

# On the Index of Sturmian Words

Jean Berstel

Institut Gaspard Monge, Université Marne-la-Vallée, F-77454 Marne-la-Vallée,  
France

**Summary.** An infinite word  $x$  has finite index if the exponents of the powers of primitive words that are factors of  $x$  are bounded. F. Mignosi has proved that a Sturmian word has finite index if and only if the coefficients of the continued fraction development of its slope are bounded. Mignosi's proof relies on a delicate analysis of the approximation of the slope by rational numbers. We give here a proof based on combinatorial properties of words, and give some additional relations between the exponents and the slope.

## 1 Introduction

Sturmian words are infinite words over a binary alphabet that have exactly  $n + 1$  factors of length  $n$  for each  $n \geq 0$ . It appears that these words admit several equivalent definitions, and can even be described explicitly in arithmetic form. For instance, every Sturmian word has a *slope* associated with it, which is an irrational number in the interval  $[0, 1]$ .

Sturmian words have a long history. A clear exposition of early work by J. Bernoulli, Christoffel, and A. A. Markov is given in the book by Venkov [30]. The term “Sturmian” has been used by Hedlund and Morse in their development of symbolic dynamics [15–17]. These words are also known as Beatty sequences, cutting sequences, or characteristic sequences. There is a large literature about properties of these sequences (see for example Coven, Hedlund [8], Series [28], Fraenkel *et al.* [14], Stolarsky [29]). From a combinatorial point of view, they have been considered by S. Dulucq and D. Gouyou-Beauchamps [13], Rauzy [25,26], Brown [6], Ito, Yasutomi [18], Crisp *et al.* [7] in particular in relation with iterated morphisms, and by Séebold [27], Mignosi [21]. Sturmian words appear in ergodic theory [24], in computer graphics [5], in crystallography [4], and in pattern recognition. Standard words, and finite factors of Sturmian words are considered in depth by De Luca [12,9,11,10], see also [3]. A survey is [1]. A more systematic presentation of Sturmian words is in preparation ([2]).

The aim of this paper is to present a new proof, with some improvements, of a theorem by Mignosi [21] cited below. Let  $x$  be an infinite word, and let  $F(x)$  be the sets of its factors (subwords). For  $w \in F(x)$ , the *index* of  $w$  in  $x$  is the greatest integer  $d$  such that  $w^d \in F(x)$ , if such an integer exists. Otherwise,  $w$  is said to have infinite index.

An infinite word  $x$  has *bounded index* if there exists an integer  $d$  such that every nonempty factor of  $x$  has an index less than or equal to  $d$ .

**Theorem 1.** *A Sturmian word has bounded index if and only if the continued fraction expansion of its slope has bounded partial quotients.*

An initial contribution to this result was by Karhumäki [20] who proved that the Fibonacci word is fourth power free. Mignosi's proof uses involved arguments from number theory. Our proof is combinatorial, and follows an argument by Mignosi and Pirillo [22] in their proof of the sharp bound for the index in the Fibonacci word.

The next section is devoted to a short introduction to Sturmian words. In particular, the standard sequence of a Sturmian word is introduced. The following section gives the proof.

## 2 Definitions

In this paper, words will be over a binary alphabet  $A = \{0, 1\}$ .

The *complexity function* of an infinite word  $x$  over some alphabet  $A$  is the function that counts, for each integer  $n \geq 0$ , the number  $P(x, n)$  of factors of length  $n$  in  $x$ . A *Sturmian* word is an infinite word  $s$  such that  $P(s, n) = n + 1$  for any integer  $n \geq 0$ . Sturmian words are aperiodic infinite words of minimal complexity. Indeed, an infinite word of lower complexity is eventually periodic. Since  $P(s, 1) = 2$ , any Sturmian word is over two letters. A *right special* (*left special*) factor of a word  $x$  is a word  $u$  such that  $u0$  and  $u1$  ( $0u$  and  $1u$ ) are factors of  $x$ . Thus a word  $x$  is Sturmian if and only if it has exactly one right special factor of each length.

