Generalized balances in Sturmian words

Isabelle Fagnot* and Laurent Vuillon†

December 1, 2000

Abstract

One of the numerous characterizations of Sturmian words is based on the notion of *balance*. An infinite word x on the $\{0, 1\}$ alphabet is balanced if, given two factors of x, w and w', having the same length, the difference between the number of $0'$s in w (denoted by $|w|_0$) and the number of $0'$s in w' is at most 1, i.e. $||w|_0 - |w'|_0| \leq 1$. It is well known that an aperiodic word is Sturmian if and only if it is balanced.

In this paper, the balance notion is generalized by considering the number of occurrences of a word u in w (denoted by $|w|_u$) and w'. The following is obtained

Theorem *Let x be a Sturmian word. Let u, w and w' be three factors of x. Then,*

$$|w| = |w'| \implies ||w|_u - |w'|_u| \leq |u|.$$

Another balance property, called equilibrium, is also given. This notion permits us to give a new characterization of Sturmian words. The main techniques used in the proofs are word graphs and return words.

1 Introduction

Sturmian words are infinite words over a binary alphabet with exactly $n+1$ factors of length n, for each $n \geq 0$. One of the numerous characterizations of Sturmian words is based on the notion of *balance*. An infinite word x on the $\{0, 1\}$ alphabet is balanced if, given two factors of x, w and w', having the same length, the difference between the number of $0'$s in w (denoted by $|w|_0$) and the number of $0'$s in w' is at most 1, i.e. $||w|_0 - |w'|_0| \leq 1$. It is well known that an aperiodic word is Sturmian if and only if it is balanced (Hedlund and Morse [10]).

The notion of balance is important in Sturmian words theory and in number theory. In particular, the structure of aperiodic balanced words in a finite alphabet containing more than 3 letters is closely related to Sturmian words (Graham [9]). In addition, the covering of integers by more than three disjoint sets of the form

$$([\alpha_i n + \beta_i])_{n \in \mathbb{N}}$$

(where all the α_i are different) leads to periodic balanced words (Tijdeman [18]). Furthermore, the balanced words appear in computer science for allocation sequences of two processes sharing a resource and in the heap model with two pieces

*Institut Gaspard Monge, Cité Descartes, 5, boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-vallée cedex 2, France, Email: fagnot@univ-mlv.fr

†LIAFA Université Paris 7, 2, place Jussieu, Case 7014, 75251 Paris Cedex 05, France, Email: Laurent.Vuillon@liafa.jussieu.fr

(Gaujal [8], Mairesse and Vuillon [14]). Recently, a paper of Cassaigne, Ferenczi and Zamboni [6] illustrates how the presence of balances is intimately connected with the underlying geometry: an Arnoux-Rauzy sequence which is totally unbalanced in the sense of Cassaigne-Ferenczi-Zamboni cannot be a natural coding of a rotation on a torus.

Berthé and Tijdeman [4] consider balance in multi-dimensional words and prove the associated double sequence to be fully periodic.

A way to extend the balance property is to consider the number of occurrences of a word u in w (denoted by $|w|_u$) and in w', both words being factors of a Sturmian word and having same length. The difference of the numbers of occurrences is studied and it is shown that it is less than $|u|$. (see Theorem 12).

More precisely, the following result is obtained (Proposition 11). (Here, we denote $\delta(u) = \max\{||v|_u - |v'|_u| \mid v, v' \in L(\mathbf{x}), |v| = |v'|\}$.)

Proposition *Let $\mathbf{x}$ be a Sturmian word and u be a factor of $\mathbf{x}$. Three cases appear.*

i. if u is non-overlapping, then $\delta(u) \leq 2$;

ii. if v is the period of u, $u = v^r$ and $v^{r+1} \notin L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then

  (a) if $|v| = 1$, then $\delta(u) \leq 2$;

  (b) otherwise, $\delta(u) \leq 3$;

iii. if v is the period of u, $u = v^r$ and $v^{r+1} \in L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then

  (a) if $|v| = 1$, then $\delta(u) \leq \max(r, 2)$;

  (b) if $|v| = 2$, then $\delta(u) \leq r + 1$;

  (c) otherwise, $\delta(u) < r + 2$.

