Sturmian and Episturmian Words
(A Survey of Some Recent Results)

Jean Berstel

Institut Gaspard Monge, Université Paris-Est, Marne-la-Vallée, France

Abstract. This survey paper contains a description of some recent results concerning Sturmian and episturmian words, with particular emphasis on central words. We list fourteen characterizations of central words. We give the characterizations of Sturmian and episturmian words by lexicographic ordering, we show how the Burrows-Wheeler transform behaves on Sturmian words. We mention results on balanced episturmian words. We give a description of the compact suffix automaton of central Sturmian words.

1 Introduction

Sturmian words are combinatorial objects that are quite remarkable by the number of different characterizations they have, formulated in terms coming from different mathematical frameworks.

Sturmian words have a geometric description as digitized straight lines. Computer representation of lines has been an active subject of research, although early theory of Sturmian words remained unnoticed in the patter recognition community. The paper by [1] is a review of recognition of straight lines with respect to interaction with other disciplines. The natural generalization would be here to digitized planes, and as counter part to Sturmian bisequences.

Sturmian words have an arithmetic description, as rotations on the torus, a combinatorial description, as aperiodic words that are balanced, a description from the point of view of dynamical systems, as aperiodic words of minimal factor (subword) complexity, and so on. Many of these descriptions are known since the years 1940 and the fundamental paper [2], and a new widely disseminated research on these words has been started about thirty years ago.

In all these cases, the description given is a characterization, that is the condition stated fully describes the set of Sturmian words. Other, less known characterizations of this kind have been given. For instance, Sturmian words are characterized by the number of their return words, or by their palindromic complexity, that is the number of palindromic factors they have.

Theoretical computer scientists have contributed the point of view of effective computation. These have been studied and developed for the class of characteristic Sturmian words, where amazing computational descriptions have been provided. The special class of characteristic Sturmian words has itself some characterizations of several kinds.

S. Bozapalidis and G. Rahonis (Eds.): CAI 2007, LNCS 4728, pp. 23–47, 2007.

The richness of the theory of Sturmian words, as the meeting point of tools from different mathematical descriptions, and as extremal point of various families of infinite words, has of course led to tentatives of generalizations to other situations, especially with the objective to capture the essence of what makes the Sturmian words so special.

One of the limitation of Sturmian words is that they are over a binary alphabet. Among the extensions to larger alphabet, the so called episturmian words have appeared to be best suited family by the number of properties of Sturmian words they share.

Another extension is to two dimensions, that is to what are discretized or digital planes. This is quite interesting from the applications to pattern recognition, and is an ongoing research topic.

Another generalization is to trees. This is just at its beginnings (see [3]). Another extension is obtained when the reversal operator is replaced by an arbitrary involutory automorphism of the free monoid, see [4].

2 Sturmian and Episturmian Words

Before starting, we give some notational conventions. Given a nonempty word w, we denote by w^- the word without its last letter. If w has at least two letters, then we write $w^=$ instead of w^{--}. Thus, for instance $abaab^= = aba$.

Given a finite or infinite word w, the set of letters that occur in w is denoted by $\mathrm{Alph}(w)$. If w is infinite, $\mathrm{Ult}(w)$ denotes the set of letters that occur infinitely many often in w.

Finally, we denote by $w(k)$ the letter at position k ($k \geq 0$) in the word w.

2.1 Complexity

Let w be an infinite word on some alphabet A. We denote by $F(w)$ the set of (finite) factors of w, and by $F_n(w) = F(w) \cap A^n$ the set of factors of length n of w. The *complexity function* c_w of w is defined by

$$c_w(n) = \mathrm{Card}(F_n(w))$$

This complexity is also called *subword* or *factor* or *block complexity*. The (right) *degree* $\deg_w(x)$ of a finite word x in w is the number of letters a such that xa is a factor of w:

$$\deg_w(x) = \mathrm{Card}\{a \in A \mid xa \in F(w)\}$$

Similarly, the left degree of w is the number of $a \in A$ with $ax \in F(w)$. Clearly, $\deg_w(x) \geq 1$ for each factor x of w. Also, $\deg_w(xy) \leq \deg_w(y)$ for all x, y. Clearly

$$c_w(n+1) = \sum_{x \in F_n(w)} \deg(x)\,.$$

A factor x is right special (left special) if its degree (left degree) is strictly greater than 1. Any suffix of a right special factor is again right special. Observe that

$$c_w(n+1) - c_w(n) = \sum_{x \in S_n(w)} \deg(x) - 1\,. \tag{1}$$

where $S_n(w)$ is the set of right special factors of length n. An infinite word w is *episturmian* if the set $F(w)$ is closed under reversal, and if, for every $n \geq 1$ there exists at most one right special factor of length n. It is *aperiodic episturmian* if it is episturmian and aperiodic, that is not eventually periodic. This is equivalent to require that there is exactly one right special factor of each length. The word w is *strict episturmian* if w is aperiodic episturmian and if all its right special factors have the same degree. If this degree is k, then it follows from (1) that for $n \geq 1$

$$c_w(n) = kn + 1\,.$$

Below, we will give a more detailed description the Tribonacci word which is a strict episturmian word. The theory of episturmian words and morphisms has been developed in three basic papers [5,6,7] by Justin and Pirillo, the first with Droubay, see also [8].

Recall that an infinite word w is *recurrent* if each factor of w occurs infinitely many often in w, and it is *uniformly recurrent* if each factor occurs infinitely many often with bounded gaps between consecutive occurrences. In other terms, w is uniformly recurrent if, for every n, there exists N such that each factor of w of length N contains all factors of w of length n, in symbols $F_n(w) = F_n(u)$ for all $u \in F_N(w)$. Any episturmian word is uniformly recurrent.

Strict episturmian words are also called *Arnoux-Rauzy* words or *AR*-words. They were introduced and studied in [9], mainly in the case of three letters. Strict episturmian words over two letters are exactly the *Sturmian* words. These are aperiodic words of minimal block complexity in view of the well-known.

Theorem 1. [2,10] *An infinite word w is eventually periodic if and only if there exists an integer $n \geq 1$ such that $c_w(n) \leq n$.*

2.2 Other Complexity Functions

Several other measures of complexity of infinite words have been defined and compared to the block complexity.

The *palindrome complexity function* p_w of an infinite word w associates to each integer $n \geq 0$ the number of distinct palindromes of length n in w.

A general exposition of palindrome complexity together with new results is given in [11]. In particular, it is shown in this paper that if w is an aperiodic infinite word, then

$$p_w(n) < \frac{16}{n} c_w\Big(n + \Big\lfloor \frac{n}{4} \Big\rfloor\Big)\,.$$

Thus in particular if $c_w(n) = O(n)$ then w has bounded palindromic complexity. This holds for Sturmian and episturmian words, for automatic words, and for words that are fixed points of primitive morphisms. For uniformly recurrent

words, there is a more precise formula given in [12]. They prove that, provided the set of factors $F(w)$ is closed under reversal,

$$p_w(n) + p_w(n+1) \leq 2 + c_w(n+1) - c_w(n)\,.$$

This is sharp for Sturmian words: these are characterized by the fact that $p_w(n) = 1$ if n is even and $p_w(n) = 2$ if n is odd [13], and also for AR-words over $r > 2$ letters: these words have palindrome complexity $p_w(n) = 1$ if n is even and $p_w(n) = r$ if n is odd [14].

Another complexity function is arithmetical complexity introduced in [15]. Given an infinite word $w = w(0)w(1)\cdots$, the *arithmetical complexity* function a_w associates to $n \geq 0$ the number of distinct words of the

$$w(k)w(k+d)w(k+2d)\cdots w(k+(n-1)d)$$

for $k \geq 0$, $d \geq 1$. The arithmetical complexity of a Sturmian word depends only on its slope (see below), since two Sturmian words have the same set of factors if and only if they have same slope. So, it is convenient to write a_α instead of a_w for a Sturmian word of slope α.

Theorem 2. [16] *For any Sturmian word of slope α, one has*

$$a_\alpha(n) \leq h(n)$$

where

$$h(n) = 2 + \binom{n+1}{3} + 2\sum_{i=1}^{n-1}(n-i)\phi(i)\,.$$

Here ϕ is Euler's totient function. In fact, the authors give the exact expression for the arithmetical complexity of Sturmian words for $1/3 < \alpha < 1/2$ (note that exchanging the two letters in a Sturmian words replaces the slope α by $1-\alpha$ without changing the complexity, so the result holds also for $1/2 < \alpha < 2/3$). Denote by (r_k) the decreasing sequence of rational numbers given by $r_k = k/(3k-1)$, for $k \geq 2$. Thus $r_2 = 2/3$, $r_3 = 3/8$.

