International Journal of Algebra and Computation, Vol. 12, Nos. 1 & 2 (2002) 371–385

RECENT RESULTS ON EXTENSIONS OF STURMIAN WORDS

JEAN BERSTEL

Institut Gaspard Monge (IGM), Université Marne-la-Vallée
5, Boulevard Descartes, 77454 Marne-la-Vallée Cédex 2, France

Communicated by J. Meakin
Received 30 November 2000
Revised 30 January 2001

AMS Mathematics Subject Classification: Primary 68Q45; Secondary 68Q45, 11B85

Sturmian words are infinite words over a two-letter alphabet that admit a great number of equivalent definitions. Most of them have been given in the past ten years. Among several extensions of Sturmian words to larger alphabets, the Arnoux–Rauzy words appear to share many of the properties of Sturmian words. In this survey, combinatorial properties of these two families are considered and compared.

1. Introduction

Sturmian words are infinite words over a two-letter alphabet that have exactly $n+1$ factors of length n for each $n \geq 0$. It appears that these words admit several equivalent combinatorial definitions. They can also be described explicitly by an arithmetic construction, providing a bridge between combinatorics and number theory. Moreover, the definition by factors makes Sturmian words define symbolic dynamical systems. The first detailed investigations of these words were done from this point of view [37]. Their numerous properties and equivalent definitions, and also the fact that the Fibonacci word is Sturmian, has lead to a great development, under various terminologies, of properties of Sturmian words.

Several attempts have been made to extend Sturmian words to words over alphabets of more than two letters. It appears that none of the three equivalent definitions given below in Theorem 2.6 carry over in a satisfactory way to larger alphabets. Another approach is that initiated by Rauzy [40] and developed by Arnoux and Rauzy in [3]. They make a seemingly restrictive hypothesis on the way the growth of the complexity function is realized, and this allows some structural description of a new family of words now called Arnoux–Rauzy words, or AR-words for short. The pertinence of this approach is confirmed by the purely combinatorial study undertaken by Justin and his coauthors, yielding other extensions of properties already known for Sturmian words.

This survey is organized as follows. After some notation, we first review briefly known definitions and properties of Sturmian words. We then present some of those properties that hold for larger alphabets, with the definition introduced by Arnoux and Rauzy.

Several reviews of properties of Sturmian words and related infinite words exist, by T. C. Brown [10], Berstel [4], de Luca [20], Parvaix [39]. For early work, see Venkov [44]. An introduction is [5]. A survey on more general sequences is by Alessandri and Berthé [1]. The (future) book by the Luminy group [7] contains, among many other topics, a detailed exposition of the theory of Sturmian sequences from a more geometric point of view.

2. Sturmian Words

An infinite word x over an alphabet A is a mapping from the nonnegative integers into A. We write $x = a_0a_1\cdots$ where $a_i = x(i)$ is the ith letter of x. A *factor* of x is a (finite) word u such that $u = a_i\cdots a_j$ for some i, j with $i \leq j$. The empty word is a factor of every word. The set of factors of x is denoted $F(x)$ and the set of factors of length n is denoted $F_n(x)$. An infinite word x is *recurrent* if each of its factors appears infinitely many times in it. It is *uniformly recurrent* if moreover the occurrences of each factor have bounded gaps. This means that if a factor v starts and ends with u and has no other occurrences of u, then its length is bounded by some integer $r(u)$ depending only on u.

2.1. *Complexity and balance*

The (*subword*) *complexity function* of an infinite word x over some alphabet A is the function P that counts, for each integer $n \geq 0$, the number $P(x, n)$ of factors of length n in x:

$$P(x, n) = \mathrm{Card}(F_n(x)) .$$

Clearly, $P(x, 0) = 1$ and $P(x, 1)$ is the number of letters appearing in x. Since every factor can be extended to the right, one has $P(x, n) \leq P(x, n+1)$. Moreover, $P(x, n+m) \leq P(x, n)P(x, m)$ as is easily checked. A *right special* factor of a word x is a word u such that can be extended in at least two ways into a factor of x. More precisely, the *right degree* of u is the number of letters a such that ua is a factor of x. A right special factor has right degree at least 2. Symmetric definitions hold for the left degree and left special factors.

Example 2.1. The *Champernowne* word

$$0110111001011101111000\cdots$$

is obtained by concatenating the binary expansions of all nonnegative integers in turn. Clearly, there are 2^n factors of length n for each n. This word is recurrent but not uniformly recurrent.

Example 2.2. The *Fibonacci* word

$$f = 01001010010010100101001001010010010\cdots$$

is the limit of the sequence of words defined by

$$f_0 = 0,\ f_1 = 01$$
$$f_{n+2} = f_{n+1}f_n\,.$$

It can be checked that $P(f,n) = n+1$ for $n \geq 0$. The left special factors are the prefixes of f.

Example 2.3. The *Tribonacci* word

$$t = 01020100102010102010010 2\cdots$$

is the limit of the sequence of words defined by

$$t_0 = 0,\ t_1 = 01,\ t_2 = 0102$$
$$t_{n+3} = t_{n+2}t_{n+1}t_n\,.$$

One has $P(t,n) = 2n+1$ for $n \geq 0$.