In this proposition, all the bounds are reached except for the case ii (b). But, we conjecture that $\delta(u) \leq 2$ in this latter case. Proposition 11 implies, in particular, that a Sturmian word whose slope has bounded coefficients in its continued fraction, has a bounded balance too.

A former result of Ostrowski [15, 13] implied that $\delta(u) \leq 2|u|$ in the general case and $\delta(u) \leq c \ln(|u|)$, with $c \in \mathbb{N}$, when the slope has bounded coefficients in its continued fraction expansion. This result is based on rotations on the unit circle and continued fraction techniques. We therefore improve these bounds using totally different means.

The generalized balance property is related to the following notion. Consider two factors z and z' of a Sturmian word such that $z = uvu$, $z' = uv'u$ and $|z|_u = |z'|_u = n$, with $n \geq 2$. The difference of lengths, $|z| - |z'|$, is called the equilibrium of the factors and Theorem 7 states that the equilibrium is bounded by the length of u (*i.e.* $||z| - |z'|| \leq |u|$). Furthermore, the equilibrium for the case where u is equal to the letter 1, permits us to give a new characterization of Sturmian words.

The article is organized as follows. Section 2 contains basic definitions and notations in combinatorics of words. Section 3 recalls some facts about Sturmian words and return words. In Section 4, it is shown that the derived word of a Sturmian word is also Sturmian. Sections 5 deals with the relative lengths of return words. Section 6 and 7 establish the main theorem using return words and combinatorics on words.

2 Definitions and notations

Let A be a finite alphabet $\{0,1\}$. The set of finite words is denoted by A^* and the set of infinite words by A^ω. The empty word is denoted by ε. Given u a finite word, its length is denoted by $|u|$.

Given $r \in \mathbb{N}$ and $u \in A^*$, we denote $\mathrm{Pref}_r(u)$ the prefix of u of length r if $|u| \geq r$, otherwise u. Likewise, we denote $\mathrm{Suff}_r(u)$ the suffix of u of length r, if $|u| \geq r$, otherwise u.

Let $\mathbf{x} = a_0 a_1 \cdots a_n (\cdots)$ be a finite or infinite word over A. For integers $i \leq j$, we define $\mathbf{x}[i,j) = a_i a_{i+1} \cdots a_{j-1}$ and $\mathbf{x}[i,j] = a_i a_{i+1} \cdots a_j$. The set of all finite factors of $\mathbf{x}$ is denoted by $L(\mathbf{x})$, i.e.

$$L(\mathbf{x}) = \{\mathbf{x}[i,j) \mid 0 \leq i \leq j\}.$$

Let u be a factor of a word w (finite or infinite). If there exist two words α and β such that $w = \alpha u \beta$, then the integer $|\alpha|$ is said to be an *occurrence* of u in w. The number of occurrences of a word v in u is denoted by $|u|_v$. An infinite word is said to be recurrent if for each factor u of $\mathbf{x}$, there are an infinite number of occurrences of u in $\mathbf{x}$.

We define the *shift operator* on infinite words, σ, as follows. If $\mathbf{x} = a_0 a_1 \cdots a_n \cdots$ is an infinite word, then $\sigma(\mathbf{x}) = a_1 \cdots a_n \cdots$. Of course, $\sigma^k(\mathbf{x}) = a_k a_{k+1} \cdots a_n \cdots$.

Let v be a finite word and r be a rational number such that $r|v|$ is an integer. We denote v^r the word $v^{\lfloor r \rfloor} \cdot v[0, \{r\}|v|)$, where $\lfloor r \rfloor$ denotes the integer part of r and $\{r\}$ its fractional part. Let u be a finite word. We say that v is *a (rational) period* of u if $u = v^r$ for some $r \in \mathbb{Q}$, and v is called *the period* of u if it is the smallest period of u. If $r \in \mathbb{N}$, the word v is said to be a *integral period* of u.

Let u be a finite word. It is said to be *overlapping* if there exist two words p and s such that $0 < |p| = |s| < |u|$ and $pu = us$. It is not difficult to see that if a word u is overlapping, it has a period v with $|v| < |u|$.

Let u be a factor of an infinite word $\mathbf{x}$ and a be a letter. We say that ua is an *(right) extension* of u if ua is also a factor of $\mathbf{x}$. Symmetrically, we say that au is an *left extension* of u if au is also a factor of $\mathbf{x}$. Obviously, if we consider infinite words over a two-letter alphabet, a factor has one or two extensions.

