# RECENT RESULTS IN STURMIAN WORDS[1]

JEAN BERSTEL
*LITP, IBP, Université Pierre et Marie Curie, 4, place Jussieu*
*F-75252 Paris Cedex 05, France*
*e-mail:* Jean.Berstel@litp.ibp.fr

ABSTRACT

In this survey paper, we present some recent results concerning finite and infinite Sturmian words. We emphasize on the different definitions of Sturmian words, and various subclasses, and give the ways to construct them related to continued fraction expansion. Next, we give properties of special finite Sturmian words, called standard words. Among these, a decomposition into palindromes, a relation with the periodicity theorem of Fine and Wilf, and the fact that all these words are Lyndon words. Finally, we describe the structure of Sturmian morphisms (i.e. morphisms that preserve Sturmian words) which is now rather well understood.

## 1 Introduction

Combinatorial properties of finite and infinite words are of increasing importance in various fields of physics, biology, mathematics and computer science. Infinite words generated by various devices have been considered [9]. We are interested here in a special family of infinite words, namely Sturmian words. Sturmian words represent the simplest family of quasi-crystals (see e.g. [3]). They have numerous other properties, related to continued fraction expansion (see e.g. [5, 8]). There are numerous relations with other applications, such as pattern recognition. Early results are reported in [25, 23].

In this survey paper, we start with the basic definitions of finite and infinite Sturmian words , for characteristic words and Christoffel words, and describe their relation with continued fraction expansion.

Next, we give a description of all Sturmian morphisms, and a characterization in terms of automorphisms of a free group.

Finally, we give various properties and characterizations of standard words. These are inductively defined, and are in fact special prefixes of characteristic words.

An *infinite word* is here a mapping

$$\mathbf{x} : \mathbb{N}_+ \rightarrow A$$

where $\mathbb{N}_+ = \{1, 2, \cdots\}$ is the set of positive integers and $A$ is an alphabet. In the sequel, we consider binary words, that is words over a two letter alphabet $A = \{a, b\}$. $A^\omega$ is the set of infinite words over $A$ and $A^\infty = A^* \cup A^\omega$.

[1]Invited paper to DLT'95. Proceedings to be published by World Scientific.

Let $f : A^* \to A^*$ be a morphism. Assume that, for some letter $a$, the word $f(a)$ starts with $a$. Then $f^{n+1}(a)$ starts with $f^n(a)$ for all $n$. If the set $\{f^n(a) \mid n \geq 0\}$ is infinite, then there exists a unique infinite word $\mathbf{x}$ such that every $f^n(a)$ is a prefix of $\mathbf{x}$. The word $\mathbf{x}$ is said to be *generated* by iterating $f$. For general results, see [19, 10]. An infinite word $\mathbf{x}$ is *morphic* if it is generated by iterating a morphism. Any morphism that generates $\mathbf{x}$ is a *generator*.

## 2 Infinite Sturmian words

In this section, we give three equivalent definitions of Sturmian words, first as aperiodic words of minimal complexity, next as words with good distributions of letters (socalled balanced word), finally as discretized straight lines.

The *complexity* function of an infinite word $\mathbf{x}$ is the function $P_{\mathbf{x}}$ where $P_{\mathbf{x}}(n)$ is the number of factors of length $n$ of $\mathbf{x}$. It is well-known (e. g. [6]) that $\mathbf{x}$ is ultimately periodic as soon as $P_{\mathbf{x}}(n) \leq n$ for some $n \geq 0$. Thus, any aperiodic words $\mathbf{x}$ has a complexity function that satisfies $P_{\mathbf{x}}(n) \geq n+1$ for all integers. This leads to the first definition of Sturmian words as those with minimal complexity:

A word $\mathbf{x}$ is *Sturmian* if $P_{\mathbf{x}}(n) = n+1$ for all $n$. Note that by definition, a Sturmian word is over two letters (because $P_{\mathbf{x}}(1) = 2$). This restriction can be overcome simply by requiring that the equality $P_{\mathbf{x}}(n) = n+1$ holds only for great enough $n$ (see e. g. [7]).

Since Sturmian words are aperiodic, the distribution of letters must be somehow irregular. This is described by the next characterization. For any finite or infinite word $w \in A^\infty$, let $\mathrm{Sub}(w)$ denotes the set of *finite factors* of $w$. Next, define the *balance* of a pair $u$ and $v$ of words of same length as the number

$$\delta(u,v) = \Big| |u|_a - |v|_a \Big| = \Big| |u|_b - |v|_b \Big|$$

(Here $|w|_a$ is the number of $a$'s in $w$.) A word $w \in A^\infty$ is *balanced* if $\delta(u,v) \leq 1$ for any $u, v \in \mathrm{Sub}(w)$ with $|u| = |v|$. Thus, in a balanced word over two letters, say $a$ and $b$, occurrences of letters are regularly distributed. In particular, the number of $b$'s between two consecutive $a$'s can take only two values.