Example 2.4. The *Thue–Morse* word

$$t = 01101001100101101001011 0\cdots$$

is the limit of the sequence u_n of words defined by

$$u_0 = 0,\ v_0 = 1$$
$$u_{n+1} = u_nv_n,\ v_{n+1} = v_nu_n\,.$$

The subword complexity function of the Thue–Morse sequence is more elaborate (see [9, 22, 23]). One has $P(t,1) = 2$, $P(t,2) = 4$ and for $n \geq 3$

$$P(t,n) = 2n+1 = \begin{cases} 6\cdot 2^{r-1} + 4s & \text{if } 0 < s \leq 2^{r-1} \\ 8\cdot 2^{r-1} + 2s & \text{if } 2^{r-1} < s \leq 2^r \end{cases}$$

where r and s defined by the decomposition $n = 2^r + s + 1$ with $0 < s \leq 2^r$.

The Fibonacci, Tribonacci and Thue–Morse words are all uniformly recurrent. They are even linearly recurrent, that is that the integer $r(u)$ is a linear function of the length $|u|$. Each of these words is also the fixed point of a morphism. The morphisms are respectively

$$\begin{matrix} 0 \mapsto 01 \\ 1 \mapsto 0 \end{matrix} \qquad \begin{matrix} 0 \mapsto 01 \\ 1 \mapsto 02 \\ 2 \mapsto 0 \end{matrix} \qquad \begin{matrix} 0 \mapsto 01 \\ 1 \mapsto 10\,. \end{matrix}$$

A *Sturmian* word is an infinite word s such that $P(s,n) = n+1$ for any integer $n \geq 0$. Thus, s is a Sturmian word if and only if it has exactly one right special factor for each length, and the degree of this factor is 2. Since $P(s,1) = 2$, any Sturmian word is over two letters. The Fibonacci word defined above is a Sturmian word. Sturmian words are aperiodic infinite words of minimal complexity ([18, 37]).

Theorem 2.5. *Let x be an infinite word. The following are equivalent*:

(i) *x is eventually periodic*,
(ii) *$P(x,n) = P(x,n+1)$ for some n*,
(iii) *$P(x,n) < n+k-1$ for some n, where k is the number of letters appearing in x*,
(iv) *$P(x,n)$ is bounded*.

There is another combinatorial description of Sturmian words, namely as balanced words. A set of words X is N-balanced if for all $x, y \in X$ of equal length, $-N \leq |x|_a - |y|_a \leq N$ for all letters a. A finite or infinite word is itself N-balanced if the set of its factors is N-balanced. The Fibonacci word is 1-balanced, and the Tribonacci word is 2-balanced. A 1-balanced word is called *balanced* for short. As we shall see below (Theorem 2.6), Sturmian and balanced words are the same.

2.2. *Mechanical words and rotations*

Given two real numbers α and ρ with $0 \leq \alpha \leq 1$, define two binary infinite words $s_{\alpha,\rho}$ and $s'_{\alpha,\rho}$ over $\{0,1\}$ by

$$\begin{aligned} s_{\alpha,\rho}(n) &= \lfloor \alpha(n+1) + \rho \rfloor - \lfloor \alpha n + \rho \rfloor \\ s'_{\alpha,\rho}(n) &= \lceil \alpha(n+1) + \rho \rceil - \lceil \alpha n + \rho \rceil \end{aligned} \qquad (n \geq 0)\,.$$

The word $s_{\alpha,\rho}$ is the *lower mechanical word* and $s'_{\alpha,\rho}$ is the *upper mechanical word* with *slope* α and *intercept* ρ. It is clear that we may assume $0 \leq \rho \leq 1$. If α is irrational, $s_{\alpha,\rho}$ and $s'_{\alpha,\rho}$ differ by at most one factor of length 2.

The terminology stems from the following graphical interpretation. Consider the straight line with equation $y = \alpha x + \rho$. The points with integer coordinates just below this line are $P_n = (n, \lfloor \alpha n + \rho \rfloor)$. Two consecutive points P_n and P_{n+1} are joined by a straight line segment that is horizontal if $s_{\alpha,\rho}(n) = 0$ and diagonal if $s_{\alpha,\rho}(n) = 1$. The same observation holds for the points located just above the line.

A special case deserves consideration, namely when $0 < \alpha < 1$ and $\rho = 0$. In this case, $s_{\alpha,0}(0) = 0$, $s'_{\alpha,0}(0) = 1$, and if α is irrational

$$s_{\alpha,0} = 0c_\alpha\,, \quad s'_{\alpha,0} = 1c_\alpha$$

where the infinite word c_α is called the *characteristic* word of α.

Mechanical words can be interpreted in several other ways. One is as cutting sequence, and is as follows. Consider again a straight line $y = \beta x + \rho$, for some $\beta > 0$ not restricted to be less than 1, and ρ not restricted to be positive.