3 Previous results

3.1 Sturmian words

There are many definitions and properties related to Sturmian words. Here, we only recall those we are going to use. For more information about Sturmian words the reader is referred to the survey of Berstel and Séébold [2].

An infinite word on $\{0,1\}$ is said to be *balanced* if, for any two factors v and w, we have

$$(|v| = |w|) \Rightarrow (||v|_0 - |w|_0| \leq 1).$$

An infinite word $\mathbf{x}$ is *Sturmian* if it is balanced and non-periodic.

Let $\mathbf{x}$ be a Sturmian word. There exists an integer $k \geq 1$, such that $\mathbf{x}$ has one of the following two forms

$$\begin{aligned} \mathbf{x} &= 0^i 1 0^{k_1} 1 0^{k_2} \cdots 1 0^{k_p} \cdots \\ \text{or } \mathbf{x} &= 1^i 0 1^{k_1} 0 1^{k_2} \cdots 0 1^{k_p} \cdots, \end{aligned}$$

where $0 \leq i \leq k+1$ and $k_p = k$ or $k_p = k+1$.

A Sturmian word $\mathbf{x}$ is *uniformly recurrent*, i.e. given any factor u of $\mathbf{x}$, it has an infinite number of occurrences and the distance between two successive occurrences of u is bounded. If $\mathbf{x}$ is a Sturmian word, then the word $\mathbf{y}$ obtained from $\mathbf{x}$ by replacing 0 by 1 and 1 by 0 is also Sturmian.

The *slope* of a Sturmian word is the real α, $0 < \alpha < 1$, defined by

$$\alpha = \lim_{n \to +\infty} \frac{|\mathbf{x}_n|_1}{|\mathbf{x}_n|},$$

where $\mathbf{x}_n = \mathbf{x}[0, n)$. The slope always exists (see [2]).

The factors of a Sturmian word are dependent of its slope. More precisely, we have the following

Proposition 1 *(Berstel and Séébold [2]) If $\mathbf{x}$ and $\mathbf{y}$ are two Sturmian words of slopes α and β respectively, then $\alpha = \beta$ if and only if $L(\mathbf{x}) = L(\mathbf{y})$.*

Moreover, the partial quotients of the continued fraction of the slope give us some information on the repetitions in the Sturmian word. Indeed, given an infinite word $\mathbf{x}$ and u a factor of $\mathbf{x}$, we define the *index* of u in $\mathbf{x}$ as the greatest integer d such that u^d is a factor of $\mathbf{x}$. The word $\mathbf{x}$ has *bounded index* if there exists an integer d such that for every factor u of $\mathbf{x}$, the index of u is less or equal to d.

Proposition 2 *(Mignosi [16]) Let $\mathbf{x}$ be a Sturmian word. Let α the slope of $\mathbf{x}$ and let $\alpha = [0, a_1, a_2, \ldots, a_n, \ldots]$ be its continued fraction expansion. The word $\mathbf{x}$ has bounded index if and only if the partial quotients $(a_n)_{n \in \mathbb{N}}$ are bounded.*

Another characterization of Sturmian words is based on *complexity*, i.e. the numbers of different factors of a given length.

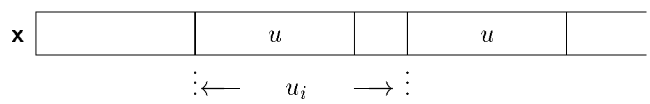
Proposition 3 *(Hedlund and Morse [10]) An infinite word is Sturmian if and only if for each $n \in \mathbb{N}$, there are exactly $n+1$ different factors of length n.*

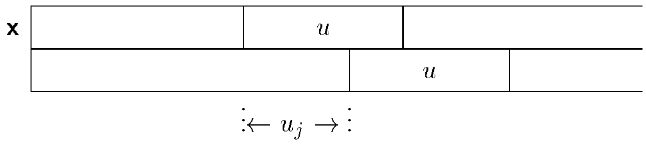
Other equivalent characterizations of Sturmian words are mainly about representations of straight lines and rotations over the unit circle.