Sturmian words are intimately related to straight lines in the plane. This characterization was called the "mechanical" by Morse and Hedlund in [14]. Let $\alpha, \rho$ be real numbers with $0 \leq \alpha < 1$. Consider the infinite words

$$\mathbf{s}_{\alpha,\rho} = a_1 \cdots a_n \cdots, \qquad \mathbf{s}'_{\alpha,\rho} = b_1 \cdots b_n \cdots \tag{1}$$

defined by

$$a_n = \begin{cases} a & \text{if } \lfloor \alpha(n+1) + \rho \rfloor = \lfloor \alpha n + \rho \rfloor \\ b & \text{otherwise} \end{cases} \tag{2}$$

and

$$b_n = \begin{cases} a & \text{if } \lceil \alpha(n+1) + \rho \rceil = \lceil \alpha n + \rho \rceil \\ b & \text{otherwise} \end{cases} \tag{3}$$

Both words are discretized straight lines. Consider indeed the straight line $y = \alpha x + \rho$. Then there are two sets of integral point associated with it: the points $L_n = (n, \lfloor \alpha n + \rho \rfloor)$ and $U_n = (n, \lceil \alpha n + \rho \rceil)$. The encoding of the line is by writing an $a$ whenever the segment $[L_n, L_{n+1}]$ (or $[U_n, U_{n+1}]$) is horizontal, and a $b$ otherwise. Observe that, even if $\alpha > 1$, the expression $\lfloor \alpha(n+1) + \rho \rfloor - \lfloor \alpha n + \rho \rfloor$ can take only two values. If $\alpha < 1$, these are 0 or 1. For $\alpha > 1$, one encounters two formulations: some authors take the function $\lfloor \alpha(n+1) + \rho \rfloor - \lfloor \alpha n + \rho \rfloor - \lfloor \alpha \rfloor$, with values 0 or 1, others consider words over the alphaber $\{k, k+1\}$, where $k$ and $k+1$ are the two values of the function.

Reverting to the case $0 < \alpha < 1$, observe also that the two infinite words $\mathbf{s}_{\alpha,\rho}$ and $\mathbf{s}'_{\alpha,\rho}$ are equal except in the case where $\alpha n + \rho$ is an integer for some $n$. If this holds, then $a_{n-1}a_n = ba$ and $b_{n-1}b_n = ab$. This happens for $n = 0$ in the important special case where $\rho = 0$. However, this is ruled out by our convention (and this is in fact the reason of this convention) that indices start at 1.

The present definition as discretized lines shows that, by changing the starting point (i. e. by replacing $\rho$ by some $\rho + m\alpha$), one gets the same kind of infinite words. For this reason, a variation of the definition is to consider twosided infinite words. This theory has been developed carefully by Coven and Hedlund [6].

The following theorem states that the three basic definitions of Sturmian words are indeed equivalent.

**Theorem 2.1** [6, 14] *Let* $\mathbf{x}$ *be an infinite binary word. The following conditions are equivalent:*

(*i*) $\mathbf{x}$ *is Sturmian;*
(*ii*) $\mathbf{x}$ *is balanced and not ultimately periodic;*
(*iii*) *there exist an irrational number* $\alpha$ $(0 < \alpha < 1)$ *and a real* $\rho$ *such that* $\mathbf{x} = \mathbf{s}_{\alpha,\rho}$ *or* $\mathbf{x} = \mathbf{s}'_{\alpha,\rho}$.

Several proofs of this result exist. It was proved first by Hedlund and Morse in 1940; another proof is by Coven and Hedlund. These are of combinatorial nature, another, based on geometric considerations, is due to Lunnon and Pleasants [15]. Many proofs of partial results have appeared in the literature. Observe that the theorem does not hold in this formulation for twosided infinite words, since $a^\omega b^\omega$ has complexity $n+1$ but is not balanced.

A Sturmian word $\mathbf{x}$ is *characteristic* if $\mathbf{x} = \mathbf{s}_{\alpha,0}$ for some irrational $\alpha$ $(0 < \alpha < 1)$. We write then $\mathbf{c}_\alpha = \mathbf{s}_{\alpha,0}$. In this case, $\mathbf{s}_{\alpha,0} = \mathbf{s}'_{\alpha,0}$. The number $\alpha$ is the *slope* of $\mathbf{x}$, and $\mathbf{x}$ is characteristic *for* the number $\alpha$. By definition, the slope is the limit of the quotients $|u|_b/|u|$, where $u$ ranges over the prefixes of $\mathbf{x}$, and $|u|_b$ is the number of $b$'s in $u$.