Consider the intersections of this line with the lines of the grid with nonnegative integer coordinates. We get a sequence of intersection points. Writing a 0 for each vertical intersection point and a 1 for each horizontal intersection point, we obtain an infinite word $K_{\beta,\rho}$ that is called the *cutting sequence*. Then

$$K_{\beta,\rho} = s_{\beta/(1+\beta),\rho/(1+\beta)}\,.$$

Indeed, the transformation $(x, y) \mapsto (x+y, x)$ of the plane maps the line $y = \beta x + \rho$ to $y = \beta/(1+\beta)x + \rho/(1+\beta)$. Thus, cutting sequences are just another formulation of mechanical words (see also [19] for a more detailed discussion).

Mechanical words can also be generated by rotations. Let $0 < \alpha < 1$. The *rotation* of angle α is the mapping $R = R_\alpha$ from $[0, 1[$ into itself defined by

$$R(z) = \{z + \alpha\}\,.$$

Iterating R, one gets $R^n(\rho) = \{n\alpha + \rho\}$. Thus, defining a partition of $[0, 1[$ by

$$I_0 = [0, 1-\alpha[\,, \quad I_1 = [1-\alpha, 1[\,,$$

one gets

$$s_{\alpha,\rho}(n) = \begin{cases} 0 & \text{if } R^n(\rho) \in I_0 \\ 1 & \text{if } R^n(\rho) \in I_1\,. \end{cases}$$

It is convenient to identify $[0, 1[$ with the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. Then, for any subinterval I of $[0, 1[$, the sets $R(I)$ and $R^{-1}(I)$ are always intervals. As an example of the use of rotations, consider a word $w = b_0 b_1 \cdots b_{m-1}$, with $b_0, b_1, \ldots$ letters. We want to know whether w is a factor of some $s_{\alpha,\rho} = a_0 a_1 \cdots$, with $a_0, a_1, \ldots$ letters. Observe that $a_{n+k} = b_i$ if and only if $R^{n+i}(\rho) \in I_{b_i}$, or equivalently, if and only if $R^n(\rho) \in R^{-i}(I_{b_i})$. Thus, for $n \geq 0$,

$$w = a_n a_{n+1} \cdots a_{n+m-1} \iff R^n(\rho) \in I_w$$

where I_w is the interval

$$I_w = I_{b_0} \cap R^{-1}(I_{b_1}) \cap \cdots \cap R^{-m+1}(I_{b_{m-1}})\,.$$

The interval I_w is non empty if and only if w is a factor of $s_{\alpha,\rho}$. Observe that this property is independent of ρ, and thus mechanical words of same slope have the same set of factors.

Mechanical words are quite naturally defined as two-sided infinite words. However, it appears that several properties, such as the theorem below, only hold with some restrictions (see e.g. [17, 18, 43])

Theorem 2.6. *Let s be a binary infinite word. The following are equivalent*:

(i) *s is Sturmian*,
(ii) *s is balanced and aperiodic*,
(iii) *s is mechanical with an irrational slope.*

There is an interesting variation of the second condition of this theorem that deserves to be mentioned. The first concerns balance. Instead of considering the number $|u|_a$ of occurrences of a letter a in a word u, one may compute more generally the number $|u|_z$ of occurrences of the word z as a factor in u. It has been shown by Fagnot and Vuillon [27] that this extended notion also characterizes Sturmian words in the following sense.

Theorem 2.7. *An infinite word x is Sturmian if and only if, for each pairs of factors u, v of equal length*

$$||u|_z - |v|_z| \leq |z|$$

for all nonempty words z.

2.3. *Factor graph*

One of the basic tools for investigating the growth of complexity in infinite words was introduced by G. Rauzy. It is called the *Rauzy graph* or the factor graph. Let x be an infinite word. The *factor graph* $G_n(x)$ of order n is the graph with vertex set $F_n(x)$ and edges defined as follows. A tuple $(p, a, b, s) \in A^* \times A \times A \times A^*$ is an edge if and only if $pa = bs \in F_{n+1}(x)$.

The outdegree of a vertex in $G_n(x)$ is the right degree of the factor. In the case of a Sturmian word, there is a unique right special factor d_n of length n. Let b be the letter such that $bd_{n-1} = d_n$. Then the edges leaving d_n are $(d_n, 0, b, d_{n-1}0)$ and $(d_n, 1, b, d_{n-1}1)$. Similarly, there is exactly one vertex with indegree 2. This is the left special factor g_n of length n. Observe that $d_n = g_n$ if and only if d_n is a palindrome word. See Fig. 1 for the factor graphs of the Fibonacci word.