3.2 Return words and derived words

The notions of return words and derived words were introduced by Durand [7] and Holton and Zamboni [11].

Let $\mathbf{x}$ be an infinite word and u be a recurrent factor of $\mathbf{x}$ of length ℓ. The factor v is a *return word* of u if there exist $i, j \in \mathbb{N}, i < j$, such that $v = \mathbf{x}[i, j)$, $\mathbf{x}[i, i+\ell) = \mathbf{x}[j, j+\ell) = u$ and $|\mathbf{x}[i, j+\ell)|_u = 2$. In other words, we define the set of return words of u to be the set of all distinct words beginning with an occurrence of u and ending exactly before the next occurrence of u in the recurrent word $\mathbf{x}$ (see examples below). We denote it by $\mathcal{H}_u(\mathbf{x})$ as in Durand [7].





For example, if $\mathbf{x}$ is the well known Fibonacci word, $\mathbf{x} = 0100101001001010010100100101001001\cdots$, we have $\mathcal{H}_\varepsilon(\mathbf{x}) = \{0, 1\}$, $\mathcal{H}_0(\mathbf{x}) = \{0, 01\}$ and $\mathcal{H}_{1001}(\mathbf{x}) = \{100, 10010\}$. In the latter example, we can see that a return word is not necessarily longer than the factor.

It is not difficult to see that if v is a return word of u, then vu is a factor of $\mathbf{x}$ and has u as prefix.

Obviously, the set $\mathcal{H}_u(\mathbf{x})$ is finite if and only if the distance between two successive occurrences of u is bounded, i.e. if $\mathbf{x}$ is uniformly recurrent. Suppose that $\mathcal{H}_u(\mathbf{x}) = \{u_1, u_2, \cdots u_n\}$. There exist a unique sequence of integers $(i_k)_{k\geq 1}$ and a unique word α such that $\mathbf{x} = \alpha u_{i_1} u_{i_2} \cdots u_{i_k} \cdots$ and such that $|\alpha|$ is the first occurrence of u in $\mathbf{x}$. The word $i_1 i_2 \cdots i_k \cdots$ is called a *derived word* of $\mathbf{x}$ with respect to u .

Obviously, the derived word obtained depends on the injective map $f : \mathcal{H}_u(\mathbf{x}) \to \mathbb{N}$. However, as all the derived words are images of each other by a letter-to-letter bijection and as the roles of 0 and 1 are symmetrical in Sturmian word theory, we will denote it ***the** derived word* $\mathcal{D}_u(\mathbf{x})$ of $\mathbf{x}$ with respect to u. In the previous example, as we have $\mathcal{D}_\varepsilon(\mathbf{x}) = \mathbf{x}$, $\mathcal{D}_0(\mathbf{x}) = 101101011010\cdots$ and $\mathcal{D}_{1001}(\mathbf{x}) = 10110101\cdots$.

4 Derived words

In this section, we are going to show that deriving a Sturmian word produces a Sturmian word. To this end, we first remind a previous result due to Berstel and Séébold.

Take E, φ and $\tilde{\varphi}$ the following morphisms on $\{0, 1\}^*$:

$$E : \begin{array}{ccc} 0 & \to & 1 \\ 1 & \to & 0 \end{array}, \qquad \varphi : \begin{array}{ccc} 0 & \to & 01 \\ 1 & \to & 0 \end{array}, \qquad \tilde{\varphi} : \begin{array}{ccc} 0 & \to & 10 \\ 1 & \to & 0 \end{array}.$$

Proposition 4 *(Berstel and Séébold [2]) Let $\mathbf{x}$ be an infinite word on $\{0, 1\}$.*

i. If $\tilde{\varphi}(\mathbf{x})$ is Sturmian and $\mathbf{x}$ starts with the letter 0, then $\mathbf{x}$ is Sturmian.

ii. Let f be a morphism that is a composition of E and φ. If $f(\mathbf{x})$ is Sturmian, then $\mathbf{x}$ is Sturmian.

Now we can state the promised result.

Proposition 5 *Let $\mathbf{x}$ be an infinite Sturmian word. For each factor u of $\mathbf{x}$, there are two and only two return words of u. Moreover, the derived word $\mathcal{D}_u(\mathbf{x})$ is also a Sturmian word.*

Remark The fact that there are exactly two return words was already shown by one of the authors in [19].