Theorem 3. [16] *For any irrational α with $1/3 < \alpha < 1/2$, one has*

$$a_\alpha(n) = \begin{cases} h(n) - 8 & \text{if } n \text{ is odd} \\ h(n) - 9 & \text{otherwise} \end{cases}$$

for $n \geq 3k$, where k is such that $r_{k-1} > \alpha > r_k$.

For other results concerning arithmetical complexity, see [17].

A more general measure is the *maximal pattern complexity*. A window τ of size k is a sequence $0 = \tau_0 < \tau_1 < \cdots < \tau_{k-1}$ of integers. The τ-pattern at position n in w is the word

$$w(n+\tau_0)w(n+\tau_1)\cdots w(n+\tau_{k-1})\,.$$

Denote by $F_\tau(w)$ the set of τ-pattern occurring in w. The τ-complexity of w is the number $c_w(\tau) = \operatorname{Card} F_\tau(w)$, and the maximal pattern complexity is

$$c_w^*(k) = \sup_{|\tau|=k} c_w(\tau)\,,$$

where $|\tau|$ denotes the size of τ. There is an analogue of Theorem 1 for the maximal pattern complexity:

Theorem 4. [18] *An infinite word w is eventually periodic if and only if $c_w^*(k) < 2k$ for some $k \geq 1$.*

Words with maximal pattern complexity $2k$ have been called pattern Sturmian words and are studied in [18]. Sturmian words are special cases of pattern Sturmian words. Generalizations are given in [19,20].

There is a variation of block complexity considered by [21,22]. Instead of counting the number of factors of given length in an infinite word, they count the number of factors of this length that occur infinitely many often in the word. If the word is uniformly recurrent, the complexities are the same. For skew words, as defined later, they are different.

2.3 Palindromic Closure

The *right palindromic closure* of a word w is the shortest palindrome which has w as a prefix. It is denoted by $w^{(+)}$. For instance, the right palindromic closure of 01011 is 0010110100. It is easy to prove that

$$w^{(+)} = uv\tilde{u}\,,$$

where v is the longest palindrome suffix of w. In the example, the longest palindrome suffix of $w = 001011$ is 11, and therefore $w^{(+)} = 0010\,11\,0100$. The notion was introduced and used by de Luca [23,24] for the analysis of finite Sturmian words.

Given a finite word d, the *right iterated palindrome* produced by d is the word $P(d)$ defined as follows. $P(\varepsilon) = \varepsilon$ and for a word d and a letter a,

$$P(da) = (P(d)a)^{(+)}\,. \tag{2}$$

For example, for the word *abbaab* one gets successively

d	$P(d)$
a	a
ab	aba
abb	$ababa$
$abba$	$ababaababa$
$abbaa$	$ababaababaababa$
$abbaab$	$ababaababaababababaababaababa$

The word d is the *directive word* of $P(d)$. A *right iterated palindrome* is a right iterated palindrome w produced by some word d. If d is over at most two letters, then the word w is *binary*.

If d is an infinite word, the *right iterated palindrome* produced by d is the infinite word which has as prefixes all right iterated palindromes produced by the finite prefixes of d. This makes sense because $P(x)$ is a prefix of $P(xy)$ for all words x, y.

If a does not occur in d, then (2) gives simply $P(da) = P(d)aP(d)^{\sim}$. There is another way to compute (2) when the letter a occurs in d. Let pa be the longest prefix of d ending with the letter a, and define the word s by $P(pa) = P(p)s$. Then $P(da) = P(d)s$. In our example, for $db = abbaab$, one has $p = ab$ and $s = baababaababa$. This computation rule is given in [25].

2.4 Justin's Formula

Justin's formula gives a useful relation between standard words and central words generated by iterated right palindromic closure. Let A be an alphabet and let $\psi : A^* \to \mathrm{End}(A^*)$ be the morphism that maps a letter a to the morphism ψ_a defined, for $b \in A$, by

$$\psi_a(b) = \begin{cases} ab & \text{if } b \neq a\,, \\ a & \text{otherwise.} \end{cases}$$

For instance, if a, b, c are letters, then

$$\psi_a(bac) = abaac\,.$$

Composition is defined for words u, v by

$$\psi_{uv} = \psi_u \circ \psi_v\,,$$

that is

$$\psi_{uv}(w) = \psi_u(\psi_v(w))\,.$$

For instance,

$$\psi_{abc}(a) = \psi_{ab}(ca) = \psi_a(bcba) = abacaba\,.$$

A word of the form $\psi_u(a)$ for some word u and some letter a is an *epistandard word*. The morphisms ψ_u are pure *epistandard morphisms*. In the binary case, these morphisms are called pure Sturmian morphisms, and the words they produce are indeed the standard words. Justin's formula establishes a relation between the morphism ψ and right palindromic closure P.

Proposition 5. (Justin's Formula) *The following holds for any words u, v:*

$$P(uv) = \psi_u(P(v))P(u)\,. \tag{3}$$

As an example, let $u = ab$, $v = ac$. Then $P(u) = aba$, $P(v) = aca$, $\psi_u(P(v)) = \psi_a(\psi_b(aca)) = \psi_a(babcba) = abaabacaba$, whereas $P(abac) = ((abaa)^{(+)}c)^{(+)} = abaabac^{(+)} = abaabacabaaba$, so indeed $P(abac) = \psi_{ab}(aca)aba$.

The formula admits several interesting special cases. First, when u is a letter, then (3) becomes

$$P(av) = \psi_a(P(v))a\,.$$

This shows that $P(av)$ is obtained from $P(v)$ by simply inserting the letter a before each letter of $P(v)$ which is not an a, and then adding a final a. For instance, since $P(ba) = bab$, one gets $P(aba) = abaaba$. Observe that $P(av)$ is also obtained from $P(v)$ by inserting the letter a *after* each non-a letter.

Another special case arises when v is just a letter. Then (3) becomes

$$P(ua) = \psi_u(a)P(u)\,. \tag{4}$$

This shows a way to compute the right palindrome closure $P(ua)$ by prefixing $P(u)$ the standard word $\psi_u(a)$. Recall that by definition $P(ua) = P(u)a\tilde{y}$, where $P(u)a = yz$ with z a maximal suffix of $P(u)a$ which is a palindrome. Since $P(u)$ and $P(ua)$ both are palindromes, one has $P(ua) = yaP(u)$ and so $\psi_u(a) = ya$.

As an example, consider the computation of $P(acbc)$. By (4), it suffices to compute $\psi_{acb}(c) = acabac$ and $P(acb) = acabaca$ to get the word

$$P(acbc) = acabacacabaca\,.$$

Finally, iteration of (4) gives, for a word $u = a_1a_2\cdots a_n$ the formula

$$P(a_1a_2\cdots a_n) = \psi_{a_1a_2\cdots a_{n-1}}(a_n)\psi_{a_1a_2\cdots a_{n-2}}(a_{n-1})\cdots\psi_{a_1a_2}(a_3)\psi_{a_1}(a_2)a_1\,.$$

For instance

$$P(acbc) = \psi_{acb}(c)\psi_{ac}(b)\psi_a(c)a = acabac\cdot acab\cdot ac\cdot a\,.$$

As an illustration of the uses of the formula, we prove the following observation.

Remark 6. A standard episturmian word w has the form $\psi_u(v)$, where u is a finite word and v is a strict standard episturmian word.

Proof. Let d be the infinite word such that $w = P(d)$. Let d' be a suffix of d such that $\mathrm{Ult}(d') = \mathrm{Alph}(d')$, and let $d = ud'$. By Justin's formula, $w = \psi_u(P(d'))$, and by construction $P(d')$ is strict.

Another remark concerns eventually periodic standard episturmian words. If w is such a word, then it is purely periodic. Indeed, by Theorem 3 in [5], one has $w = P(va^\omega)$ for some word v and some letter a, and consequently $w = \psi_v(P(a^\omega)) = \psi_v(a^\omega) = (\psi_v(a))^\omega$.

