-divided finite words $x_i \dots x_{i+m-1}$ for all $i \geqslant 1$ and $m \geqslant 2$.

**Corollary 4.1.** *Let $s$ be a characteristic sturmian word. Then the factorization of $s$, $s = \prod_{n \geqslant 0} x_n$ with $x_n = \ell_n^{c_{2n+1}}$ is an ω-division for $s$ (w.r.t. the lexicographical order).*

De Luca [6] showed that sturmian words are $\omega$-divided words using a different factorization. Corollary 4.1 is a consequence of Theorem 3.3 together with a result by Reutenauer [16] according to which the decreasing factorization of a finite word $w=\ell_1^{n_1} \dots \ell_m^{n_m}$ into *distinct* Lyndon words is an $m$-division of that word. This $\omega$-division, in the particular case of the Fibonacci word, also appears in a work by Pirillo [14]. In [12, Proposition 15], we give a more general result according to which any infinite word having a non-ultimately periodic factorization of type (1), is $\omega$-divided. Compare with [18, Theorem 3.7].

### 4.2. Lyndon factors of sturmian words

Using Theorem 3.3 we give a short proof of a result by Berstel and de Luca [3]. We say that a finite word $v \in A^*$ is a *factor* of an infinite word $s$ if $s = uvt$ (where $u \in A^*$ is finite, and $t$ is infinite). Denote by $St$ the set of factors of sturmian words; thus $St = \{v \in A^*: \exists$ a sturmian word $s$ such that $v$ is a factor of $s\}$. Recall that $St$ coincides with the set of factors of all characteristic sturmian words (cf. Section 3). Let $L \cap St$ denote the factors of sturmian words that qualify as Lyndon words.

**Corollary 4.2.** *The set $L \cap St$ of factors of sturmian words that qualify as Lyndon words is equal to the set CP of* primitive Christoffel *words.*

We recall the definition of primitive Christoffel words. Associate with any word $w \in \{a,b\}^*$ a path in the discrete plane $Z \times Z$, starting at the origin: to a letter $a$ corresponds a horizontal segment $(i,j) \to (i+1,j)$ and to a letter $b$ corresponds a vertical segment $(i,j) \to (i,j+1)$ (see Fig. 3).

**Definition 4.3.** A *primitive Christoffel word* is a word such that its path is below the line segment joining the end points of the path, and such that the region thus formed does not contain points with integer coordinates. By convention, letters are primitive Christoffel words.

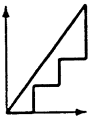

Fig. 3. The word *abababb* $\in CP$ with associated slope $\frac{4}{3}$.

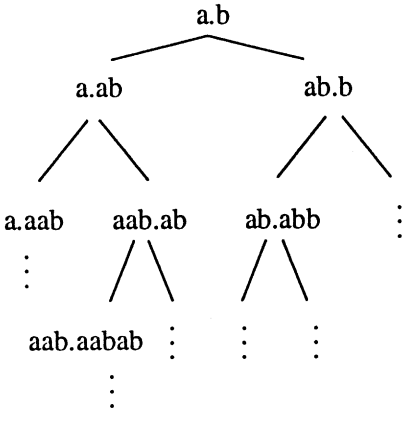


Fig. 4. The Christoffel tree.

These words play a central role in algorithmic number theory (see [4]). Primitive Christoffel words are easily obtained: draw a line segment with rational slope $m = p/q$ (in reduced form) joining two points in $Z \times Z$. The unique primitive Christoffel words corresponding to the given line segment is obtained by forming the unique path crossing the points $(i, \lfloor i * p/q \rfloor)$ $(i = 1, \ldots, p+q)$. One can show that primitive Christoffel words are Lyndon words and that the standard factorization of a primitive Christoffel word corresponds to its unique factorization into a product of two primitive Christoffel words. Moreover, a primitive Christoffel word is intimately linked to the slope $m$ of the line segment joining the end points of its path: it is maximal amongst all Lyndon words having an associated line segment with slope $m$ (see [4,9]). Christoffel primitive words may be generated using a recursive process.

Again, the following observations are borrowed from [3]. This recursive process is a tree process and we may build an infinite rooted planar binary tree having at its root the primitive Christoffel word $ab$ (with standard factorization $(a)(b)$). Suppose the primitive Christoffel word $w$ attached to a node has standard factorization $w = w'w''$; then the primitive Christoffel words attached to the left and right sons of this nodes are $w'w$ and $ww''$, respectively (see Fig. 4). This tree contains every primitive Christoffel word exactly once. Moreover, the unique rational corresponding to a given primitive Christoffel word sits at the corresponding node in the Stern–Brocot tree, as the reader may check.