Sometimes, a variation of characteristic words are considered. *Christoffel words* are the words $a\mathbf{s}$ and $b\mathbf{s}$, where $\mathbf{s}$ is a characteristic word. In other terms, Christoffel words are discretized straight lines, where the indices start at 0.

The most famous characteristic word is the Fibonacci word

$$\mathbf{f} = abaababaabaab\cdots$$

generated by the morphism

$$\begin{aligned} a &\mapsto ab \\ b &\mapsto a \end{aligned}$$

Its slope is $1/\phi^2$, where $\phi$ is the golden ratio. In view of the preceding theorem, it is clear that for any Sturmian word $\mathbf{s}$, one of the words $a\mathbf{s}$ or $b\mathbf{s}$ is Sturmian. Characteristic words are also described by

**Proposition 2.2** *A Sturmian word* $\mathbf{s}$ *is characteristic iff both* $a\mathbf{s}$ *and* $b\mathbf{s}$ *are Sturmian.*

If $\mathbf{s}$ is characteristic, then $a\mathbf{s}$ and $b\mathbf{s}$ are Christoffel words and $ba\mathbf{s}$ and $ab\mathbf{s}$ are Sturmian. Characteristic words have also be called *homogeneous spectra*, and Sturmian words *inhomogeneous spectra*.

A popular equivalent formulation of the "mechanical" definition of Sturmian words is by rotation. Let $\alpha$ be irrational, $0 < \alpha < 1$. The *rotation* of angle $\alpha$ is the mapping

$$R_\alpha : x \mapsto x + \alpha \bmod 1$$

from $\mathbb{R}/\mathbb{Z}$ into itself. Iterating $R_\alpha$, one gets

$$R_\alpha^n(x) = \{n\alpha + x\}$$

where $\{z\} = z - \lfloor z \rfloor$ denotes the fractional part of $z$. Since

$$\lfloor (n+1)\alpha + x \rfloor = \lfloor n\alpha + x \rfloor \iff R_\alpha^n(x) \in [0, 1-\alpha[$$

the Sturmian word $\mathbf{s}_{\alpha,\rho} = a_1 \cdots a_n \cdots$ is also defined by

$$a_n = \begin{cases} a & \text{if } R_\alpha^n(\rho) \in [0, 1-\alpha[ \\ b & \text{otherwise} \end{cases}$$

Finally, there is a definition of Sturmian words by *cutting sequences*. This notion is exploited by C. Series [22] and Crisp *et al.* [8]. We consider here only the homogeneous case. Consider the square grid consisting of all vertical and horizontal lines through integer points in the first quadrant. Consider a line $y = \beta x$, where $\beta$ is any positive irrational. Label the intersections of $y = \beta x$ with the grid using $a$ if the grid line crossed is vertical, and $b$ if it is horizontal. The sequence of labels, read from the origin out, is the *cutting sequence* of $y = \beta x$ and is denoted by $\mathbf{S}_\beta$. The following proposition (see e. g. [8]) shows the relation of cutting sequences with characteristic words.

**Proposition 2.3** *Let* $0 < \alpha < 1$ *be irrational. Then* $\mathbf{c}_\alpha = \mathbf{S}_\beta$*, where* $\beta = \alpha/(1-\alpha)$*.*

Cutting sequences are equivalent to *billiard sequences*: consider a billiard ball hitting the sides of a square billiard, the reflection being without side-effect. Denoting $a$ the hitting of a vertical side, and $b$ the hitting of a horizontal side, one gets merely the same as a cutting sequence, provided the initial angle of the direction of the ball is irrational.

## 3 Subwords of Sturmian words

Subwords of infinite words are important because of their relation to dynamical systems. Recall that a symbolic *dynamical system* is a set of infinite words that is both closed under the shift operator (the operator that removes the first letter) and topologically closed (for the usual topology, where two words are "close" if they share a long common prefix). It is known that two infinite words $\mathbf{x}$ and $\mathbf{y}$ generate the same dynamical system iff $\mathrm{Sub}(\mathbf{x}) = \mathrm{Sub}(\mathbf{y})$.

**Proposition 3.1** *The dynamical system of a Sturmian word is minimal.*

A system is minimal if it does not strictly contain another system. Minimal systems have an interesting combinatorial characterization: they are exactly those generated by uniformly recurrent words, i. e. infinite words $\mathbf{x}$ such that, for any $n > 0$, there exists an integer $N > 0$ with the property that any subword of $\mathbf{x}$ of length $N$ contains *all* subwords of $\mathbf{x}$ of length $n$.