The factor graph of order n of a Sturmian word x is composed of three paths: the first is from g_n to d_n, both vertices included. This path is never empty. There

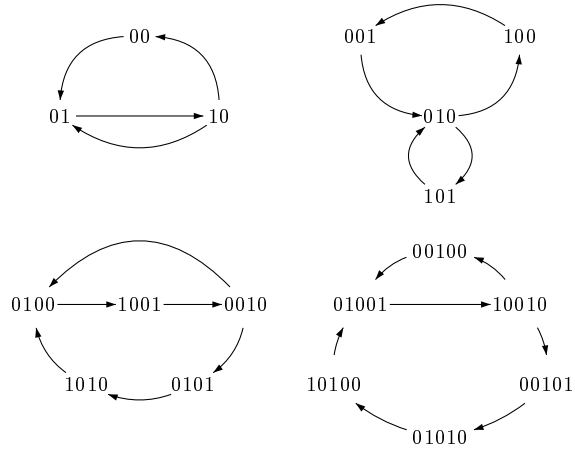

Fig. 1. Factor graphs for the Fibonacci word.

are two other paths, from d_n to g_n, one through vertex $d_{n-1}0$, the other through $d_{n-1}1$. We consider that the endpoints d_n and g_n are not part of these paths. Then such a path may be empty. This happens if and only if $d_{n-1}0 = g_n$ or $d_{n-1}1 = g_n$ which in turn is the case if and only if $d_{n-1} = g_{n-1}$ because g_{n-1} is a prefix of g_n. An infinite word y is *accepted* by $G_n(x)$ if there is an infinite path in the graph with label y. Of course, x itself is accepted by $G_n(x)$. The set of accepted infinite words is a sofic dynamical system, and the minimal system generated by x is the intersection of theses systems, for $n \geq 0$.

Factor graphs have been used by V. Berthé [6] to compute the explicit values of the frequencies of the factors of a Sturmian word.

2.4. *Return words*

Let x be an infinite word, and let u be a factor of x. A *return word* for u in x is a factor of x that starts at some occurrence of u in x and that stops just before the next occurrence of u in x. Thus, a return word is a nonempty factor w such that u is a prefix of wu and moreover wu contains exactly two occurrences of the word u. If x is uniformly recurrent, then every factor u of x has finitely many return words. The notion of return word was introduced independently by Durand [26] and by Holton and Zamboni [30] in order to study primitive substitutive sequences.

Consider for example the Thue–Morse word $t = 011010011001\cdots$ defined earlier. The return words for $u = 01$ are 01, 010, 011, and 0110. On the contrary, in the Fibonacci word $f = 0100101001001\cdots$, the word 010 for instance has two return words 01 and 010. This is a characteristic property.

Theorem 2.8. *A uniformly recurrent word is Sturmian if and only if each of its prefixes has exactly* 2 *return words.*

This was proved by Vuillon [45]. A simplified proof appears in [35]. As we shall see below, this characterization holds not longer for Arnoux–Rauzy words.

2.5. *Palindromic closure*

The *palindromic closure* of a finite word u is the shortest palindrome word $u^{(+)}$ having u as a prefix. For instance, for $u = 0100101$, one gets $u^{(+)} = 01001010010$. The following characterization is due to de Luca [20].

Proposition 2.9. *An infinite binary word s is a characteristic Sturmian word if and only if there exists an infinite binary word $\Delta(s) = a_0a_1\cdots$ with infinitely many occurrences of both letters such that*

$$s = \lim_{n\to\infty} u_n$$

where $u_0 = \varepsilon$ and $u_{n+1} = (u_na_n)^{(+)}$, for $n \geq 0$.

The word $\Delta(s)$ is the *directive sequence* of s. As we shall see, the same result holds for AR-words.

2.6. *Palindromes*

The first palindromes that are factors of the Fibonacci word are easily computed. They are ε, 0, 1, 00, 010, 101, 1001. It appears that there is just one palindrome of even length, and that there are two palindromes of odd length. This is a general property that even characterizes Sturmian words, as shown by Droubay and Pirillo [25].

Theorem 2.10. *An infinite word is Sturmian if and only if, for each nonnegative integer n, there is exactly one palindrome of length n, if n is even, and there are exactly two palindromes of length n, if n is odd.*

2.7. *Decimation*

Given an infinite binary word x over the alphabet $\{a, b\}$, we number the occurrences of each letter from left to right, starting with number 1. For instance, in the Fibonacci word $f = abaababaaba\cdots$, the fourth a appears at position 5, and the fifth at position 7. The k-decimation of x is the word obtained in withdrawing all letters whose occurrence has a number that is not a multiple of k. It appears that the Fibonacci word is invariant under all decimation. This is a general result, already mentioned by Rauzy and proved by Justin and Pirillo [34].

Theorem 2.11. *An infinite word is invariant under all decimations if and only if it is a characteristic mechanical word* (*that is of rational or irrational slope*).

Parvaix [38] studied extensions of the decimation operation.