Given two real numbers  $\alpha$  and  $\rho$  with  $\alpha$  irrational and  $0 < \alpha < 1$ , we define two infinite words

$$s_{\alpha, \rho} : \mathbf{N} \rightarrow A, \quad s'_{\alpha, \rho} : \mathbf{N} \rightarrow A$$

by

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

The numbers  $\alpha$  and  $\rho$  are *slope* and the *intercept*. Words  $s_{\alpha, \rho}$  and  $s'_{\alpha, \rho}$  are called *mechanical*.

**Theorem 2.** [17] *Let  $s$  be an infinite word. The following are equivalent:*

- (i)  $s$  is Sturmian;
- (ii)  $s$  is mechanical.

A special case deserves consideration, namely when  $\rho = 0$ . In this case,  $s_{\alpha,0}(0) = [\alpha] = 0$ ,  $s'_{\alpha,0}(0) = [\alpha] = 1$ , and

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

where the infinite word  $c_\alpha$  is called the *characteristic* word of  $\alpha$ . It can be shown that a Sturmian word is characteristic if and only if every prefix is left special.

There is a close relation between the slope of a characteristic word and the combinatorial structure of this word.

Let  $(d_1, d_2, \dots, d_n, \dots)$  be a sequence of integers, with  $d_1 \geq 0$  and  $d_n > 0$  for  $n > 1$ . To such a sequence, we associate a sequence  $(s_n)_{n \geq -1}$  of words by

$$s_{-1} = 1, \quad s_0 = 0, \quad s_n = s_{n-1}^{d_n} s_{n-2} \quad (n \geq 1) \quad (1)$$

The sequence  $(s_n)_{n \geq -1}$  is a *standard sequence*, and the sequence  $(d_1, d_2, \dots)$  is its *directive sequence*. Observe that if  $d_1 > 0$ , then any  $s_n$  ( $n \geq 0$ ) starts with 0; on the contrary, if  $d_1 = 0$ , then  $s_1 = s_{-1} = 1$ , and  $s_n$  starts with 1 for  $n \neq 0$ . Every  $s_{2n}$  ends with 0, every  $s_{2n+1}$  ends with 1.

*Example 1.* The directive sequence  $(1, 1, \dots)$  gives the standard sequence defined by  $s_n = s_{n-1}s_{n-2}$ , that is the sequence of finite Fibonacci words. Observe that the directive sequence  $(0, 1, 1, \dots)$  results in the sequence of words obtained from Fibonacci words by exchanging 0 and 1.

**Proposition 1.** [14] *Let  $\alpha = [0, 1 + d_1, d_2, \dots]$  be the continued fraction expansion of some irrational  $\alpha$  with  $0 < \alpha < 1$ , and let  $(s_n)$  be the standard sequence associated to  $(d_1, d_2, \dots)$ . Then every  $s_n$  is a prefix of  $c_\alpha$  and*

$$c_\alpha = \lim_{n \rightarrow \infty} s_n.$$

*Example 2.* Consider  $\alpha = (\sqrt{3} - 1)/2 = [0, 2, 1, 2, 1, \dots]$ . The directive sequence is  $(1, 1, 2, 1, 2, 1, \dots)$ , and the standard sequence starts with  $s_1 = 01$ ,  $s_2 = 010$ ,  $s_3 = 01001001$ ,  $\dots$ , whence

$$c_{(\sqrt{3}-1)/2} = 0100100101001001001001001001001 \dots$$

Due to the periodicity of the development, we get for  $n \geq 2$  that  $s_{n+2} = s_{n+1}^2 s_n$  if  $n$  is odd, and  $s_{n+2} = s_{n+1} s_n$  if  $n$  is even.

### 3 Index

As usual, a word of the form  $w = (xy)^n x$  is written as  $w = u^r$ , with  $u = xy$  and  $r = n + |x|/|u|$ . The rational number  $r$  is the *exponent* of  $u$ , and if  $u$  is primitive, it is the root of the fractional power  $w$ .

Let  $x$  be an infinite word. For  $w \in F(x)$ , the *index* of  $w$  in  $x$  is the number

$$\text{ind}(w) = \sup\{r \in \mathbf{Q} \mid w^r \in F(x)\}$$

if such an integer exists. Otherwise,  $w$  is said to have infinite index. We also define the *prefix index*  $\text{pind}(w)$  to be the greatest number  $r$  such that  $w^r$  is a prefix of  $x$ . The prefix index is always finite, provided  $x$  is not periodic, and it is zero when the first letter of  $w$  differs from the first letter of  $x$ .