Concerning the sets of subwords in Sturmian words, the first observation is that they depend only on the slope:

**Proposition 3.2** *Let* $\mathbf{s}$ *and* $\mathbf{t}$ *be Sturmian words.*
(1) *If* $\mathbf{s}$ *and* $\mathbf{t}$ *have the same slope, then* $\mathrm{Sub}(\mathbf{s}) = \mathrm{Sub}(\mathbf{t})$.
(2) *If* $\mathbf{s}$ *et* $\mathbf{t}$ *have distinct slopes, then* $\mathrm{Sub}(\mathbf{s}) \cap \mathrm{Sub}(\mathbf{t})$ *is finite.*

In particular, for any $\rho$, one has $\mathrm{Sub}(\mathbf{s}_{\alpha,\rho}) = \mathrm{Sub}(\mathbf{c}_\alpha)$. Next, since in any Sturmian word $\mathbf{s}$, there are exactly $n+1$ subwords of length $n$, there exists, for each $n$, exactly one subword of length $n$ that can by extended in two ways into a subword of length $n+1$. More precisely, call a word $w$ a *special subword* for $\mathbf{s}$ if $wa, wb \in \mathrm{Sub}(\mathbf{s})$. Then there is exactly one special subword of length $n$ for each $n$ in a Sturmian word. Special words have been determined by F. Mignosi [16]:

**Proposition 3.3** *The special subwords of a Sturmian word* $\mathbf{s}_{\alpha,\rho}$ *are exactly the reversals of the prefixes of the characteristic word* $\mathbf{c}_\alpha = \mathbf{s}_{\alpha,0}$.

## 4 Characteristic words

Characteristic words have numerous additional properties, mainly related to the continued fraction expansion of their slope. They can also be generated systematically. The corresponding formulae are slightly different if one considers characteristic or Christoffel words.

Before describing these properties, we start with the description of the relation between characteristic words and the famous Beatty sequences (see e. g. [23]).

A *Beatty sequence* is a set

$$B = \{\lfloor sn \rfloor \mid n \geq 1\}$$

for some irrational $s$. Two Beatty sequences $B$ and $B'$ are *complementary* if $B$ and $B'$ form a partition of $\mathbb{N}_+ = \{1, 2, \ldots\}$.

**Theorem 4.1** (Beatty) *The sets* $\{\lfloor sn \rfloor \mid n \geq 1\}$ *and* $\{\lfloor s'n \rfloor \mid n \geq 1\}$ *are complementary iff*

$$\frac{1}{s} + \frac{1}{s'} = 1\,.$$

The relation between characteristic words and Beatty sequences is described by the following

**Proposition 4.2** *Let* $\alpha = 1/s$, *and* $\mathbf{c}_\alpha = a_1 a_2 \cdots a_n \cdots$. *Then*

$$\{\lfloor sn \rfloor \mid n \geq 1\} = \{k \mid a_k = b\}\,.$$

Let $E$ be the morphism that exchanges the letters $a$ and $b$:

$$E \;:\; \begin{array}{ccc} a & \mapsto & b \\ b & \mapsto & a \end{array}$$

Then it is easy to check that

$$E(c_\alpha) = \mathbf{c}_{1-\alpha} \tag{4}$$

Indeed, setting $\beta = 1-\alpha$, one has $\alpha n + \beta n = n$ for all $n$, whence $\lfloor \alpha n \rfloor + \lfloor \beta n \rfloor = n-1$ and $(\lfloor \alpha(n+1) \rfloor - \lfloor \alpha n \rfloor) + (\lfloor \beta(n+1) \rfloor - \lfloor \beta n \rfloor) = 1$. This constitutes a proof of Beatty's theorem.

We now turn to the relation between a characteristic word and the continued fraction expansion of its slope. The basic observation is:

**Proposition 4.3** *Let* $\alpha = [0; 1+d_1, d_2, \ldots]$ *be the continued fraction of the irrational* $\alpha$, *with* $0 < \alpha < 1$. *Define a sequence* $(s_n)_{n \geq -1}$ *of words by*

$$s_{-1} = b, \quad s_0 = a, \qquad s_n = s_{n-1}^{d_n} s_{n-2}, \quad (n \geq 1) \tag{5}$$

*Then every* $s_n$, *for* $n \geq 1$, *is a prefix of* $\mathbf{c}_\alpha$ *and*

$$\mathbf{c}_\alpha = \lim_{n \to \infty} s_n\,.$$

The sequence $(d_n)_{n \geq 1}$ is the *directive sequence* of $\mathbf{c}_\alpha$, and the sequence $(s_n)_{n \geq -1}$ is the *standard sequence* of $\mathbf{c}_\alpha$.