2.8. *Doubling map*

For each $0 < \alpha < 1$, the set S^α of Sturmian words of slope α over the alphabet $\{0, 1\}$ is a minimal dynamical system. Each word $x \in S^\alpha$ can be viewed as the binary expansion of some real number $r(x)$ in the interval $[0, 1[$. It is not difficult to check that $x < x'$ for the lexicographic order if and only if $r(x) < r(x')$. Also, provided α is irrational, the smallest and greatest words in S^α (for the lexicographic order) are the two words $0c_\alpha$ and $1c_\alpha$, where c_α is the characteristic word of slope α. Clearly, $r(1c_\alpha) = 1/2 + r(0c_\alpha)$, and in fact the set $r(S^\alpha)$ is entirely contained in the interval $[r(0c_\alpha), r(1c_\alpha)]$ of length $1/2$. Moreover, the set $r(S^\alpha)$ is a dynamical system for the operation induced on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ by the shift, which is the well-known "doubling map" $\sigma : t \mapsto 2t \bmod 1$. This observation admits a converse. For each interval $C_\mu = [\mu, 1/2 + \mu]$, there is a unique α such that $r(S^\alpha) \subset C_\mu$, and moreover, there is no other dynamical system for the doubling map that is a subset of C_μ. This correspondence is described in detail by Bullett and Sentenac [11] where they consider it as a reformulation of the results in [37]. They also prove that, for irrational α, the number $\mu = r(0c_\alpha)$ is transcendental.

Considering probability measures, the first result may be restated by saying that there is a unique probability measure with support contained in C_μ and invariant for the doubling map σ. It is quite natural to call it the Sturmian measure of parameter

α, where α is the unique number such that $r(S^\alpha) \subset C_\mu$. If α is rational, this measure is the finite measure m_α defined by $m_\alpha(x) = 1/\mathrm{Card}(r(S^\alpha))$ for $x \in r(S^\alpha)$. Sturmian measures can also be defined directly by considering the mechanical words (Bousch and Mairesse [8]).

3. Arnoux–Rauzy Words

The definition of Sturmian words by their complexity implies that they are over a binary alphabet. Equivalent words over more than two letters can be defined by relaxing the minimality condition for the complexity (see e.g. [17]). Also, balance could be considered, but this seems to be quite complex. Another approach was introduced by Rauzy [40], and investigated in depth by Arnoux and Rauzy [3], and resulted in a family of words now called Arnoux–Rauzy sequences or AR-words for short. However, as we shall see, none of the oldest characterizations of Sturmian words (growth, balance, mechanical property) holds for this family.

An infinite word x over a k letter alphabet is an *Arnoux–Rauzy* word if there is a unique left special factor and an unique right special factor for each length, and if, moreover, these factors have right degree (respectively left degree) exactly k.

The Tribonacci word is a typical example of an AR-word over 3 letters. The prefixes of this word are the left special factors, and the reversals of these prefixes are the right special factors.

The Rauzy graphs (or factor graphs) of an AR-word x over k letters are quite similar to those of a Sturmian word. For a given integer n, the graph $G_n(x)$ has exactly one vertex d with outdegree k, and one vertex g with indegree k. These may be identical (if and only if the right special factor is a palindrome). All other vertices have indegree 1 and outdegree 1, so there are k nontrivial paths in the graph from the vertex d to the vertex g and one path from g to d. This path is the null path if and only if $d = g$. Any AR-word is uniformly recurrent and its set of factors is closed under reversal.

A more general definition has been introduced and studied by Droubay, Justin and Pirillo [24] (also [33]). They only require that the right special and the left special factors are unique, but the degree of these factors may be less than k. They call these words *episturmian*. AR-words appear as strict episturmian words. There is a growing interest in AR-words and in episturmian words ([2, 14, 16, 31, 36, 47].

An AR-word x is *characteristic* if its prefixes are the left special factors. Again, the Tribonacci word is characteristic. Observe that, in the binary case, characteristic words were defined by a geometric property which appeared to be equivalent to the combinatorial definition.

Rauzy's rules [41] are one of the methods for describing characteristic AR-words. For simplicity, we give them here the ternary case. There are three of them:

$$(x, y, z) \mapsto (x, xy, xz)$$
$$(x, y, z) \mapsto (yx, y, yz)$$
$$(x, y, z) \mapsto (zx, zy, z)\,.$$

As an example, if you apply the rules in turn to (0, 1, 2), the result is:

0	1	2
0	01	02
010	01	0102
0102010	010201	0102
0102010	0102010010201	01020100102
	...	

which clearly converges to the Tribonacci sequence. This behavior can also be expressed by iterating morphisms (Rauzy [41]). Given an alphabet A, define morphisms $\tau_a : A^* \mapsto A^*$ for $a \in A$ by setting

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

For a word $w = a_1 \cdots a_n$, set $\tau_w = \tau_{a_1} \cdots \tau_{a_n}$. As an example, $\tau_{012}(0) = \tau_{01}(20) = \tau_0(1210) = 0102010$.

Proposition 3.1. *Any characteristic AR-word over an alphabet A is obtained as the limit of any of the sequences $(\tau_{d_n}(a))_{n\geq 0}$ for $a \in A$, where d_n is the prefix of length n of some infinite word Δ that contains infinitely many occurrences of each letter in A.*

Palindromic closure is another equivalent way to generate characteristic AR-words [24].