**Lemma 4.4.** *Let $s$ be a characteristic sturmian word with directive sequence $(c_n)_{n\geqslant 0}$, factorizing as in Eq.* (4), *and let $w \in L$ be a Lyndon word. If $w$ is a factor of $s$ then it is equal to $\ell_n$, $as_{2n}\bar{s}_{2n+1}$ or to $a\bar{s}_{2n+1}$, for some $n \geqslant 0$.*

**Proof.** This is a consequence of Proposition 2.4, according to which $w$ is either equal to or is a factor of one of $\ell_n$, for some $n \geqslant 0$. Suppose $w$ is a proper factor of $\ell_n$. Observe that $\ell_n = (a\bar{s}_{2n+1})^{c_{2n}-1} as_{2n}\bar{s}_{2n+1}$ and $\ell + n' = a\bar{s}_{2n+1} = (as_{2n}\bar{s}_{2n+1})'$ (by Lemma 3.4), so that by Corollary 2.3 the left and right factorization $\ell_n$ coincide. This implies, using Proposition 2.4, that $w$ is a factor of $a\bar{s}_{2n+1}$ or a factor of $as_{2n}\bar{s}_{2n+1}$. Eqs. (5) and (6) show that these may be expressed in terms of $\ell_{n-1}$ and $\ell'_{n-1}$. An easy induction concludes the proof. □

**Proof of Corollary 4.2.** We first show that the Lyndon words $\ell_n$, $as_{2n}\bar{s}_{2n+1}$ and $a\bar{s}_{2n+1}$ are primitive Christoffel words. Observe that if $w$ is a primitive Christoffel word, then for any $p, q \geqslant 1$, $(w')^p w$ and $w(w'')^q$ are Christoffel words. In the Christoffel tree, one goes from $w$ to either one of these by following a left or right extreme path of length $p$ or $q$. Hence, the result follows from an easy induction using Eqs. (5) and (6) and $\ell_n = (a\bar{s}_{2n+1})^{c_{2n}-1} as_{2n}\bar{s}_{2n+1}$.

Let $w \in CP$ be any Christoffel primitive word. As suggested by the Christoffel tree, there exists a unique sequence of integers $a_0, a_1, \ldots, a_n$ such that $w$ is obtained by forming the sequence of words:

$$u_0 = a, \quad v_0 = b, \quad u_{i+1} = u_i v_i^{\alpha_{2i}}, \quad v_{i+1} = u_{i+1}^{\alpha_{2i+1}} v_i.$$

Each word $u_i$ or $v_j$ is a primitive Christoffel word and $w$ is either $u_n$ or $v_n$ (according to the parity of $n$). Now, suppose $a_0 \neq 0$. Let $(c_n)_{n\geqslant 0}$ be any directive sequence satisfying $c_0 = 0, c_1 = a_0, \ldots, c_{n+1} = a_n$. One computes $u_i = a\bar{s}_{2i+1}$, and $v_j = \ell_j$. The case $a_0 = 0$ is similar. This shows that the primitive Christoffel word $w$ is a Lyndon factor of a characteristic sturmian word $s$. Lemma 4.4 helps us conclude that the set $L \cap St$ of Lyndon factors of sturmian words is exactly the set $CP$ of primitive Christoffel words. □

## References

[1] J. Berstel, Tracés de droite, fractions continues et morphismes itérées, in: Mots — Mélanges offerts à M.P. Schützenberger, Hermès, 1990, pp. 298–309.

[2] J. Berstel, Recent results in sturmian words, invited paper to DLT'95, World Scientific, Singapore, 1996, in preparation. http://www-igm.univ-mlv.fr/berstel/index.html.

[3] J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1997) 171–203.

[4] J.P. Borel, F. Laubie, Quelques mots sur la droite projective réelle, J. Théorie Nombres Bordeaux 5 (1993) 23–51.

[5] K.T. Chen, R.H. Fox, R.C. Lyndon, Free differential calculus, IV — the quotient groups of the lower central series, Ann. Math. 68 (1958) 81–95.

[6] A. de Luca, A division property of the Fibonacci word, Technical Report 95/1, Università degli Studi di Roma "La Sapienza", 1995.

[7] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics, Theoret. Comput. Sci. (Special Issue on Formal Languages), to appear.
[8] J.P. Duval, Factorizing words over an ordered alphabet, J. Algorithms 4 (1983) 363–381.
[9] E. Laurier, Opérations sur les mots de Christoffel, Ph.D. Thesis, Université de Limoges, 1995.
[10] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.
[11] G. Melançon, Combinatorics of Hall trees and Hall words, J. Combin. Theory Ser. A 59 (2) (1992) 285–308.
[12] G. Melançon, Lyndon factorization of infinite words, in: C. Puech, R. Reischuk (Eds.), STACS '96, 13th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science, vol. 1046, Springer, Berlin, 1996, pp. 147–154.
[13] G. Melançon, Viennot factorizations of infinite words, Inform. Process. Lett. 60 (1996) 53–57.
[14] G. Pirillo, Some remarks on the Fibonacci infinite word, in: Kokyuroku 910 Semigroups Formal Languages and Combinatoris on Words, RIMS Kyoto University, Kyoto, Japan, 1994.
[15] G. Rauzy, Automata in infinite words, in: M. Nivat, D. Perrin (Eds.), Mots infinis en arithmétique, Lecture Notes in Computer Science, vol. 192, Springer, Berlin, 1984, pp. 164–171.
[16] C. Reutenauer, Mots de Lyndon et un théorème de Shirshov, Ann. Sci. Math. du Québec 10(2) (1986) 237–245.
[17] R. Siromoney, L. Matthew, V.R. Dare, K.G. Subramanian, Infinite Lyndon Words, Inform. Process. Lett. 50 (1994) 101–104.
[18] S. Varricchio, Factorizations of free monoids and unavoidable regularities, Theoret. Comput. Sci. 73 (1990) 81–89.