Example 7. The *Tribonacci* word is a generalization of the Fibonacci word. Finite Tribonacci words are the words t_n defined over three letters a, b, c by

$$t_{-1} = c,\ t_0 = a,\ t_1 = ab,\ t_n = t_{n-1}t_{n-2}t_{n-3}\ (n \geq 2)\,.$$

Thus

$$t_2 = abac$$
$$t_3 = abacaba$$
$$t_4 = abacabaabacab$$
$$t_5 = abacabaabacababacabaabac$$

The infinite Tribonacci word t is the limit of the words t_n. An equivalent definition of the t_n is through the morphism

$$\psi : a \mapsto ab,\ b \mapsto ac,\ c \mapsto a\,.$$

Indeed, it is easy to check that $t_n = \psi^n(a)$ for $n \geq 0$. Finally, one has also

$$t = P((abc)^\omega)$$

showing that t is a strict standard episturmian word. Indeed, denote by δ_n the prefix of length n of $(abc)^\omega$ and set $u_n = P(\delta_n)$. Then it can be shown that $u_n = t_{n-1}u_{n-1}$ for $n \geq 1$. Thus $t = \lim u_n$. Also

$$u_n = t_{n-1}t_{n-2}\cdots t_0\,.$$

This formula has been extended to more general words in [26]. For other properties of the Tribonacci word, see [27,28] and the chapter by Allouche and Berthé in [29].

3 Sturmian Words

Sturmian words have particular properties related to their geometric interpretation. This holds especially for finite Sturmian words.

3.1 Mechanical Words

Sturmian words have a geometric interpretation as *cutting sequences* of straight lines (this word comes from [30]) and therefore are closely related to digitization and pattern recognition. An equivalent formulation is through *mechanical words* (as they are called in [2]) or as *rotation words* (this is the name given for instance in [31]).

Consider a straight line in the plane. At each intersection point with the integer grid, write the letter a if the line intersects grid vertically, and write the letter b otherwise, see Figure 1. This is the definition of Sturmian words as cutting sequences. By a "shear", that is the mapping $(x, y) \mapsto (x + y, y)$, one gets the definition as "mechanical words". These are infinite words defined, for reals $0 < \alpha < 1$ and $0 \leq \rho \leq 1$, by

$$s_{\alpha,\rho}(n) = \begin{cases} a & \text{if } \lfloor (n+1)\alpha + \rho \rfloor = \lfloor n\alpha + \rho \rfloor, \\ b & \text{otherwise.} \end{cases}$$

$$s'_{\alpha,\rho}(n) = \begin{cases} a & \text{if } \lceil (n+1)\alpha + \rho \rceil = \lceil n\alpha + \rho \rceil, \\ b & \text{otherwise.} \end{cases}$$

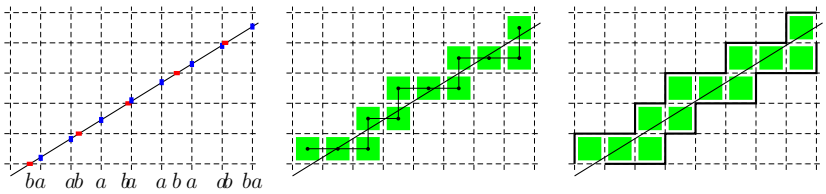


Fig. 1. A Sturmian word defined as a cutting sequence by intersection or by adjacent squares, and the upper and the lower mechanical word

for $n \geq 0$. The word $s_{\alpha,\rho}$ ($s'_{\alpha,\rho}$) is called the lower (upper) *mechanical word* with *slope* α and *intercept* ρ.

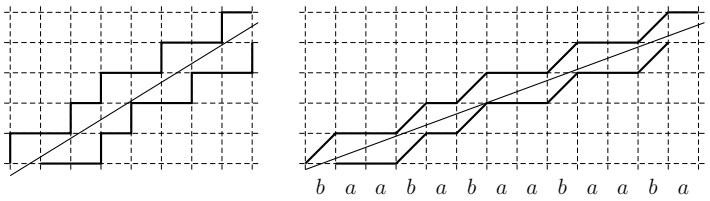


Fig. 2. "Shear" of the cutting sequence

There is an equivalent definition by rotation. Consider indeed the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ of reals modulo 1, and partition $\mathbb{T}$

$$I_a = [0, 1-\alpha),\ I_b = [1-\alpha, 1),\ I'_a = (0, 1-\alpha],\ I'_b = (1-\alpha, 1]\,,$$

and let $R_\alpha : \mathbb{T} \to \mathbb{T}$ be the rotation of angle α. Then

$$s_{\alpha,\rho}(n) = \begin{cases} a & \text{if } R^n_\alpha(\rho) \in I_a, \\ b & \text{otherwise.} \end{cases}, \quad s'_{\alpha,\rho}(n) = \begin{cases} a & \text{if } R^n_\alpha(\rho) \in I'_a, \\ b & \text{otherwise.} \end{cases}$$

This is why mechanical words are also called rotation words. They are rational words when α is rational, and irrational words when α is irrational. It is known [2] that irrational mechanical words are exactly Sturmian words. It is also known that two Sturmian words with the same slope have the same set of factors. When $\rho = \alpha$, one has $s_{\alpha,\rho} = s'_{\alpha,\rho}$. This word is called the *characteristic* word of slope α, and is denoted by c_α. For a systematic exposition, see [32] and [33].

4 Finite Sturmian Words

In this section, all words are binary over the alphabet $A = \{a, b\}$.

A finite word is *Sturmian* if it is a factor of some infinite Sturmian word. Among finite Sturmian words, particular classes are the standard words, the central words, and the Christoffel words.

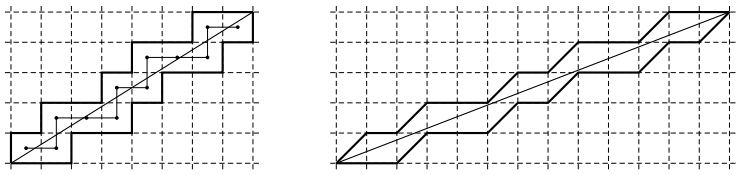


Fig. 3. The central word corresponding to the point $(8, 5)$ is $x = abaababaaba$. The upper and lower Christoffel words are $bxa = babaababaabaa$ and $axb = aabaababaabab$. Two standard words are associated with them, namely $xab = abaababaabaab$ and $xba = abaababaababa$.

The mechanical words $s_{\alpha,\rho}$ and $s'_{\alpha,\rho}$ are purely periodic when α is rational. Moreover, if $\alpha = p/(p+q)$ for $p \perp q$, then $s_{\alpha,0} = w^\omega$ and $s'_{\alpha,0} = w'^\omega$ where w and w' are precisely the lower and upper Christoffel words defined by p and q. It is easily checked that for $0 \leq n < p + q$,

$$\left\lfloor (n+1)\frac{p}{q} \right\rfloor = \left\lfloor n\frac{p}{q} \right\rfloor \iff np \bmod p+q \ < \ (n+1)p \bmod p+q\,.$$

So the lower Christoffel word is obtained simply by considering consecutive values in the sequence $np \bmod p + q$. For $p = 5$ and $q = 8$, one gets the sequence

$$0 \xrightarrow{a} 5 \xrightarrow{a} 10 \xrightarrow{b} 2 \xrightarrow{a} 7 \xrightarrow{a} 12 \xrightarrow{b} 4 \xrightarrow{a} 9 \xrightarrow{b} 1 \xrightarrow{a} 6 \xrightarrow{a} 11 \xrightarrow{b} 3 \xrightarrow{a} 8 \xrightarrow{b} 0$$

This is the construction as given by Christoffel in [34]. Another equivalent definition is by directive sequences and will be given below.

A finite word w is *balanced* if, for each pair of factors x, y of w of equal length, $\Big||x|_a - |y|_a\Big| \leq 1$ for the letter a. Here $|x|_a$ denotes the number of occurrences of a in x.

4.1 Standard and Central Words

A *directive sequence* $d = (d_0, d_1, \ldots, d_k)$ is a sequence of integers with $d_0 \geq 0$ and $d_i > 0$ for $i \geq 1$. The *standard word* produced by d is the word $S(d) = s_{k+1}$, where

$$s_{-1} = b\,,\ s_0 = a\,,\ s_{n+1} = s_n^{d_n} s_{n-1}\,,\ n \geq 0\,.$$

Example 8. For $d = (3, 1, 2, 1)$, one gets $s_1 = a^3b$, $s_2 = a^3ba$, $s_3 = a^3ba^4ba^4b$, $S(d) = s_4 = a^3ba^4ba^4ba^3ba$.

The standard word produced by the empty sequence is a, the standard word produced by (0) is b.