Theorem 3.2. *Any characteristic AR-word over an alphabet A is obtained as the limit of a sequence $(u_n)_{n\geq 0}$ of words, where $u_{n+1} = (u_n a_n)^{(+)}$ is the palindromic closure of $u_n a_n, u_0 = \varepsilon$, and the infinite word $\Delta = a_0 a_1 \cdots$ contains infinitely occurrences of each of the letters in A.*

We call the words u_n described in this theorem A-central words. Again, the Tribonacci word is a good example. Given $\Delta = 012012\cdots$, we get

$$\mathit{0}10\mathit{2}0100\mathit{1}0201010\mathit{2}01001020102\cdots$$

where the symbols of the word Δ are italicized.

It happens that the palindromes occurring in such a sequence have a special property that allows an alternate description and construction. We report the formulation of Risley and Zamboni [42]. Instead of palindromic closure, the basic operation is *suffix replication*. For this, the letters of the word Δ are marked when they are added at the end of the word, and $u_{n+1} = u_n a_n v_n$, where v_n is the longest

suffix of u_n containing no marked occurrence of a_n. As and example, consider again the Tribonacci word.

$$\begin{aligned} u_0 &= \varepsilon \\ u_1 &= 0 \\ u_2 &= 010 \\ u_3 &= 01020\,10 \\ u_4 &= 0102010\,0102010 \\ u_5 &= 0102010\,0102010\,1020100102010\,. \\ &\dots \end{aligned}$$

In other words, the suffix replication constructs precisely the palindromic closure. The relationship between these formulations is contained in the formula from [33] (again, $d_n = a_0 \cdots a_{n-1}$ and $u_{n+1} = (u_n a_n)^{(+)}$)

$$u_{n+1} = \tau_{d_n}(a_n) u_n\,.$$

Because of these numerous combinatoric properties, some geometric and number-theoretic characterizations were looked for. In particular, rotations on the two dimensional torus $\mathbb{T}^2$ seemed to be good candidates. Such a rotation was given by Rauzy in 1982 for the Tribonacci word [40], and they are known for large families of other AR-words. It was believed that every ternary AR-word codes the orbit of a point under a rotation on $\mathbb{T}^2$. However, this does not hold, as proved by Cassaigne, Ferenczi and Zamboni [12]. They prove the proposition below, and then use another result by Rauzy to get the negative answer.

Proposition 3.3. *There exist AR-words x that are imbalanced, in the following sense. For each integer n, there are factors u, v of x equal length with $||u|_a - |v|_a| > n$ for some letter a.*

It can be shown that the Tribonacci word is 2-balanced, and more generally, if x is an AR-word that is linearly recurrent, then x is N-balanced for some N (a word is linearly recurrent if there is a constant K such that for every factor u and for every return word v of u, one has $K^{-1}|u| \leq |v| \leq K|u|$). Linearly recurrent AR-words were completely characterized in [15, 42].

3.1. *Return words*

Theorem 2.8 extends partially to AR-words [35].

Proposition 3.4. *In a k-letter AR-word, the number of return words of any nonempty factor is exactly k.*

This property however is not characteristic of AR-words. Consider for instance the modified Chacon word (this has complexity $2n + 1$, see Ferenczi [28] for other

properties). It is the fix point of the morphism

$$0 \mapsto 01$$
$$1 \mapsto 201$$
$$2 \mapsto 2201$$

and starts as follows

$$c = 01201220101201220122010120 1 \cdots .$$

For instance, the factor 12 has the return words 120, 1220, 12010 and 122010.

Proposition 3.4 has been extended to regular interval exchange transformations by Vuillon [46]. An interval exchange transformation is a piecewise affine transformation that maps a partition into intervals of the unit interval into another partition, according to a permutation. This transformation could be more complicated than a rotation. A transformation is *regular* if no endpoint of the intervals of the partition is non-trivially mapped into another endpoint. Vuillon shows that infinite word that codes the orbit of a point under a regular interval exchange transformation of k intervals has the property that each of its nonempty factors has exactly k return words.

3.2. *A theorem of Fine and Wilf*

A *period* of a words $w = a_1 \cdots, a_n$, where $a_1, \ldots, a_n$ are letters, is an integer $p \leq n$ such that $a_i = a_{i+p}$ for $1 \leq i \leq n - p$. A remarkable theorem, due to Fine and Wilf [29] states that if w has two periods p, q and $n \geq p + q - \gcd(p, q)$, then $\gcd(p, q)$ is also a period of w. The bound of this result is known to be sharp, and the description of words of maximal length for which the theorem fails to hold shows a surprising relationship with prefixes of characteristic Sturmian words. It suffices to consider the case where $\gcd(p, q) = 1$. It was shown by de Luca and Mignosi [21] that the words of maximal length for which Fine and Wilf's fails are precisely the words u_n of Theorem 3.2 over two letters, that is the central words over a two letter alphabet. This result was extended to three letters by Castelli, Mignosi and Restivo [13] and to the general case by Justin [32]. We just sketch the construction.