**Proposition 2.** *Every nonempty factor of a Sturmian word  $s$  has finite index in  $s$ .*

*Proof.* Assume the contrary. There exist a Sturmian word  $s$  and a nonempty factor  $u$  of  $s$  such that  $u^n$  is a factor of  $s$  for every  $n \geq 1$ . Consequently, the periodic word  $u^\omega$  is in the dynamical system generated by  $s$ . Since this system is minimal,  $F(s) = F(u^\omega)$ , a contradiction.  $\square$

An infinite word  $x$  has *bounded index* if there exists a number  $d$  such that every nonempty factor of  $x$  has an index less than or equal to  $d$ . If  $x$  has bounded index, the upper bound might be irrational. For instance, Mignosi and Pirillo [22] have shown that in the case of the Fibonacci word, this bound is  $2 + \tau$ , where  $\tau = (1 + \sqrt{5})/2$ . We will get this as a consequence of our investigations.

We start with a notation. Let  $(s_n)_{n \geq -1}$  be the standard sequence of the characteristic word  $c_\alpha$ , with  $\alpha = [0, 1 + d_1, d_2, \dots]$ . For  $n \geq 3$  (and for  $n = 2$  if  $d_1 \geq 1$ ), define

$$t_n = s_{n-1}^{d_n-1} s_{n-2} s_{n-1}$$

and for  $n \geq 0$  set

$$p_n = s_{n-1}^{d_n} s_{n-2}^{d_{n-1}} \dots s_0^{d_1}$$

In particular  $p_0 = \varepsilon$ . In view of (1), the word  $t_n$  is just a conjugate of  $s_n$ . More precisely

**Lemma 1.** (i) *For  $n \geq 3$  (and for  $n = 2$  if  $d_1 \geq 1$ ), one has*

$$s_n s_{n-1} = s_{n-1} t_n, \quad s_{n-1} s_n = s_n t_{n-1}.$$

(ii) *For  $n \geq 0$ ,*

$$s_n s_{n-1} = \begin{cases} p_n 10 & \text{if } n \text{ is odd} \\ p_n 01 & \text{if } n \text{ is even.} \end{cases}$$

$$s_{n-1} s_n = \begin{cases} p_n 01 & \text{if } n \text{ is odd} \\ p_n 10 & \text{if } n \text{ is even.} \end{cases}$$

*Proof.* (i) First,

$$s_n s_{n-1} = s_{n-1}^{d_n} s_{n-2} s_{n-1} = s_{n-1} t_n.$$

Next

$$\begin{aligned} s_{n-1} s_n &= s_{n-1}^{d_n} s_{n-1} s_{n-2} \\ &= s_{n-1}^{d_n} s_{n-2} t_{n-1} = s_n t_{n-1} \end{aligned}$$

(ii) Since  $s_0 = 0$  and  $s_{-1} = 1$ , the equations hold for  $n = 0$ . Also,

$$s_1 s_0 = s_0^{d_0} 10 = p_0 10, \quad s_0 s_1 = s_0^{d_0} 01 = p_0 01$$

Next, for  $n \geq 2$  and even,

$$s_n s_{n-1} = s_{n-1}^{d_n} s_{n-1} s_{n-2} = s_{n-1}^{d_n} p_{n-1} 01$$

and since  $p_n = s_{n-1}^{d_n} p_{n-1}$ , one gets the first formula. The other equations are verified in the same manner.  $\square$

**Corollary 1.** *The words  $s_n$  and  $t_n$  differ only by their last two letters.*

*Proof.* In view of the previous lemma

$$\begin{aligned} s_{n+1} &= s_n^{d_{n+1}-1} s_n s_{n-1} = s_n^{d_{n+1}-1} p_n ab \\ t_{n+1} &= s_n^{d_{n+1}-1} s_{n-1} s_n = s_n^{d_{n+1}-1} p_n ba \end{aligned}$$

where  $ab = 01$  or  $ab = 10$ . This proves the claim.  $\square$

*Example 3.* Consider again  $\alpha = (\sqrt{3}-1)/2 = [0, 2, 1, 2, 1, \dots]$  and its directive sequence  $(1, 1, 2, 1, 2, 1, \dots)$ . The sequences start with