If $k \geq 0$, the sequences $d = (d_0, d_1, \ldots, d_k, 1)$ and $d' = (d_0, d_1, \ldots, d_k+1)$ produce the same word up to the last two letters which are interchanged, because

$$S(d) = s_k^{d_k} s_{k-1} s_k\,,\ S(d') = s_k^{d_k} s_k s_{k-1}\,,$$

and $s_{k-1}s_k$ and $s_k s_{k-1}$ are easily seen to be the same up to the last two letters, by induction.

A *central word* is a standard word without its two last letters: a word x is central if and only if $x = s^{=}$ for some standard word s.

A upper (lower) *Christoffel word* is a word of the form bxa (axb) for some central word x.

The relation between the mechanical definition and the description by the directive sequence is through the continued fraction expansion of the slope. Let again p and q be positive integers with $p \perp q$. The rational number q/p has two expansions into continued fractions, say

$$[d_0, d_1, \ldots, d_k, 1] = [d_0, d_1, \ldots, d_k + 1]\,.$$

These are the directive sequences for the two standard words with q letters a and p letters b. For example, if $q = 5$ and $p = 8$, then $q/p = [1, 1, 1, 1, 1] = [1, 1, 1, 2]$. Also, for the word $s_4 = a^3ba^4ba^4ba^3ba$ produced by the directive sequence $d = (3, 1, 2, 1)$ given above, one has $q/p = [3, 1, 2, 1]$ with $p = |s_4|_a = 4$ and $q = |s_4|_b = 15$.

Proposition 9. *Let x be a word. Then the following are equivalent*

1. *x is a central word;*
2. *xab is a standard word;*
3. *xba is a standard word;*
4. *bxa is an upper Christoffel word;*
5. *axb is a lower Christoffel word.*

As a consequence, every characterization of central words translates automatically into a characterization of standard words and of Christoffel words. In particular, we may speak about the central word produced by a directive sequence, and as mentioned above, the sequences $d = (d_0, d_1, \ldots, d_k, 1)$ and $d' = (d_0, d_1, \ldots, d_k + 1)$ produce the same central word.

4.2 Characterizations of Central Words

Proposition 10. [35] *A word x is central if and only if the words axb and bxa are conjugate.*

Proposition 11. [36] *A word is central if and only if it is a palindrome prefix of a characteristic Sturmian word.*

Proposition 12. [23] *A word is central if and only if it is a binary right iterated palindrome.*

Proposition 13. [36] *A word w is central if and only if wab or wba is a standard Sturmian word.*

Proposition 14. [36] *A word w is central if and only if it is a palindrome and wab (or wba) is a product of two palindromes.*

Proposition 15. [37] *A word w is a conjugate of a standard Sturmian word if and only if it is primitive and all its conjugates are balanced.*

Proposition 16. [37] *A word w is a conjugate of a standard Sturmian word if and only if the circular word w has $k+1$ factors of length k for $0 \leq k < |w|$, and this holds if and only if w is primitive and has $|w| - 1$ factors of length $|w| - 2$.*

Proposition 17. [36] *A word w is central if and only if the words awa, awb, bwa, bwb are balanced.*

In fact, a weaker condition is sufficient.

Proposition 18. [36] *A word w is central if and only if the words awb and bwa are balanced.*

Proposition 19. [23,38] *A word w is central if and only if it is a palindrome and the words wa and wb are balanced.*

Denote by π_w the minimal period of w. Then one has

Proposition 20. [38] *A word w is central if and only if it is a power of a letter or it is a palindrome and its prefix of length $\pi_w - 2$ is a right special factor of w.*

Example 21. Consider the word $w = baaabaaab$ has minimal period 4. Its prefix of length 2 is ba which is not a right special factor of w. So, according to Proposition 20, this word is not central. The conclusion follows also from Proposition 19, since $wb = baaabaaabb$ is not balanced.

The next proposition is actually a consequence of a result of [23].

Proposition 22. [39] *A word w is central if it is a power of a single letter or it satisfies the equation $w = w_1abw_2 = w_2baw_1$ with $w_1, w_2 \in a, b^*$. Moreover, in this latter case w_1 and w_2 are central words, $p = |w_1| + 2$ and $q = |w_2| + 2$ are co-prime periods of w and $\min p, q$ is the minimal period of w.*

Proposition 23. [36] *A word w is central if and only if there exist integers $p \perp q$ with $|w| = p + q - 2$ such that w has periods p, q.*

There is a duality between periods and number of letters in central words as already described in [23] and in [40]. Further results are in [24]. This duality has been developed recently in [41].

Proposition 24. [40] *A word w is central if and only if the word awb is a balanced Lyndon word.*

A *Sturmian palindrome* is a finite Sturmian word which is a palindrome. Every central word is a Sturmian palindrome but the converse is false. For instance, $baab$ is a Sturmian palindrome (it is a factor of the infinite Fibonacci word $f = abaab\cdots$) but is it not central in view of Proposition 18 since $bbaaba$ is not balanced. The following characterization holds.

Theorem 25. [37,23,5,38] *A word is a Sturmian palindrome if and only if it is a median factor of a central word.*

There are much more Sturmian palindromes than central words. The number of central words of length n is $\phi(n+2)$ since a central word of length n is described by two positive integers $p \perp q$ with $p+q = n+2$. On the contrary, one has

Theorem 26. [38] *Denote by $h(n)$ the number of Sturmian palindromes of length n. Then*

$$h(2n) = 1 + \sum_{i=1}^{n} \phi(2i)\,, \quad h(2n+1) = 1 + \sum_{i=1}^{n} \phi(2i+1)\,.$$

4.3 Directive Word and Directive Sequence

Given a directive sequence $d = (d_0, d_1, \ldots)$, the word $S(d)$ produced by d is a standard word if d is finite, a characteristic word if d is infinite. Define a directive word δ by $\delta = a^{d_0} b^{d_1} a^{d_2} \cdots c^{d_n}$, where $c = a$ if n is even, and $c = b$ otherwise. The relation between directive words and directive sequences in the binary case is the following.

Proposition 27. *Let d and δ be as above. Then $S(d) = \psi_\delta(\bar{c})$ where $\bar{c}$ is the opposite letter of c and moreover $S(d)^{=} = P(\delta)$.*

5 Balance

Let $\ell \geq 1$ be an integer. A set X of words over an alphabet A is *ℓ-balanced* if, for each x, y in X of equal length, $\Big||x|_a - |y|_a\Big| \leq \ell$ for all letters a. Here $|x|_a$ denotes the number of occurrences of a in x. A word is ℓ-balanced if the set of its factors is balanced. Binary balanced words are precisely 1-balanced words. A word w is *strongly balanced* if w is primitive and w^2 is balanced. A word w such that w^2 is balanced, without being necessarily primitive is called *cyclically balanced* in [42]. Thus a word is cyclically balanced if it is a power of some strongly balanced word. For instance, *abba* is balanced but is not strongly balanced because the square *abbaabba* contains both factors *aa* and *bb*. The word *ababab* is cyclically balanced. A *finite Sturmian word* is a word which is a factor of some (infinite) Sturmian word.

Proposition 28. *A finite binary word is balanced if and only if it is a finite Sturmian word.*

Proposition 29. [43,42,44] *A finite binary word is strongly balanced if and only if it is a conjugate of some standard Sturmian word.*

For infinite words, we recall the following characterization of Sturmian words.

Proposition 30. *A binary infinite word is Sturmian if and only if it is balanced and aperiodic.*

It is easy to find balanced eventually periodic words, such as ab^ω. These are not Sturmian. We discuss this in the next section.

In fact, Sturmian words share a stronger balance property. Denote by $|x|_u$ the number of distinct occurrences of the word u as a factor in the word x, counting also overlaps. For instance, $|abbabaab|_{ba} = 2$ and $|abaababa|_{aba} = 3$. Then, one has

Theorem 31. [45] *A binary infinite word w is Sturmian if and only if for each word u,*

$$\Big||x|_u - |y|_u\Big| \leq |u|$$

for each pair of factors x, y of the same length of w,

A characterization of episturmian words by a balance property like Proposition 30 does not exist. It is known that the Tribonacci word t is 2-balanced. However, when applying a well chosen pure epistandard morphism, it does not remain 2-balanced. For instance, the word $\mu(t)$ with $\mu = \psi_{aabbac}$, contains the factors *baabaaabaabaabaaabaab* and *aacaabaabaaabaabaacaa* of length 21. Indeed, the first is a factor of $\mu(bab)$ and the second is a factor of $\mu(aca)$. The number of b in these factors are 7 and 4, so their balance is 3. It has been proved by [46] that there exist AR-sequences which not ℓ-balanced for any ℓ.