For a sequence $p = (p_1, \ldots, p_n)$ of nonnegative integer which are not all zero, set $|p| = p_1 + \cdots + p_n$ and $\gcd(p) = \gcd(p_1, \ldots, p_n)$. Further, let $\mu(p) = \min\{p_1, \ldots, p_n \mid p_i \neq 0\}$ and let $\lambda = \lambda(p)$ be any index such that $p_\lambda = \mu(p)$ is the minimum. Define a mapping $R : p \mapsto p' = R(p)$ by

$$p'_i = \begin{cases} p_i & \text{if } p_i = 0 \text{ or } i = \lambda \\ p_i - p_\lambda & \text{otherwise}. \end{cases}$$

The mapping R just achieves one step in the Euclidean algorithm in its additive form. Set $p^{(0)} = p$, and $p^{(k+1)} = R(p^{(k)})$, and let $m(i) = m(i, p)$ be the smallest

exponent k such that $p^{(k)}$ contains exactly i zero entries. Finally, set $h(i) = h(i,p) = |p^{m(i)}|$. It follows from Euclid's algorithm that $h(n-1) = \gcd(p)$. Justin extends a result by Castelli, Mignosi and Restivo for three periods and shows

Theorem 3.5. *If a words w has periods $p_1, \ldots, p_n$ and has length at least $f(p)$, where*

$$f(p) = \frac{|p|}{n-1} + \sum_{i=1}^{n-2} \frac{h(i,p)}{(n-i)(n-i-1)} - \gcd(p)$$

then w has period $\gcd(p)$.

Say that a tuple p is *good* if $m(1,p) = \cdots = m(n-1,p)$ and $\gcd(p) = 1$. For a good p, the formula for f reduces to $f(p) = (|p|-1)/(n-1)$. Then Justin shows the following:

Theorem 3.6. *Let A be an n-letter alphabet. The set of words w of length $f(p)-1$ having n periods $p_1, \ldots, p_n$ such that $p = (p_1, \ldots, p_n)$ is good is the set of A-central words.*

Example 3.7. Consider the triple $p = (7, 11, 13)$. Euclid's algorithm gives

$$\begin{aligned} p^{(0)} &= (\underline{7}, 11, 13) \\ p^{(1)} &= (7, \underline{4}, 6) \\ p^{(2)} &= (3, 4, \underline{2}) \\ p^{(3)} &= (\underline{1}, 2, 2) \\ p^{(4)} &= (1, 1, 1) \end{aligned}$$

where the minimal entry is underlined. The triple (7,11,13) is good. Consider the word obtained by concatenating the positions of the underlined entries, here $d = 0120$. The central word generated by this word is the prefix 01020100102010 of the Tribonacci word. This word has indeed the three periods 7, 11, 13, and has length 14. Thus it is a maximal word where the generalized Fine and Wilf theorem fails.

Acknowledgments

I thank Jacques Justin and Jean Mairesse for helpful comments of an earlier version of this paper.

References

1. P. Alessandri and V. Berthé, *Three distance theorems and combinatorics on words*, L'enseignement mathématique **44** (1998), 103–132.
2. P. Arnoux and S. Ito, *Pisot substitution and Rauzy fractal*, Bull. Belg. Math. Soc. Simon Sterin **8** (2001), 181–207.
3. P. Arnoux and G. Rauzy, *Représentation géométrique de suites de complexité* $2n+1$, Bull. Soc. Math. France **119** (1991), 199–215.

4. J. Berstel, *Recent results on Sturmian words*, in Developments in Language Theory II, eds. J. Dassow and A. Salomaa, World Scientific, 1996, pp. 13–24.
5. J. Berstel and P. Séébold, *Sturmian words*, in Algebraic Combinatorics on Words, ed. Lothaire, Chap. 2, to appear.
6. V. Berthé, *Fréquences des facteurs des suites sturmiennes*, Theoret. Comput. Sci. **165** (1996), 295–309.
7. V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel (eds.), *Introduction to finite automata and substitution dynamical systems*, Technical Report 2000-43, Institut Mathématique de Luminy, 2000.
8. T. Bousch and J. Mairesse, *Asymptotic height optimization for topical IFS, Tetris heaps, and the finiteness conjecture*, J. Amer. Math. Soc. **15** (2002), 77–111.
9. S. Brlek, *Enumeration of the factors in the Thue–Morse word*, Discrete Appl. Math. **24** (1989), 83–96.
10. T. C. Brown, *Descriptions of the characteristic sequence of an irrational*, Can. Math. Bull. **36**(1) (1993), 15–21.
11. S. Bullett and P. Sentenac, *Ordered orbits of the shift, square roots, and the devil's staircase*, Math. Proc. Cambridge Philos. Soc. **115** (1994), 451–481.
12. J. Cassaigne, S. Ferenczi and L. Q. Zamboni, *Imbalances in Arnoux-Rauzy sequences*, Ann. Inst. Fourier. **50** (2000), 1265–1276.
13. M. G. Castelli, F. Mignosi and A. Restivo, *Fine and Wilf's theorem for three periods and a generalization of Sturmian words*, Theoret. Comput. Sci. **218** (1999), 83–94.
14. N. Chekhova, *Algorithme d'approximation et propriétés ergodiques des suites d'Arnoux–Rauzy*, manuscript, 1999.
15. N. Chekhova, *Fonctions de récurrence des suites d'Arnoux–Rauzy et réponse à une question de Hedlund et Morse*, manuscript, 1999.
16. N. Chekhova, P. Hubert and A. Messaoudi, *Propriétés combinatoires, ergodiques et arithmétiques de la substitution de Tribonacci*, Technical Report 98-24, IML, 1998.
17. E. M. Coven, *Sequences with minimal block growth II*, Math. Syst. Theory **8** (1974), 376–382.
18. E. M. Coven and G. A. Hedlund, *Sequences with minimal block growth*, Math. Syst. Theory **7** (1973), 138–153.
19. D. Crisp, W. Moran, A. Pollington and P. Shiue, *Substitution invariant cutting sequences*, J. Théor. Nombres Bordeaux **5** (1993), 123–137.
20. A. de Luca, *Sturmian words*: *Structure, combinatorics, and their arithmetics*, Theoret. Comput. Sci. **183** (1997), 45–82.
21. A. de Luca and F. Mignosi, *On some combinatorial properties of Sturmian words*, Theoret. Comput. Sci. **136** (1994), 361–385.
22. A. de Luca and S. Varricchio, *On the factors of the Thue–Morse word on three symbols*, Inf. Proc. Lett. **27** (1988), 281–285.
23. A. de Luca and S. Varricchio, *Some combinatorial properties of the Thue–Morse sequence*, Theoret. Comput. Sci. **63** (1989), 335–348.
24. X. Droubay, J. Justin and G. Pirillo, *Episturmian words and some constructions of de Luca and Rauzy*, Theoret. Comput. Sci. **255** (2001), 539–553.
25. X. Droubay and G. Pirillo, *Palindromes and Sturmian words*, Theoret. Comput. Sci. **223** (1999), 73–85.
26. F. Durand, *A characterization of substitutive sequences using return words*, Discrete Math. **179** (1998), 89–101.
27. I. Fagnot and L. Vuillon, *Generalized balance in Sturmian words*, Technical Report 2000-02, Liafa, 2000.