**Example** The directive sequence $(d_n)$ for the Fibonacci word is $(1,1,1,\ldots)$, since $1/\phi^2 = [0; 2, 1, 1, \ldots]$, and the standard sequence is the sequence of finite Fibonnaci words.

**Example** Since $1/\phi = [0; 1, 1, 1, \ldots]$, the corresponding standard sequence is $s_1 = b$, $s_2 = ba$, $s_3 = bab, \ldots$. The sequence is obtained from the Fibonacci sequence by exchanging $a$'s and $b$'s, in concordance with equation (4).

**Example** Consider $\alpha = (\sqrt{3}-1)/2 = [0; 2, 1, 2, 1, \ldots]$. The directive sequence is $(1, 1, 2, 1, 2, 1, \ldots)$, and the standard sequence starts with $s_1 = ab$, $s_2 = aba$, $s_3 = abaabaab, \ldots$, whence

$$\mathbf{c}_{(\sqrt{3}-1)/2} = abaabaababaabaabaababaabaabaab \cdots$$

Consider, as in Proposition 2.3, the irrational $\beta = \alpha/(1-\alpha)$, and set $\beta = [e_0; e_1, \ldots]$. Then $e_0 = 0$, and $e_n = d_n$ for $n \geq 1$, if $d_1 > 0$, i.e. if $\alpha < 1/2$, and $e_n = d_{n+2}$ for $n \geq 0$ otherwise. Define

$$t_{-2} = a, \quad t_{-1} = b, \qquad t_n = t_{n-1}^{e_n} t_{n-2} \quad (n \geq 0)$$

Then $t_n = s_n$ or $t_n = s_{n+2}$, and $\mathbf{c}_\alpha = \mathbf{S}_\beta = \lim t_n$. Because of the complete correspondence of the continued fraction for $\beta$ and the construction of the sequence $(t_n)$, this second expression is sometimes preferred.

A similar construction to that of Proposition 4.3 for characteristic words exists for Christoffel words (see e. g. [4, 1]).

**Proposition 4.4** *Let $\alpha = [0; 1+d_1, d_2, \ldots]$ be the continued fraction of the irrational $\alpha$, with $0 < \alpha < 1$. Define three sequences $(u_n)_{n\geq -1}$, $(v_n)_{n\geq -1}$ and $(w_n)_{n\geq -1}$ of words by*

$$u_{-1} = v_{-1} = w_{-1} = b, \quad u_0 = v_0 = w_0 = a$$

*and*

$$\begin{array}{rcll} u_{2n} & = & u_{2n-2}(u_{2n-1})^{d_{2n}} & n \geq 1 \\ u_{2n+1} & = & (u_{2n})^{d_{2n+1}} u_{2n-1} & n \geq 0 \end{array}, \qquad \begin{array}{rcll} v_{2n} & = & (v_{2n-1})^{d_{2n}} v_{2n-2} & n \geq 1 \\ v_{2n+1} & = & v_{2n-1}(v_{2n})^{d_{2n+1}} & n \geq 0 \end{array}$$

$$w_n = w_{n-2}(w_{n-1})^{d_n} \quad n \geq 1$$

*Then*

$$\begin{array}{rclrcl} a\mathbf{c}_\alpha & = & \lim_{n\to\infty} u_n, & b\mathbf{c}_\alpha & = & \lim_{n\to\infty} v_n \\ ab\mathbf{c}_\alpha & = & \lim_{n\to\infty} w_{2n} & ba\mathbf{c}_\alpha & = & \lim_{n\to\infty} w_{2n+1} \, . \end{array}$$

These sequences of words are related altogether, and can be derived from a more basic sequence called the *palindrome sequence.*

**Proposition 4.5** *Let $\alpha = [0; 1+d_1, d_2, \ldots]$ be the continued fraction of the irrational $\alpha$, with $0 < \alpha < 1$. Define a sequence $(\pi_n)_{n\geq -1}$ by $\pi_{-1} = a^{-1}$, $\pi_0 = b^{-1}$ and*

$$\begin{array}{rcll} \pi_{2n} & = & \pi_{2n-2}(ba\pi_{2n-1})^{d_{2n}} & n \geq 1 \\ \pi_{2n+1} & = & (\pi_{2n}ba)^{d_{2n+1}} \pi_{2n-1} & n \geq 0 \end{array}$$

*The words $\pi_n$, for $n \geq 1$ are palindromes; moreover,*

$$\begin{array}{rclrclrcl} s_{2n} & = & \pi_{2n}ba, & u_n & = & a\pi_n b, & w_{2n} & = & ab\pi_{2n}, \\ s_{2n+1} & = & \pi_{2n+1}ab, & v_n & = & b\pi_n a, & w_{2n+1} & = & ba\pi_{2n+1}. \end{array}$$

The palindromes appearing in these sequences have interesting properties, described below.