$s_0 = 0$	$p_0 = \varepsilon$
$d_1 = 1 \quad s_1 = 01$	$p_1 = 0$
$d_2 = 1 \quad s_2 = 010$	$p_2 = 010$
$d_3 = 2 \quad s_3 = 01001001$	$p_3 = 010010010$
$d_4 = 1 \quad s_4 = 01001001010$	$p_4 = 01001001010010010$
$d_5 = 2 \quad s_5 = 0100100101001001001001001001001$	

The importance of the sequence  $p_n$  comes from the following observation. Consider the sequence of integers  $(q_n)$  defined by

$$q_{-1} = q_0 = 1 \quad q_{n+1} = d_{n+1} q_n + q_{n-1}$$

so that  $q_n$  is precisely the length of  $s_n$ .

**Proposition 3.** *The word  $p_{n+1}$  is the highest rational power of  $s_n$  that is a prefix of the standard word  $c_\alpha$ . The prefix index of  $s_n$  is  $1 + d_{n+1} + (q_{n-1} - 2)/q_n$ .*

*Proof.* Clearly,  $s_{n+1} s_n$  is a prefix of the characteristic word  $c_\alpha$ . Since

$$s_{n+1} s_n = s_n^{d_{n+1}} s_{n-1} s_n = s_n^{1+d_{n+1}} t_{n-1}$$

the word  $s_n^{1+d_{n+1}}$  is a prefix of  $c_\alpha$ . Observe that  $t_{n-1}$  is not a prefix of  $s_n$ . Indeed, the word  $s_{n-1}$  is prefix of  $s_n$  and has the same length as  $t_{n-1}$ . Thus the longest common prefix  $h_n$  of  $t_{n-1}$  and  $s_{n-1}$  has length  $q_{n-1} - 2$ , and since

$$s_{n+1} s_n = s_n^{1+d_{n+1}} t_{n-1} = p_{n+1} ab$$

the longest power of  $s_n$  that is prefix of  $c_\alpha$  is  $p_{n+1}$ . Since  $|p_{n+1}| = q_{n+1} + q_n - 2 = (d_{n+1} + 1)q_n + q_{n-1} - 2$ , the exponent of  $s_n$  is  $1 + d_{n+1} + (q_{n-1} - 2)/q_n$ .  $\square$

*Example 4.* Consider first the Fibonacci sequence, where all  $d_n$  are 1. The  $q_n$  are the Fibonacci numbers. The formula shows that the prefix index of  $s_n$  is  $q_{n+2}/q_n - 2/q_n$ , and since  $q_{n+2}/q_n \rightarrow \tau^2 = 1 + \tau$ , where  $\tau$  is the golden ratio, the prefix index is bounded by the sequence  $1 + \tau$ .

*Example 5.* Consider again  $\alpha = (\sqrt{3} - 1)/2 = [0, 2, 1, 2, 1, \dots]$ . Since its directive sequence has bounded values, the exponent of  $s_n$  as a prefix is bounded. It is not difficult to see that the sequence  $1 + d_{n+1} + (q_{n-1} - 2)/q_n$  has two accumulation points, namely  $2 + \sqrt{3}$  and  $1 + (\sqrt{3} - 1)/2$ , and thus the prefix index is bounded by  $3 + 2\alpha$ .

**Proposition 4.** *Any Sturmian word contains cubes. More precisely, the index of  $s_n$  as a factor in the characteristic word  $c_\alpha$  is at least  $2 + d_{n+1} + (q_{n-1} - 2)/q_n$ .*

*Proof.* Set  $\delta_n = 1 + d_{n+1} + (q_{n-1} - 2)/q_n$ . We show that  $s_{n+4}$  contains a power of  $s_n$  of exponent  $1 + \delta_n$ . Indeed,

$$s_{n+4} = s_{n+3}^{d_{n+4}-1} s_{n+3} s_{n+2} = s_{n+3}^{d_{n+4}-1} s_{n+2} t_{n+3}$$