Proof We show both properties by induction on the length of u in the same time.

The base case: $u = \varepsilon$. As both 0 and 1 appear in $\mathbf{x}$, we have $\mathcal{H}_\varepsilon(\mathbf{x}) = \{0, 1\}$ and $\mathcal{D}_\varepsilon(\mathbf{x}) = \mathbf{x}$ which is clearly Sturmian too.

Now, let us consider the induction. Let $u = va$, where a is a letter. There are two cases:

i. If v has only one extension, then the occurrences of u are exactly those of v. Consequently, the return words of u are the same as those of v, i.e. $\mathcal{H}_u(\mathbf{x}) = \mathcal{H}_v(\mathbf{x})$. Obviously, we have also $\mathcal{D}_u(\mathbf{x}) = \mathcal{D}_v(\mathbf{x})$ and, consequently, by induction hypothesis, $\mathcal{D}_u(\mathbf{x})$ is a Sturmian word.

ii. If both words $v0$ and $v1$ are factors of $\mathbf{x}$, let v_1 and v_2 be the two return words of v. By definition v_1v and v_2v are factors of $\mathbf{x}$ and have the prefix v. Thus, there exist two words t_1 and t_2 such that

$$\begin{aligned} v_1v &= vt_1 \\ v_2v &= vt_2. \end{aligned}$$

As v has two extensions, we have $\mathrm{Pref}_1(t_1) \neq \mathrm{Pref}_1(t_2)$. Suppose that $a = \mathrm{Pref}_1(t_1)$. The occurrences of $u = va$ are thus those of v_1v, that is, if to compute $\mathcal{D}_v(\mathbf{x})$ we replace v_1 by 0 and v_2 by 1, the occurrences of u correspond to the $0'$s in $\mathcal{D}_v(\mathbf{x})$.

By induction hypothesis, we know that $\mathcal{D}_v(\mathbf{x})$ is a Sturmian word. Again there are two cases:

(a) There exists $k \geq 1$ such that $\mathcal{D}_v(\mathbf{x}) = 0^i10^{k_1}10^{k_2}\cdots10^{k_p}\cdots$ where $0 \leq i \leq k+1$ and $k_p = k$ or $k_p = k+1$. Therefore $\mathbf{x}$ has the form $wv_1^iv_2v_1^{k_1}v_2v_1^{k_2}\cdots v_2v_1^{k_p}\cdots$, where $|w|$ is the first occurrence of v. Then $\mathcal{H}_u(\mathbf{x}) = \{v_1, v_1v_2\}$. If to compute $\mathcal{D}_u(\mathbf{x})$, we replace v_1 by 1 and v_1v_2 by 0, we have $\mathcal{D}_u(\mathbf{x}) = 1^{i-1}01^{k_1-1}01^{k_2-1}\cdots01^{k_p-1}\cdots$ if $i > 0$, and $\mathcal{D}_u(\mathbf{x}) = 1^{k_1-1}01^{k_2-1}\cdots01^{k_p-1}\cdots$ otherwise. Then $\varphi(\mathcal{D}_u(\mathbf{x})) = \sigma^p(\mathcal{D}_v(\mathbf{x}))$ where $p = 0$ or $p = 1$. Since $\sigma^p(\mathcal{D}_v(\mathbf{x}))$ is clearly Sturmian, we have by Proposition 4(ii) that $\mathcal{D}_u(\mathbf{x})$ is Sturmian too.

(b) There exists $k \geq 1$ such that $\mathcal{D}_v(\mathbf{x}) = 1^i01^{k_1}01^{k_2}\cdots01^{k_p}\cdots$ where $0 \leq i \leq k+1$ and $k_p = k$ or $k_p = k+1$. Therefore $\mathbf{x}$ has the form $wv_2^iv_1v_2^{k_1}v_1v_2^{k_2}\cdots v_1v_2^{k_p}\cdots$, where $|w|$ is the first occurrence of v. Then $\mathcal{H}_u(\mathbf{x}) = \{v_1v_2^k, v_1v_2^{k+1}\}$. If to compute $\mathcal{D}_u(\mathbf{x})$ we replace $v_1v_2^k$ by 0 and $v_1v_2^{k+1}$ by 1, we have $\mathcal{D}_u(\mathbf{x}) = x_{k_1}x_{k_2}\cdots x_{k_p}\cdots$, where $x_{k_p} = 0$ if $k_p = k$ and $x_{k_p} = 1$ if $k_p = k+1$.