All these words have the same length. More precisely, let $\alpha = [0; 1+d_1, d_2, \ldots]$ be the continued fraction of the irrational $\alpha$, and define integers by

$$q_{-1} = 1, \quad q_0 = 1, \qquad q_n = d_n q_{n-1} + q_{n-2}, \quad (n \geq 1) \, .$$

Then of course

$$|s_n| = |u_n| = |v_n| = |w_n| = 2 + |\pi_n| = q_n$$

There is a nice interpretation of $\mathbf{c}_\alpha$ in a number system associated to $(q_n)$, (see T. C. Brown [5]). Any integer $m \geq 0$ can be written in the form

$$m = z_h q_h + \cdots + z_0 q_0, \qquad (0 \leq z_i \leq d_{i+1}) \tag{6}$$

and the representation is unique (but we do not need this here) provided

$$z_i = d_{i+1} \implies z_{i-1} = 0 \qquad (i \geq 1)$$

**Proposition 4.6** (Brown) *If $m = z_h q_h + \cdots + z_0 q_0$ as in eq. (6), then the prefix of $\mathbf{c}_\alpha$ of length $m$ has the form*

$$s_h^{z_h} \cdots s_0^{z_0} .$$

There is another relation between characteristic words and Christoffel words, related to lexicographic order. Let $\mathbf{x} = a_1 a_2 \cdots$ and $\mathbf{y} = b_1 b_2 \cdots$ be two infinite words. We write $\mathbf{x} < \mathbf{y}$ when $\mathbf{x}$ is *lexicographically less* than $\mathbf{y}$, i. e. when there is an integer $n \geq 1$ such $a_k = b_k$ for $1 \leq k < n$ and $a_n = a$, $b_n = b$. First, we observe that

**Proposition 4.7** *Let $0 \leq \rho, \rho' < 1$ and let $0 < \alpha < 1$, $\alpha$ be irrational. Then*

$$\mathbf{s}_{\alpha,\rho} < \mathbf{s}_{\alpha,\rho'} \iff \rho < \rho'.$$

Also, the two Christoffel words are extremes for Sturmian words of given slope.

**Proposition 4.8** *Let $0 < \alpha < 1$ be irrational. For any $0 < \rho < 1$, one has*

$$a\mathbf{c}_\alpha < \mathbf{s}_{\alpha,\rho} < b\mathbf{c}_\alpha.$$

It is quite natural to extend the notion of *Lyndon* word to infinite words as follows: a word $\mathbf{x}$ is an infinite Lyndon word iff it is lexicographically less than all its proper suffixes. Borel and Laubie have shown [4]:

**Proposition 4.9** *Let $0 < \alpha < 1$ be irrational. The word $a\mathbf{c}_\alpha$ is lexicographically smaller than all its suffixes, i. e. is an infinite Lyndon word, and $b\mathbf{c}_\alpha$ is lexicographically greater than all of its suffixes.*

Characteristic words are not Lyndon words. In that case, Melançon [18] has proved:

**Theorem 4.10** *Let $0 < \alpha < 1$ be irrational, and let $(d_n)$ be the directive sequence and $(s_n)$ be the standard sequence of $\mathbf{c}_\alpha$. Then*

$$\mathbf{c}_\alpha = \ell_0^{d_2} \ell_1^{d_4} \cdots \ell_n^{d_{2n+2}} \cdots$$

*where the sequence*

$$\ell_n = a s_{2n}^{d_{2n+1}-1} s_{2n-1} s'_{2n}$$

*is a strictly decreasing sequence of finite Lyndon words, and $s'_{2n}$ is just $s_{2n}$ without its last letter.*

Observe that, since $s_{2n+1} = s_{2n}^{d_{2n+1}} s_{2n-1} = s_{2n} s_{2n}^{d_{2n+1}-1} s_{2n-1}$, the Lyndon word $\ell_n$ is a conjugate of $s_{2n+1}$.

## 5 Finite Sturmian words

*Finite Sturmian words* are defined as finite subwords of (infinite) Sturmian words. The following shows that one of the characterizations of Sturmian words also holds for finite words.