There is a closed formula for the number of finite balanced words, that is of factors of Sturmian words.

Proposition 32. *The number of balanced binary words of length n is*

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

where ϕ is the Euler's totient function.

The first proof of this formula is perhaps [47]. Other proofs are in [48,36,23,49,50]. Related results also appear in [51,52]. An exact formula for the number $g_\ell(n)$ of ℓ-balance words of length n seems not to be known. It was shown already in [47] that it is exponential for $\ell \geq 2$ (whereas usual number theory shows that $g_1(n) = N^3/\pi^2 + O(n^2)$) and more exactly that

$$g_\ell(n) = \Theta\binom{\ell + 1}{\lfloor \ell/2 \rfloor}^{n/(\ell+1)}$$

which gives $g_2(n) = \Theta(3^{n/3})$. Heinis provided independently in [53] a lower bound, and Tarannikov [54] shows that

$$g_\ell(n) = \Theta(n^2(2\cos\frac{\pi}{\ell + 2})^n)$$

which is better for $\ell \geq 3$.

On the other hand, the number of factors of length n of strict episturmian words (or equivalently of Arnoux-Rauzy words) has been considered. A *bispecial factor* is a word that is both a left and a right special factor of the same Arnoux-Rauzy word.

Proposition 33. [55] *The number of factors of length n of strict episturmian words over a k-letter alphabet is*

$$k + (n-1)k(k-1) + (k-1)^2 \sum_{i=1}^{n-2} (n-i-1)b(i)$$

where $b(m)$ is the number of bispecial factors of length m of Arnoux-Rauzy words.

The number of bispecial factors is evaluated, in [55], in terms of a generalized Euclidean algorithm.

We already mentioned that episturmian words are not balanced in general. In fact, almost the opposite is true: episturmian words are never balanced, except in simple cases. More precisely, the following holds.

Theorem 34. [56] *Let x be a standard episturmian word over the alphabet $A = \{1, 2, \ldots, k\}$ with $k \geq 3$. Then x is balanced if and only if its directive word δ can be written in one of the following forms, up to a permutation of the alphabet.*

1. $123 \cdots k1^\omega$
2. $1^n 23 \cdots (k-1)k^\omega$ *for some* $n \geq 1$
3. $12 \cdots \ell 1(\ell+1) \cdots (k-1)k^\omega$ *for some* $1 \leq \ell < k$.

For $k = 5$, an example of the last case is 121345^ω with $\ell = 2$. All episturmian words of the theorem are eventually periodic.

6 Lexicographic Ordering

Every (total) order on an alphabet A defines a lexicographic order on (right) infinite words. Given an infinite word x, we denote by $\min(x)$ and by $\max(x)$ the minimal and the maximal word, for the lexicographic order, of the orbit of x. This is simply defined by the condition that, for each integer n, the prefix of length n of $\min(x)$ (of $\max(x)$) is the smallest (largest) word in $F_n(x)$. For Sturmian words, it is easily seen that $s_{\alpha,\rho} < s_{\alpha,\rho'}$ if and only if $\rho < \rho'$ (recall that α is irrational). Thus for the ordering $a < b$, one gets that $\min(s_{\alpha,\rho}) = s_{\alpha,0} = ac_\alpha$ and $\max(s_{\alpha,\rho}) = bc_\alpha$, where c_α denotes the characteristic word with slope α. As an example, consider the Fibonacci word $f = abaababaabaab \cdots$. Then $\min(f) = af$ and $\max(f) = bf$.

The comparison of words for the lexicographic order is well suited for the study of balanced infinite words, and can be extended to the case of more than two letters. It will be convenient to use the following old terminology from [2]. A *Sturmian trajectory* is an infinite binary word whose (finite) factors are finite Sturmian words. Thus, Sturmian trajectories are precisely balanced binary words.

Similarly, we call *episturmian trajectory* an infinite word whose finite factors are finite episturmian words. Episturmian trajectories are called episturmian words *in the wide sense* in [57].

It is known since [2] that Sturmian trajectories can be partitioned into three classes:

1. aperiodic words: these are exactly all Sturmian words or equivalently all irrational mechanical words;
2. (purely) periodic words : these are the rational mechanical words; they are of the form w^ω, where w is a conjugate of some standard word;
3. eventually periodic but not purely periodic words. These are called *skew* words. They are not mechanical words. It has been shown that they are those suffixes of the words of the form $\mu(a^n ba^\omega)$, for some pure standard Sturmian morphism μ and some integer $n \geq 0$ which are not suffixes of $\mu(a^\omega)$.

The three classes of Sturmian trajectories can be grouped together in three manners. First, group (1) is compose of aperiodic words whereas groups (2)+(3) are eventually periodic words. Next, words in (1) + (2) are uniformly recurrent, whereas words of type (3) are not recurrent. Finally, words of type (1) + (3) are precisely the words called *fine* by Pirillo in [58] and that we will describe in a moment. First we give the following characterization.

Theorem 35. *A binary infinite word x over $\{a, b\}$ with $a < b$ is a Sturmian trajectory if and only if there is an infinite y such that $ay \leq \min(x)$ and $\max(x) \leq by$.*

This is a corollary of the next theorem, and appears also, under a different guise, in [59]. We denote by $\min(A)$ the smallest letter in the alphabet A for the given order.

Theorem 36. [57] *An infinite word x over A is an episturmian trajectory if and only if there exists an infinite word y such that $\min(A)y \leq \min(x)$ for every order over A.*

Episturmian trajectories are either episturmian words or belong to the family of so-called *episkew* words. These are exactly the episturmian trajectories which are not recurrent. It is quite interesting to note that the characterization of skew Sturmian trajectories carries over, with some complications, to episkew words. This is done in [60], see also [57].

Proposition 37. *An infinite word x with $A = \mathrm{Alph}(x)$ is episkew if and only if there is a letter a, a standard episturmian word y on $B = A \setminus \{a\}$, a finite prefix p of y and a pure epistandard morphism μ such that $zx = \mu(\tilde{p}ay)$ for some proper prefix z of $\mu(\tilde{p}a)$.*

If, in the proposition, the word y is strict, then the word x itself is called strict episkew. Observe also that in the Sturmian case the word $\tilde{p}ay$ indeed reduces to a word of the form $a^p ba^\omega$.

In the case of characteristic words or of epistandard words, one has stronger conditions.

Theorem 38. [61] *A binary word x over $A = \{a, b\}$ with $a < b$ is a characteristic Sturmian word if and only if $ax = \min(x)$ and $\max(x) = bx$.*

A result similar to Theorem 38 holds for strict episturmian (or Arnoux-Rauzy) words.

Proposition 39. [6] *An infinite word x over some alphabet A is a strict epistandard word if and only if* $\min(A)x = \min(x)$ *for any order on A.*

This is related to the following.

Proposition 40. [61] *An infinite word x over some alphabet A is an epistandard word if and only if* $\min(A)x \leq \min(x)$ *for any order on A.*

Let x be an infinite word and let $A = \mathrm{Alph}(x)$. The word x is *fine* if there exists an infinite word y such that $\min(x) = \min(A)y$ holds for any lexicographic order. As announced, we have the following.

Proposition 41. [58] *A binary word is fine if and only if it is a Sturmian word or a skew Sturmian word.*

Thus, the Sturmian trajectories which are not fine are precisely the rational mechanical words. This has been extended to episturmian trajectories.

Proposition 42. [60] *A word x is fine if and only if it is a strict epistandard word or a strict skew episturmian word.*

7 Burrows-Wheeler Transformation

The Burrows-Wheeler transformation, introduced in [62], is a reversible transformation that produces a permutation $BWT(w)$ of an input sequence w. It appears that the transform is easier to compress than the original sequence because there is some clustering effect in the transformed word. BWT is used in the BZIP2 data compression algorithm. The Burrows-Wheeler transformation has a strong relation to a transformation called the Gessel-Reutenauer transform, introduced in [63]. This connection has been described in [64]. As has been shown in [65] the Burrows-Wheeler transformation takes a very particular form when applied to standard Sturmian words. Recent results are given in the forthcoming paper [66].

The Burrows-Wheeler transformation takes as input a word w, and produces as output a permutation $BWT(w)$, obtained as follows. Let $M(w)$ be the matrix composed of all conjugates of w, ordered lexicographically. Then $BWT(w)$ is the last column of $M(w)$.