The suffix  $s_{n+2} t_{n+3}$  of  $s_{n+4}$  contains the desired power. Indeed,  $s_{n+2}$  ends with  $s_n$ , and  $t_{n+3}$  shares a prefix of length  $q_{n+3} - 2$  with  $s_{n+3}$ . Now  $p_{n+1}$  is the prefix of  $c_\alpha$  of length  $q_{n+1} + q_n - 2$ , and since  $q_{n+1} + q_n < q_{n+3}$ ,  $s_n p_{n+1}$  is a factor of  $s_{n+2} t_{n+3}$ , and also of  $s_{n+4}$ .  $\square$

A weak converse also holds.

**Proposition 5.** *Let  $w$  be a primitive factor in  $c = c_\alpha$ , and assume that  $\text{ind}(w) \geq 4$ . Assume further that  $w$  has the maximal index among its conjugates. Then  $w$  is one of the  $s_n$ .*

*Proof.* Set  $1 + d = \text{ind}(w)$ , set  $w = za$  with  $a$  a letter, and let  $b$  be the letter preceding the occurrence of  $w^{d+1}$ . Then  $a \neq b$  since otherwise the conjugate  $az$  would have greater index. Thus  $aw^d$  and  $bw^d$  are factors of  $c$ . This means that  $w^d$  is a left special factor, and therefore it is a prefix of  $c$ .

Let  $n$  be the greatest integer such that  $s_n$  is a prefix of  $w^2$  (recall that  $2 \leq d$ ). Since  $s_n$  is primitive,  $s_n \neq w^2$ . If  $w$  is a prefix of  $s_n$ , then either  $w = s_n$  and the proposition is proved, or there is a factorization  $w = uv$ , for non empty  $u, v$ , such that  $s_n = wu$ . Next,  $s_n w$  is a prefix of  $w^3$ . Thus  $s_n w z = w^3$  for some word  $z$ . It follows from  $s_n w z = w u v z$  and  $w^3 = w u v u v$  that  $uv = vu$ , which is impossible because  $w$  is primitive.

Thus  $s_n$  is a proper prefix of  $w$ . Now,  $w^2$  is a proper prefix of  $s_{n+1}$ , thus also of  $p_{n+1}$ ; thus  $w^2 z = s^k$  for some  $k > 2$ . Thus  $w$  and  $s_n$  are powers of the same word, and since they are primitive, they are equal.  $\square$

*Proof of Theorem 1.* Since a Sturmian word has the same factors as the characteristic word of same slope, it suffices to prove the result for characteristic

words. Let  $c$  be the characteristic word of slope  $\alpha = [0, 1 + d_1, d_2, \dots]$ . Let  $(s_n)_{n \geq -1}$  be the associated standard sequence.

To prove that the condition is sufficient, observe that  $s_n^{d_{n+1}}$  is a prefix of  $c$  for each  $n \geq 1$ . Consequently, if the sequence  $(d_n)$  of partial quotients is unbounded, the infinite word  $c$  has factors of arbitrarily great index.

Conversely, assume that  $c$  has unbounded index. Then there are words  $w$  of arbitrarily high index. By the preceding proposition, there are, in the standard sequence, words  $s_n$  of arbitrarily high prefix index. Since the prefix index of  $s_n$  is  $1 + d_{n+1} + (q_{n-1} - 2)/q_n$ , this means that the partial quotients  $d_{n+1}$  are unbounded. This completes the proof.  $\square$

## 4 Concluding remark

There might be a more precise correspondence between the factors and the prefixes of the characteristic word. Also, the precise value of the index of a Sturmian word seems to be more complicated to compute in the general case as for the Fibonacci word.