**Proposition 5.1** *A word $w$ is a finite Sturmian word iff it is balanced.*

A careful analysis of the property of being balanced shows that a word $w$ is not balanced iff it admits one of the factorizations

$$w = xauayb\tilde{u}bz, \quad \text{or} \quad w = xbubya\tilde{u}az$$

for some word $u$. It follows that [13]:

**Theorem 5.2** (Dulucq, Gouyou-Beauchamps) *The complement of the set of finite Sturmian words is context-free.*

This remarkable property however does not extend to unambiguity: the language is inherently ambiguous because its generating function is transcendental. Indeed, one has the following:

**Theorem 5.3** *The number of finite Sturmian words of length $n$ is*

$$1 + \sum_{i=1}^{n} \phi(i)(n-i+1)$$

*where $\phi$ is Euler's function.*

Several proofs of this result exist. See e. g. [16].

## 6 Sturmian morphisms

A morphism $f : A^* \to A^*$ is a *Sturmian morphism* if $f(\mathbf{x})$ is Sturmian for all Sturmian words. The following are known for Sturmian morphisms.

**Theorem 6.1** [17] *Every Sturmian morphism is a composition of the three morphisms*

$$E : \begin{array}{l} a \mapsto b \\ b \mapsto a \end{array} \qquad D : \begin{array}{l} a \mapsto ab \\ b \mapsto a \end{array} \qquad G : \begin{array}{l} a \mapsto ba \\ b \mapsto a \end{array}$$

*in any order and number.*

**Theorem 6.2** [2] *A morphism $f$ is Sturmian if $f(x)$ is Sturmian for some finite Sturmian word $x$.*

Morphisms that map a characteristic word to a characteristic word are a subclass:

**Theorem 6.3** [8] *Let $\mathbf{c}_\alpha$ and $\mathbf{c}_\beta$ be characteristic words. If $\mathbf{c}_\alpha = f(\mathbf{c}_\beta)$, then the morphism $f$ is a composition of $E$ and $D$.*

We call a Sturmian morphism *standard* if it is a composition of $E$ and $D$. An explicit description of standard Sturmian morphisms will be given below.

There is an interesting relation between Sturmian morphisms and automorphisms of a free group that has been discovered by Wen and Wen [26]. Denote by $F$ the free group generated by $\{a, b\}$ and let, as usual $Aut(F)$ be the group of automorphisms of $F$. It is well known the $Aut(F)$ is generated, as a group, by the three morphisms $E, G, D$ given above. Thus, any automorphism is generated by these morphisms or their inverses. Call an automorphism $\tau \in Aut(F)$ a *substitution* if $\tau(a) \in A^*$ and $\tau(b) \in A^*$. Then

**Theorem 6.4** [26] *Sturmian morphisms are exactly those automorphisms that are substitutions.*

## 7 Standard words

Consider two function $\gamma$ and $\delta$ from $A^* \times A^*$ into itself defined by

$$\gamma(u, v) = (u, uv), \qquad \delta(u, v) = (vu, v)$$

The family $\mathcal{R}$ of *standard pairs* is the smallest set of pairs of words such that

(1) $(a, b) \in \mathcal{R}$;
(2) $\mathcal{R}$ is closed under $\gamma$ and $\delta$.

The components of standard pairs are called *standard words*. Their set is denoted $S$. Observe that the two components of a standard pair always end with different letters. It is easily seen that the set $S$ of standard words is exactly the set of all words $s_n$ appearing in standard sequences. More precisely:

**Proposition 7.1** *Let $\alpha = [0; 1 + d_1, d_2, \ldots]$ be the continued fraction of an irrational $\alpha$, with $0 < \alpha < 1$, and let $(s_n)_{n \geq -1}$ be its standard sequence. Then $(s_{2n-1}, s_{2n})$ and $(s_{2n+1}, s_{2n})$ are standard pairs for $n \geq 0$, and*

$$\gamma^{d_{2n+2}}(s_{2n+1}, s_{2n}) = (s_{2n+1}, s_{2n+2}), \qquad \delta^{d_{2n+1}}(s_{2n-1}, s_{2n}) = (s_{2n+1}, s_{2n}).$$

Standard pairs and standard words have numerous properties. First, the relation between Sturmian morphisms and standard pairs is the following ([12]):

**Proposition 7.2** *Let $f : A^* \to A^*$ be a morphism. The following are equivalent:*
(*i*) *The morphism $f$ is standard (i. e. is a product of $E$ and $D$).*
(*ii*) *The set $(f(a), f(b))$ or the set $(f(b), f(a))$ is a standard pair.*
(*iii*) *The morphism $f$ preserves standard words.*
(*iv*) *The morphism $f$ preserves characteristic words.*

A full description of general Sturmian morphisms has been obtained recently by Séébold [21].