Example 43. For the input word $w = abraca$, the matrix is

$$M(w) = \begin{bmatrix} a & a & b & r & a & c \\ a & b & r & a & c & a \\ a & c & a & a & b & r \\ b & r & a & c & a & a \\ c & a & a & b & r & a \\ r & a & c & a & a & b \end{bmatrix}$$

and the output is the last column, that is $BWT(w) = caraab$.

Clearly, two words u and v are conjugate if and only if $M(u) = M(v)$. In particular, $BWT(u) = BWT(v)$. In order to make the transformation injective, the position of the input word in the matrix is added to the transform. If $u = v^m$ for some integer m and $BWT(v) = a_0a_1 \cdots a_{n-1}$, the $BWT(u) = a_0^m a_1^m \cdots a_{n-1}^m$. In fact, the matrix $M(u)$ has every row repeated m times and every column duplicated m times.

The Burrows-Wheeler Transform is reversible: given $x = BWT(w)$ and an index i, it is possible to recover w. To do this, one first recovers the first column of $M(w)$ by ordering lexicographically the letters of the word $BWT(w)$. Next, one defines a permutation τ on the set $\{0, \ldots, n-1\}$ that maps a position in the first column of $M(w)$ to the corresponding position in x. This permutation gives the word w, when started in the position i.

Example 44. Consider $x = caraab$, and let us compute w such that $BWT(w) = x$. The matrix $M(w)$ has the form

$$M(w) = \begin{bmatrix} a & \cdot & \cdot & \cdot & \cdot & c \\ a & \cdot & \cdot & \cdot & \cdot & a \\ a & \cdot & \cdot & \cdot & \cdot & r \\ b & \cdot & \cdot & \cdot & \cdot & a \\ c & \cdot & \cdot & \cdot & \cdot & a \\ r & \cdot & \cdot & \cdot & \cdot & b \end{bmatrix}$$

The correspondence τ is

$$\tau = \begin{pmatrix} 0\,1\,2\,3\,4\,5 \\ 1\,3\,4\,5\,0\,2 \end{pmatrix} = \big(1\,3\,5\,4\,0\big)$$

Thus

$$\begin{array}{r} 1\,3\,5\,2\,4\,0 \\ w = a\;b\;r\;a\;c\;a \end{array}$$

The main observation concerning the relation with Sturmian words is the following remarkable theorem. Recall that a binary word is strongly balanced if and only if it is a conjugate of a standard Sturmian word.

Theorem 45. [66] *A word w over $\{a, b\}$, with $a < b$ is the power of a strongly balanced word if and only if its Burrows-Wheeler Transform is of the form $b^q a^p$. Moreover, in the matrix $M(w)$, each row is obtained from the preceding by replacing a factor ab by a factor ba, and all columns also are conjugates.*

Example 46. Consider the strongly balanced word *abaabab*. The matrix is

$$M(abaabab) = \begin{bmatrix} a & a & b & a & b & a & b \\ a & b & a & a & b & a & b \\ a & b & a & b & a & a & b \\ a & b & a & b & a & b & a \\ b & a & a & b & a & b & a \\ b & a & b & a & a & b & a \\ b & a & b & a & b & a & a \end{bmatrix} \tag{5}$$

Observe that the first (last) row is the lower (upper) Christoffel word, and these rows are composed of the central word bordered by a, b and b, a respectively.

The matrix $M(w)$ defined for the Burrows-Wheeler transform has also been considered in [43] in the process of giving characterizations of strongly balanced binary words. Denote by $P(w)$ the matrix of partial sums of $M(w)$ where $P(w)_{i,j}$ is defined to be the number of b in the prefix of length j of the ith row in $M(w)$. For instance, the matrix in (5) has the matrix of partial sums.

$$P(abaabab) = \begin{bmatrix} 0\,0\,1\,1\,2\,2\,3 \\ 0\,1\,1\,1\,2\,2\,3 \\ 0\,1\,1\,2\,2\,2\,3 \\ 0\,1\,1\,2\,2\,3\,3 \\ 1\,1\,1\,2\,2\,3\,3 \\ 1\,1\,2\,2\,2\,3\,3 \\ 1\,1\,2\,2\,3\,3\,3 \end{bmatrix}$$

They prove the following

Theorem 47. [43] *A word w over $\{a, b\}$, with $a < b$ is the power of a strongly balanced word if and only if every column in the matrix of partial sum is increasing when read from top to bottom.*

Let us mention briefly the connection of the Burrows-Wheeler and the Gessel-Reutenauer transformation [63]. The Burrows-Wheeler transformation is the inverse of the Gessel-Reutenauer transformation. Define the *standardization* associated to a word $w = a_1 \cdots a_n$ over an ordered alphabet A as the permutation σ given by

$$\sigma(i) < \sigma(j) \quad \text{iff} \quad a_i < a_j \text{ or } (a_i = a_j \text{ and } i < j)$$

Example 48. Consider the word *ccbbbcacaaabba*. After a lexicographic sort, the symbols a are at positions 1–5, symbols b at positions 6–10. The symbols c appear in position 11 to 14. This gives the permutation.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ c & c & b & b & b & c & a & c & a & a & a & b & b & a \\ 11 & 12 & 6 & 7 & 8 & 13 & 1 & 14 & 2 & 3 & 4 & 9 & 10 & 5 \end{pmatrix}$$

After cycle decomposition, one gets

$$\begin{matrix} (1\ 11\ 4\ 7)\ (2\ 12\ 9)\ (3\ 6\ 13\ 10)\ (5\ 8\ 14) \\ c\ a\ b\ a \quad c\ b\ a \quad b\ c\ b\ a \quad b\ c\ a \end{matrix}$$

The result is $(caba)(cba)(bcba)(bca)$

Theorem 49. [63] *The standardization σ induces a bijection between all words over A and the family of multisets of conjugacy classes of primitive words over A.*

Define a new order on finite order on words by

$$u \preceq v \quad \text{if and only if} \quad u^\omega < v^\omega \text{ or } (u^\omega = v^\omega \text{ and } |u| \leq |v|)$$

For example, $aba \prec ab$ because $abaaba\cdots < ababab\cdots$. Recovering the word w from its decomposition S into conjugacy classes is done as follows: One sorts the conjugates of words in S by $\prec$. Then the word w is the sequence of last letters in this table.

Example 50. Consider the set $S = \{caba, bcba, bca, cba\}$. The conjugates of all words in S are ordered with respect to the new order $\prec$. This gives the sequence $(abac, abc, abcb, acab, acb, babc, baca, bac, bca, bcba, caba, cab, cbab, cba)$. The word composed of the last letters of the words in this sequence is *ccbbbcacaaabba*.

Conversely, to get the decomposition S from w, one sorts the word w alphabetically, then computes the letter-correspondence permutation and then outputs the permutation in cycle form, and computes the multiset.

Example 51. Starting with *ccbbbcacaaabba*, one gets the table

$$\begin{pmatrix} a & a & a & a & a & b & b & b & b & b & c & c & c & c \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 7 & 9 & 10 & 11 & 14 & 3 & 4 & 5 & 12 & 13 & 1 & 2 & 6 & 8 \\ c & c & b & b & b & c & a & c & a & a & a & b & b & a \end{pmatrix}$$

In cycle form, one gets

$$\begin{matrix} (1\ 7\ 4\ 11) & (2\ 9\ 12) & (3\ 10\ 13\ 6) & (5\ 14\ 8) \\ a\ b\ a\ c & a\ b\ c & a\ b\ c\ b & a\ c\ b \end{matrix}$$

The output is the set $S = \{caba, bca, bcba, cba\}$.

8 Sturmian Graphs

Given a standard or a central Sturmian word it appears interesting to consider, in this special case, some well known constructs, such as the (compacted) suffix tree or the suffix automaton (also called DAWG for directed acyclic word graph). A compacted version of the minimal suffix automaton has been considered by [67] for the Fibonacci word and by [68] for arbitrary central words.

The CDWAG (compact directed acyclic word graph) $G(w)$ of a word w is the minimal automaton recognizing the set of suffixes of w, after removing non-final states with out-degree 1.

The terminology DWAG stems from [69]. See also [70].

Example 52. For $w = abaababaaba$, the automaton $G(w)$ (all states are final) is given in Figure 4.