## References

1. J. Berstel, Recent results on Sturmian words, in *Developments in Language Theory*, J. Dassow et A. Salomaa (éd.), World Scientific, 1996.
2. J. Berstel, P. Séébold, Sturmian Words, in: M. Lothaire *Algebraic Combinatorics on Words*, in preparation.
3. J. Berstel et P. Séébold, A remark on morphic Sturmian words, *Informatique théorique et applications* **28** (1994), 255–263.
4. E. Bombieri, J. E. Taylor, Which distributions of matter diffract? An initial investigation, *J. Phys.* **47** (1986), Colloque C3, 19–28.
5. J. E. Bresenham, Algorithm for computer control of a digital plotter, *IBM Systems J.* **4** (1965), 25–30.
6. T. C. Brown, A characterization of the quadratic irrationals, *Canad. Math. Bull.* **34** (1991), 36–41.
7. D. Crisp, W. Moran, A. Pollington, P. Shiue, Substitution invariant cutting sequences, *J. Théorie des Nombres de Bordeaux* **5** (1933), 123–137.
8. E. Coven, G. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973), 138–153.
9. A. De Luca, Sturmian words: structure, combinatorics, and their arithmetics, *Theoret. Comput. Sci.* **183** (1997), 45–82.
10. A. De Luca, Combinatorics of standard Sturmian words, in: J. Mycielski, G. Rozenberg, A. Salomaa (eds.) *Structures in Logic and Computer Science, Lect. Notes Comp. Sci.* Vol. 1261, Springer-Verlag, 1997, pp 249–267.
11. A. De Luca, Standard Sturmian morphisms, *Theoret. Comput. Sci.* **178** (1997), 205–224.
12. A. De Luca et F. Mignosi, Some combinatorial properties of Sturmian words, *Theoret. Comput. Sci.* **136** (1994), 361–385.

13. S. Dulucq, D. Gouyou-Beauchamps, Sur les facteurs des suites de Sturm, *Theoret. Comput. Sci.* **71** (1990), 381–400.
14. A. S. Fraenkel, M. Mushkin, U. Tassa, Determination of  $[n\theta]$  by its sequence of differences, *Canad. Math. Bull.* **21** (1978), 441–446.
15. G.A. Hedlund, Sturmian minimal sets, *Amer. J. Math* **66** (1944), 605–620.
16. M. Morse, G.A. Hedlund, Symbolic dynamics, *Amer. J. Math* **60** (1938), 815–866.
17. M. Morse, G.A. Hedlund, Sturmian sequences, *Amer. J. Math* **61** (1940), 1–42.
18. S. Ito, S. Yasutomi, On continued fractions, substitutions and characteristic sequences, *Japan. J. Math.* **16** (1990), 287–306.
19. J. Karhumäki, On strongly cube-free  $\omega$ -words generated by binary morphisms, in *FCT '81*, pp. 182–189, *Lect. Notes Comp. Sci.* Vol. 117, Springer-Verlag, 1981.
20. J. Karhumäki, On cube-free  $\omega$ -words generated by binary morphisms, *Discr. Appl. Math.* **5** (1983), 279–297.
21. F. Mignosi, On the number of factors of Sturmian words, *Theoret. Comput. Sci.* **82** (1991), 71–84.
22. F. Mignosi et G. Pirillo, Repetitions in the Fibonacci infinite word, *Theoret. Inform. Appl.* **26**,3 (1992), 199–204.
23. F. Mignosi, P. Séébold, Morphismes sturmiens et règles de Rauzy, *J. Théorie des Nombres de Bordeaux* **5** (1993), 221–233.
24. M. Queffélec, *Substitution Dynamical Systems – Spectral Analysis*, Lecture Notes Math., vol. 1294, Springer-Verlag, 1987.
25. G. Rauzy, Suites à termes dans un alphabet fini, *Sémin. Théorie des Nombres* (1982–1983), 25-01,25-16, Bordeaux.
26. G. Rauzy, Mots infinis en arithmétique, in: *Automata on infinite words* (D. Perrin ed.), *Lect. Notes Comp. Sci.* **192** (1985), 165–171.
27. P. Séébold, Fibonacci morphisms and Sturmian words, *Theoret. Comput. Sci.* **88** (1991), 367–384.
28. C. Series, The geometry of Markoff numbers, *The Mathematical Intelligencer* **7** (1985), 20–29.
29. K. B. Stolarsky, Beatty sequences, continued fractions, and certain shift operators, *Canad. Math. Bull.* **19** (1976), 473–482.
30. B. A. Venkov, *Elementary Number Theory*, Wolters-Noordhoff, Groningen, 1970.