28. S. Ferenczi, *Les transformations de Chacon*: *Combinatoire, structure géométrique, liens avec les systèmes de complexité* $2n + 1$, Bull. Soc. Math. France **123** (1995), 271–292.
29. N. J. Fine and H. S. Wilf, *Uniqueness theorem for periodic functions*, Proc. Amer. Math. Soc. **16** (1965), 109–114.
30. C. Holton and L. Q. Zamboni, *Geometric realizations of substitutions*, Bull. Soc. Math. France **126** (1998), 149–179.
31. S. Ito and M. Kimura, *On Rauzy fractal*, Japan J. Indust. Appl. Math. **8** (1991), 461–486.
32. J. Justin, *On a paper by Castelli, Mignosi, Restivo*, Theoret. Inf. Appl. **34** (2000), 373–377.
33. J. Justin and G. Pirillo, *Episturmian words and episturmian morphisms*, manuscript, November 2000.
34. J. Justin and G. Pirillo, *Decimation and Sturmian words*, Theoret. Inf. Appl. **31** (1997), 271–290.
35. J. Justin and L. Vuillon, *Return words in Sturmian and episturmian words*, Theoret. Inf. Appl. **34** (2000), 343–356.
36. A. Messaoudi, *Propriétés arithmétiques et dynamiques du fractal de Rauzy*, J. Théor. Nombres Bordeaux **10** (1998), 135–162.
37. M. Morse and G. A. Hedlund, *Symbolic dynamics II*: *Sturmian sequences*, Amer. J. Math. **61** (1940), 1–42.
38. B. Parvaix, *Propriétés d'invariance des mots sturmiens*, J. Théor. Nombres Bordeaux **9** (1997), 351–369.
39. B. Parvaix, *Contribution à l'étude des Mots Sturmiens*, Vol. 25, Publications du Lacim, Montreal, 1998.
40. G. Rauzy, *Suites à termes dans un alphabet fini, in Sémin. Théorie des Nombres*, University of Bordeaux, 1982-1983, Exposé No 25, pp. 1–16
41. G. Rauzy, *Mots infinis en arithmétiquae*, in Automata on Infinite Words, eds. M. Nivat and D. Perrin, Lecture Notes in Computer Science 192, Springer-Verlag, 1985, pp. 165–171.
42. R. Risley and L. Q. Zamboni, *A generalization of Sturmian sequences*: *Combinatorial structure and transcendence*, Acta Arith. **95** (2000), 167–184.
43. R. Tijdeman, *On the minimal complexity of infinite words*, Indag. Math. **10** (1999), 123–129.
44. B. A. Venkov, *Elementary Number Theory*, Wolters-Noordhoff, Groningen, 1970.
45. L. Vuillon, *A characterization of Sturmian words by return words*, European J. Combin. **22** (2001), 263–275.
46. L. Vuillon, *On the number of return words in infinite words with complexity* $2n + 1$, Technical Report 2000-15, Liafa, 2000.
47. N. Wozny and L. Q. Zamboni, *Frequencies of factors in Arnoux–Rauzy sequences*, Acta Arith. **96** (2001), 261–278.