Any CDAWG is homogeneous, that is all edges leading to a state have the same label. For the description of the method of construction, we use $u[d]$ to

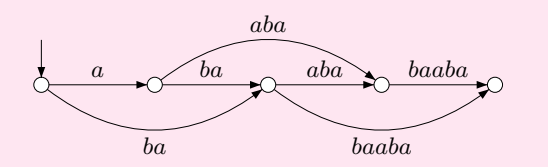


Fig. 4. The automaton $G(abaababaaba)$

denote the reversal of the standard word with directive sequence d. Thus $u[21] = abaa$ because the standard word produced by $(2, 1)$ is $aaba$. We write $c[d]$ for the central word produced by the directive sequence d. We use the identity

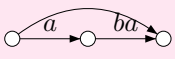
$$c[d_0 d_1 \cdots d_n 1] = u[\varepsilon]^{d_0} u[d_0]^{d_1} u[d_0 d_1]^{d_2} \cdots u[d_0 d_1 \cdots d_{n-1}]^{d_n} .$$

The CDAWG of a central word c with directive sequence d is constructed by induction. The method goes as follows. Set $d = d'\delta 1$.

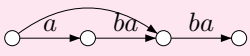
1. if $\delta \neq 1$, repeat the last edge of the graph of $d'(\delta - 1)1$.
2. otherwise (that is d ends with 11), set $d = d''\delta' 11$, take the graph of $d'1$, add a new state and $1 + \delta'$ edges to this state. The common label of these fresh edges is $u[d'']$.

Example 53. In order to compute the graph of 12311, we start with $d = 11$, $c[11] = a$, and the graph

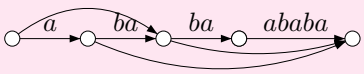
Then, using the second rule with $d = 111$, $c[111] = a|ba$, we get



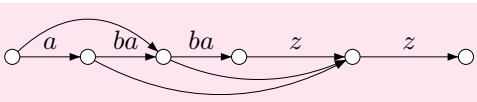
Now the first rule is applied for $d = 121$, $c[121] = a|ba|ba$. This gives



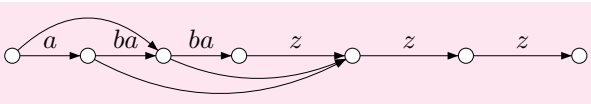
For $d = 1211$ and $c[1211] = a|ba|ba|ababa$, the second rule gives



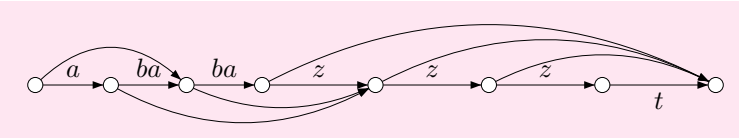
For $d = 1221$, and setting $z = ababa$, one gets $c[1221] = a|ba|ba|z|z$ and the graph is



For $d = 1231$, one has $c[1231] = a|ba|ba|z|z|z$ and



Finally, for $d = 12311$, one gets $c[12311] = a|ba|ba|z|z|z|t$ with $t = bazzz$ and the graph



The length of the central word c defined by $d = (d_0, d_1, \ldots, d_k)$ is $|\ell_k| - 2$, where $\ell_n = |s_n| - 2$ and $\ell_{-1} = \ell_0 = 1$, $\ell_{n+1} = d_n\ell_n + \ell_{n-1}$. Let $H(c)$ be the graph obtained from the $G(c)$ by replacing each label by its length. Then $H(c)$ *counts from* 0 *to* $|c|$ in the following sense: each integer h with $0 \leq h \leq |c|$ is the sum of the weights of exactly one path in $H(c)$ starting at the initial state. In

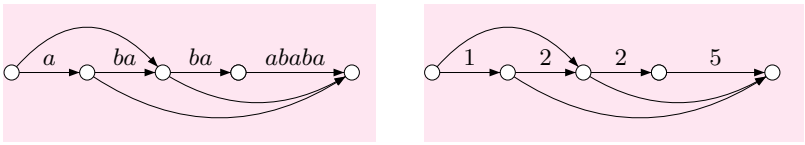


Fig. 5. The CDAWG for 1211 and the corresponding counting graph

other words, the set of weights in $H(c)$ is complete and unambiguous base for representing integer up to $|c|$, provided the representation is a path in the graph. For example, the graph on the right of Figure 5 counts up to 10.

Problem 54. What is the minimal size of a graph with out-degree at most 2 counting from 0 to n?

If the size of the labels increase exponentially, like for the Fibonacci word, then the size is $O(\log n)$. It is conjectured that the bound $O(\log n)$ always holds. This is related to the following number-theoretic conjecture (see [68] for details).

Conjecture 55 (Zaremba). There exists an integer K such that for all positive m, there exists some $i \perp m$, $i < m$ such that all partial quotients in the continued fraction expansions of i/m are bounded by K.

Acknowledgments

I thank Amy Glen, Aldo de Luca for their helpful comments, and Alessandro De Luca for sending me several preprints.

References

1. Klette, R., Rosenfeld, A.: Digital straightness—a review. Discrete Appl. Math. 139, 197–230 (2004)
2. Morse, M., Hedlund, G.A.: Symbolic dynamics II. Sturmian trajectories. Amer. J. Math. 62, 1–42 (1940)

3. Berstel, J., Boasson, L., Carton, O., Fagnot, I.: A first investigation of Sturmian trees. In: Thomas, W., Weil, P. (eds.) STACS 2007. LNCS, vol. 4393, pp. 73–84. Springer, Heidelberg (2007)
4. de Luca, A., De Luca, A.: Pseudopalindrome closure operators in free monoids. Theoret. Comput. Sci. 362(1-3), 282–300 (2006)
5. Droubay, X., Justin, J., Pirillo, G.: Epi-Sturmian words and some constructions of de Luca and Rauzy. Theoret. Comput. Sci. 255(1-2), 539–553 (2001)
6. Justin, J., Pirillo, G.: On a characteristic property of Arnoux-Rauzy sequences. Theor. Inform. Appl. 36(4), 385–388 (2002)
7. Justin, J., Pirillo, G.: Episturmian words and episturmian morphisms. Theoret. Comput. Sci. 276(1-2), 281–313 (2002)
8. Justin, J.: Episturmian words and morphisms (results and conjectures). In: Crapo, H., Senato, D. (eds.) Algebraic Combinatorics and Computer Science, pp. 533–539. Springer, Heidelberg (2001)
9. Arnoux, P., Rauzy, G.: Représentation géométrique de suites de complexité $2n+1$. Bull. Soc. Math. France 119, 199–215 (1991)
10. Coven, E.M., Hedlund, G.A.: Sequences with minimal block growth. Math. Systems Theory 7, 138–153 (1973)
11. Allouche, J.P., Baake, M., Cassaigne, J., Damanik, D.: Palindrome complexity. Theoret. Comput. Sci. 292(1), 9–31 (2003)
12. Baláži, P., Masáková, Z., Pelantová, E.: Factor versus palindromic complexity of uniformly recurrent infinite words. Theoret. Comput. Sci. 380, 266–275 (2007)
13. Droubay, X., Pirillo, G.: Palindromes and Sturmian words. Theoret. Comput. Sci. 223(1-2), 73–85 (1999)
14. Damanik, D., Zamboni, L.Q.: Combinatorial properties of Arnoux-Rauzy subshifts and applications to Schrödinger operators. Rev. Math. Phys. 15(7), 745–763 (2003)
15. Avgustinovich, S.V., Fon-Der-Flaas, D.G., Frid, A.E.: Arithmetical complexity of infinite words. In: Words, Languages and Combinatorics. Proc. 3rd Conf. Words, Languages and Combinatorics, Kyoto, March 2000, vol. III, pp. 51–62. World Scientific, Singapore (2003)
16. Cassaigne, J., Frid, A.E.: On the arithmetical complexity of Sturmian words. Theoret. Comput. Sci. 380, 304–316 (2007)
17. Avgustinovich, S.V., Cassaigne, J., Frid, A.E.: Sequences of low arithmetical complexity. Theor. Inform. Appl. 40(4), 569–582 (2006)
18. Kamae, T., Zamboni, L.Q.: Maximal pattern complexity for discrete systems. Ergodic Theory Dynam. Systems 22(4), 1201–1214 (2002)
19. Kamae, T., Rao, H., Tan, B., Xue, Y.M.: Language structure of pattern Sturmian words. Discrete Math. 306(15), 1651–1668 (2006)
20. Kamae, T., Rao, H.: Maximal pattern complexity of words over l letters. European J. Combin. 27(1), 125–137 (2006)
21. Nakashima, I., Tamura, J.I., Yasutomi, S.I.: Modified complexity and $*$-Sturmian word. Proc. Japan Acad. Ser. A Math. Sci. 75(3), 26–28 (1999)
22. Nakashima, I., Tamura, J.I., Yasutomi, S.I.: $*$-Sturmian words and complexity. J. Theor. Nombres Bordeaux 15(3), 767–804 (2003)
23. de Luca, A.: Sturmian words: structure, combinatorics, and their arithmetics. Theoret. Comput. Sci. 183(1), 45–82 (1997)
24. de Luca, A.: Combinatorics of standard Sturmian words. In: Mycielski, J., Rozenberg, G., Salomaa, A. (eds.) Structures in Logic and Computer Science. LNCS, vol. 1261, pp. 249–267. Springer, Heidelberg (1997)
25. Risley, R., Zamboni, L.Q.: A generalization of Sturmian sequences: combinatorial structure and transcendence. Acta Arith. 95, 167–184 (2000)