Let f be the morphism:

$$f : \begin{array}{lcl} 0 & \rightarrow & 01^k \\ 1 & \rightarrow & 01^{k+1}. \end{array}$$

We have that $f(\mathcal{D}_u(\mathbf{x})) = \sigma^i(\mathcal{D}_v(\mathbf{x}))$, but we cannot conclude so easily.

Let g and h be the following Sturmian morphisms

$$g = E \circ \tilde{\varphi} : \begin{array}{lcl} 0 & \rightarrow & 01 \\ 1 & \rightarrow & 1 \end{array}, \qquad h = \varphi \circ E : \begin{array}{lcl} 0 & \rightarrow & 0 \\ 1 & \rightarrow & 01. \end{array}$$

First, we can check that $f = g^k \circ h$. We have obviously that $h(\mathcal{D}_u(\mathbf{x}))$ starts with the letter 0, and then for all ℓ, the word $g^\ell \circ h(\mathcal{D}_u(\mathbf{x}))$ starts with 0 too.

Thus we have $E \circ \left(\tilde{\varphi} \circ g^{k-1} \circ h(\mathcal{D}_u(\mathbf{x}))\right) = f(\mathcal{D}_u(\mathbf{x})) = \sigma^i(\mathcal{D}_v(\mathbf{x}))$ which implies that $\tilde{\varphi} \circ g^{k-1} \circ h(\mathcal{D}_u(\mathbf{x}))$ is Sturmian (Proposition 4 (ii)). As $g^{k-1} \circ$

$h(\mathcal{D}_u(\mathbf{x}))$ starts with 0 (see below), by Proposition 4 (i), we have that $g^{k-1} \circ h(\mathcal{D}_u(\mathbf{x}))$ is Sturmian. By induction, we can prove similarly that $g^\ell \circ h(\mathcal{D}_u(\mathbf{x}))$ is Sturmian for $\ell, 0 \le \ell \le k$. Thus $h(\mathcal{D}_u(\mathbf{x}))$ is Sturmian and by Proposition 4 (ii), $\mathcal{D}_u(\mathbf{x})$ is Sturmian. This completes the induction.

∎

5 Relative lengths of return words

In this section we use the word graph associated with the factors of a Sturmian word $\mathbf{x}$ (see Arnoux and Rauzy [1], Berstel and Séébold [2], Berthé [3] and Cassaigne [5]) in order to study the relative lengths of return words.

We begin by stating some notations about word graphs (for more information see Arnoux and Rauzy [1]).

In the graph of length n, the vertices are words of length n. There is an edge between the vertices u and v if and only if there exist two letters a and b such that ua and bv are factors of $\mathbf{x}$ and $ua = bv$ (we label the edge by a, $u \to_a v$). As $\mathbf{x}$ is a Sturmian word, there exists for each n a unique word R_n (resp. L_n) of length n with two right extensions (resp. with two left extensions). The other words have a unique right extension (resp. left extension).

Consequently, the word graph for Sturmian words is composed by three paths: the first and the second ones from R_n to L_n, the third one from L_n to R_n The first path is

$$R_n \to_{a_1} f_1 \to_{a_2} f_2 \to \cdots f_{\ell_1 - 1} \to_{a_{\ell_1}} L_n,$$

with length equal to ℓ_1.

The second path is

$$R_n \to_{b_1} g_1 \to_{b_2} g_2 \to \cdots g_{\ell_2 - 1} \to_{b_{\ell_2}} L_n,$$

with length equal to ℓ_2.

The third path is from L_n to R_n, namely

$$L_n \to_{c_1} h_1 \to_{c_2} h_2 \to \cdots h_{\ell_3 - 1} \to_{c_{\ell_3}} R_n$$

with length equal to ℓ_3.

By construction, we have $\ell_1 \ge 1$ and $\ell_2 \ge 1$. The third path has length 0 if $L_n = R_n$ (see [1, 3, 5] for general properties on word graphs associated with Sturmian words).

Now, we are ready to state the proposition.