It appears ([11]) that every standard word $w$ is either a letter or of the form $w = pxy$, with $p$ a palindrome word, and $x$, $y$ distinct letters. More precisely, let $P$ be the set of palindromes over the alphabet $A$. Then

$$S = A \cup (P^2 \cap P\{ab, ba\})$$

The palindromes appearing here play a central role. Let $\Pi$ be this set. Then $S = A \cup \Pi\{ab, ba\}$. The set $\Pi$ has the following properties:

**Theorem 7.3**

(1) *The set $\Pi$ is the set of strictly bispecial factors of Sturmian words, i.e of those words $w$ such that all four words in $AwA$ are Sturmian.*

(2) *The set $\Pi$ is the set of all words $w$ having two periods $p$, $q$ which are coprime and such that $|w| = p + q - 2$.*

(3) *Every word in $a\Pi b$ is a Lyndon word.*

The first two characterization are from [11]. The last property is due to Borel and Laubie [4].

## Acknowledgements

I thank Patrice Séébold for helpful discussions during the symposium DLT'95. These greatly improved the paper.

## References

[1] P. Arnoux et G. Rauzy, Représentation géométrique de suites de complexité $2n+1$, *Bull. Soc. Math. France* **119** (1991), 199–215.

[2] J. Berstel, P. Séébold, A characterization of Sturmian morphisms, *in*: A. Borzyskowski, S. Sokolowski (eds.) *MFCS'93*, Lect. Notes Comp. Sci. **711** 1993, 281–290.

[3] E. Bombieri, J. E. Taylor, Which distributions of matter diffract? An initial investigation, *J. Phys.* **47** (1986), Colloque C3, 19–28.

[4] J.-P. Borel, F. Laubie, Quelques mots sur la droite projective réelle, *J. Théorie des nombres de Bordeaux* **5** (1993), 23–52.

[5] T. C. Brown, Descriptions of the characteristic sequence of an irrational, *Canad. Math. Bull.* **36**,1 (1993), 15–21.

[6] E. Coven and G. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973), 138–153.

[7] E. M. Coven et G. Hedlund, Sequences with minimal block growth II, *Math. Systems Theory* **8** (1973), 376–382.

[8] D. Crisp, W. Moran, A. Pollington, P. Shiue, Substitution invariant cutting sequences, *J. Théorie des nombres de Bordeaux* **5** (1993), 123–138.

[9] K. Culik II and J. Karhumäki, Iterative devices generating infinite words, *Intern. J. Algebra Comput.* **5**, 1, (1994), 69–97.

[10] K. Culik II and A. Salomaa, On infinite words obtained by iterating morphisms, *Theoret. Comput. Sci.* **19** (1982), 29–38.

[11] A. De Luca and F. Mignosi, Some combinatorial properties of Sturmian word, *Theoret. Comput. Sci.* **136** (1994), 361–385.

[12] A. De Luca, On standard Sturmian morphisms, *submitted.*

[13] S. Dulucq et D. Gouyou-Beauchamps, Sur les facteurs des suites de Sturm, *Theoret. Comput. Sci.* **71** (1990), 381–400.

[14] G. Hedlund and M. Morse, Symbolic dynamics II: Sturmian sequences, *Amer. J. Math.* **61** (1940), 1–42.

[15] W. F. Lunnon et P. A. B. Pleasants, Characterization of two-distance sequences, *J. Austral. Math. Soc. (Series A)* **53** (1992), 198–218.

[16] F. Mignosi, On the number of factors of Sturmian words, *Theoret. Comput. Sci.* **82** (1991), 71–84.

[17] F. Mignosi, P. Séébold, Morphismes sturmiens et règles de Rauzy, *J. Théorie des nombres de Bordeaux*, **5** (1993), 221–233.

[18] G. Melançon, Lyndon factorization of Sturmian words, Techn. Report, LaBRI, Université Bordeaux I, december 1995.

[19] A. Salomaa, Morphisms on free monoids and language theory, in *Formal Language Theory : Perspectives and Open Problems*, pp. 141–166, Academic Press, 1980.

[20] A. SALOMAA, *Jewels of Formal Language Theory*, Computer Science Press, 1981.

[21] P. Séébold, On the conjugation of standard morphisms, in preparation.

[22] C. Series, The geometry of Markoff numbers, *Math. Intell.* **7** (1985), 20–29.

[23] K. B. Stolarsky, Beatty sequences, continued fractions, and certain shift operators, *Cand. Math. Bull.* **19** (1976), 473–482.

[24] G. Rauzy, Mots infinis en arithmétique, in: *Automata on infinite words* (D. Perrin ed.), *Lect. Notes Comp. Sci.* **192** (1985), 165–171.

[25] B. A. Venkov, *Elementary Number Theory*, Wolters-Noordhoff, Groningen, 1970.

[26] Z.-X. Wen and Z.-Y. Wen, Local isomorphisms of invertible substitutions, *C. R. Acad. Sci. Paris* (1994).