26. Glen, A.: Powers in a class of ⊣-strict standard episturmian words. Theoret. Comput. Sci. 380, 330–354 (2007)
27. Tan, B., Wen, Z.Y.: Some properties of the Tribonacci sequence. European J. Combin. (2007)
28. Chekhova, N., Hubert, P., Messaoudi, A.: Propriétés combinatoires, ergodiques et arithmétiques de la substitution de Tribonacci. J. Theor. Nombres Bordeaux 13(2), 371–394 (2001)
29. Lothaire, M.: Applied Combinatorics on Words. Encyclopedia of Mathematics and its Applications, vol. 105. Cambridge University Press, Cambridge (2005)
30. Series, C.: The geometry of Markoff numbers. Math. Intelligencer 7(3), 20–29 (1985)
31. Berthé, V., Ei, H., Ito, S., Rao, H.: Invertible substitutions and Sturmian words: an application to Rauzy fractals. Theor. Inform. Appl. (to appear, 2007)
32. Lothaire, M.: Algebraic Combinatorics on Words. Encyclopedia of Mathematics and its Applications, vol. 90. Cambridge University Press, Cambridge (2002)
33. Pytheas Fogg, N.: Substitutions in dynamics, arithmetics and combinatorics. In: Berthé, V., Ferenczi, S., Mauduit, C., Siegel, A. (eds.) Lecture Notes in Mathematics, vol. 1794, Springer, Heidelberg (2002)
34. Christoffel, E.B.: Observatio arithmetica. Annali di Mathematica 6, 145–152 (1875)
35. Pirillo, G.: A curious characteristic property of standard Sturmian words. In: Crapo, H., Senato, D. (eds.) Algebraic Combinatorics and Computer Science. A tribute to Gian-Carlo Rota., pp. 541–546. Springer, Heidelberg (2001)
36. de Luca, A., Mignosi, F.: Some combinatorial properties of Sturmian words. Theoret. Comput. Sci. 136(2), 361–385 (1994)
37. Borel, J.P., Reutenauer, C.: Palindromic factors of billiard words. Theoret. Comput. Sci. 340(2), 334–348 (2005)
38. de Luca, A., De Luca, A.: Combinatorial properties of Sturmian palindromes. Internat. J. Found. Comput. Sci. 17(3), 557–573 (2006)
39. Carpi, A., de Luca, A.: Codes of central Sturmian words. Theoret. Comput. Sci. 340(2), 220–239 (2005)
40. Berstel, J., de Luca, A.: Sturmian words, Lyndon words and trees. Theoret. Comput. Sci. 178(1-2), 171–203 (1997)
41. Berthé, V., de Luca, A., Reutenauer, C.: On an involution of Christoffel words and Sturmian morphisms. In: European J. Combinatorics (in press, 2007)
42. Chuan, W.F.: Moments of conjugacy classes of binary words. Theoret. Comput. Sci. 310(1-3), 273–285 (2004)
43. Jenkinson, O., Zamboni, L.Q.: Characterisations of balanced words via orderings. Theoret. Comput. Sci. 310(1-3), 247–271 (2004)
44. de Luca, A., De Luca, A.: Some characterizations of finite Sturmian words. Theoret. Comput. Sci. 356(1-2), 118–125 (2006)
45. Fagnot, I., Vuillon, L.: Generalized balances in Sturmian words. Discrete Appl. Math. 121(1-3), 83–101 (2002)
46. Cassaigne, J., Ferenczi, S., Zamboni, L.Q.: Imbalances in Arnoux-Rauzy sequences. Ann. Inst. Fourier (Grenoble) 50(4), 1265–1276 (2000)
47. Lipatov, E.P.: A classification of binary collections and properties of homogeneity classes. Problemy Kibernet 39, 67–84 (1982)
48. Mignosi, F.: On the number of factors of Sturmian words. Theoret. Comput. Sci. 82(1), 71–84 (1991)
49. Berstel, J., Pocchiola, M.: A geometric proof of the enumeration formula for Sturmian words. Internat. J. Algebra Comput. 3(3), 349–355 (1993)

50. Berstel, J., Pocchiola, M.: Random generation of finite Sturmian words. In: Proceedings of the 5th Conference on Formal Power Series and Algebraic Combinatorics (Florence, 1993), vol. 153, pp. 29–39 (1996)
51. Berenstein, C.A., Lavine, D.: On the number of digital straight line segments. IEEE Trans. Pattern Anal. Mach. Intell. 10(6), 880–887 (1988)
52. Koplowitz, J., Lindenbaum, M., Bruckstein, A.M.: The number of digital straight lines on an $n \times n$ grid. IEEE Transactions on Information Theory 36(1), 192–197 (1990)
53. Heinis, A.: On low-complexity bi-infinite words and their factors. J. Theor. Nombres Bordeaux 13(2), 421–442 (2001)
54. Tarannikov, Y.: On the bounds for the number of ℓ-balanced words. Technical report, Mech. & Math. Department, Moscow State University (2007)
55. Mignosi, F., Zamboni, L.Q.: On the number of Arnoux-Rauzy words. Boolean Calculus of Differences 101(2), 121–129 (2002)
56. Paquin, G., Vuillon, L.: A characterization of balanced episturmian sequences. Electronic J. Combinatorics 14(1) R33, pages 12 (2007)
57. Glen, A., Justin, J., Pirillo, G.: Characterizations of finite and infinite episturmian words via lexicographic orderings. European Journal of Combinatorics (2007)
58. Pirillo, G.: Morse and Hedlund's skew Sturmian words revisited. Annals Combinatorics (to appear, 2007)
59. Gan, S.: Sturmian sequences and the lexicographic world. Proc. Amer. Math. Soc. 129, electronic, 1445–1451 (2001)
60. Glen, A.: A characterization of fine words over a finite alphabet. Theoret. Comput. Sci. CANT conference, Liege, Belgium, May 8-19, 2007, 8–19 (to appear, 2007)
61. Pirillo, G.: Inequalities characterizing standard Sturmian and episturmian words. Theoret. Comput. Sci. 341, 276–292 (2005)
62. Burrows, M., Wheeler, D.J.: A block sorting data compression algorithm. Technical report, Digital System Research Center (1994)
63. Gessel, I., Reutenauer, C.: Counting permutations with given cycle structure and descent set. J. Comb. Theory A 64, 189–215 (1993)
64. Crochemore, M., Désarménien, J., Perrin, D.: A note on the Burrows-Wheeler transformation. Theoret. Comput. Sci. 332, 567–572 (2005)
65. Mantaci, S., Restivo, A., Sciortino, M.: Burrows Wheeler transform and Sturmian words. Inform. Proc. Letters 86, 241–246 (2003)
66. Mantaci, S., Restivo, A., Rosone, G., Sciortino, M.: An extension of the Burrows Wheeler transform. Theoret. Comput. Sci. (2007)
67. Rytter, W.: The structure of subword graphs and suffix trees of Fibonacci words. Theoret. Comput. Sci. 363(2), 211–223 (2006)
68. Epifanio, C., Mignosi, F., Shallit, J., Venturini, I.: On Sturmian graphs. Discrete Appl. Math 155, 1014–1030 (2007)
69. Blumer, A., Blumer, J.A., Haussler, D., Ehrenfeucht, A., Chen, M.T., Seiferas, J.I.: The smallest automaton recognizing the subwords of a text. Theoret. Comput. Sci. 40(1), 31–55 (1985) (Special issue: Eleventh international colloquium on automata, languages and programming, Antwerp, (1984)
70. Crochemore, M., Rytter, W.: Jewels of stringology. World Scientific Publishing Co. Inc, River Edge, NJ (2003)