Proposition 6 *Let $\mathbf{x}$ be a Sturmian word. Assume that $\mathcal{H}_u = \{u_1, u_2\}$ with u a factor of $\mathbf{x}$. Then*

$$||u_1| - |u_2|| \le |u|.$$

Proof This proof has the same structure as the proof of Proposition 5. (In particular, we use intermediate results about return words of Proposition 5.)

We show the proposition by induction on the length of u. Let G be the word graph of length $|u|$.

The base case: $u = \varepsilon$. By definition, $\mathcal{H}_\varepsilon(\mathbf{x}) = \{0, 1\}$. We find that $||u_1| - |u_2|| = ||0| - |1|| = 0 = |\varepsilon|$.

Suppose that $u = va$, where a is a letter.

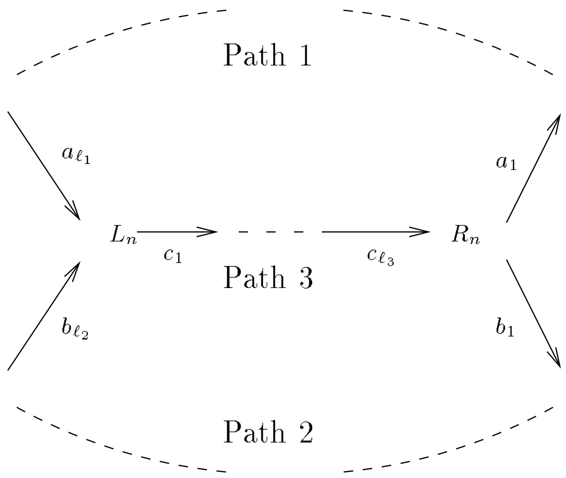


Figure 1: Graph of words.

- If v has only one extension, then $\mathcal{H}_u(\mathbf{x}) = \mathcal{H}_v(\mathbf{x})$. Consequently, by induction hypothesis, $||u_1| - |u_2|| \leq |v| < |u|$.

- If v has two extensions, then both words $v0$ and $v1$ are factors of $\mathbf{x}$. Let v_1 and v_2 the two return words of v. By an argument in the proof of Proposition 5, either $\mathcal{H}_u(\mathbf{x}) = \{v_1, v_1v_2\}$ or $\mathcal{H}_u(\mathbf{x}) = \{v_1v_2^k, v_1v_2^{k+1}\}$. In both cases, we have $||u_1| - |u_2|| = |v_2|$. Thus, it is sufficient to show that $|v_2| \leq |u| = |v| + 1$ to prove the statement.

  Consider G the word graph of length $|v|$. As v has two extensions, we find that $R_{|v|} = v$. By construction of the return words, one return word is given by the concatenation of the labels of the path

  $$R_{|v|} \rightarrow_{a_1} f_1 \rightarrow_{a_2} f_2 \rightarrow \cdots f_{\ell_1-1} \rightarrow_{a_{\ell_1}} L_{|v|} \rightarrow_{c_1} h_1 \rightarrow_{c_2} h_2 \rightarrow \cdots h_{\ell_3-1} \rightarrow_{c_{\ell_3}} R_{|v|}.$$

  The other return word is given by the concatenation of the label of the path

  $$R_{|v|} \rightarrow_{b_1} g_1 \rightarrow_{b_2} g_2 \rightarrow \cdots g_{\ell_2-1} \rightarrow_{b_{\ell_2}} L_{|v|} \rightarrow_{c_1} h_1 \rightarrow_{c_2} h_2 \rightarrow \cdots h_{\ell_3-1} \rightarrow_{c_{\ell_3}} R_{|v|}$$

  In consequence, $|v_2| \leq \max(\ell_1 + \ell_3, \ell_2 + \ell_3)$.

  Recall that, in the graph of length $|v|$, the vertices are words of length $|v|$ and there are exactly $|v| + 1$ such vertices. (Indeed, in a Sturmian word, the number of distinct words with length n is $n + 1$.) Furthermore, the number of edges is $|v| + 2$. (Because the number of distinct words with length $n + 1$ is $n + 2$.) That is, $\ell_1 + \ell_2 + \ell_3 = |v| + 2$. By construction, $\ell_1 \geq 1$ and $\ell_2 \geq 1$.

  In consequence, $|v_2| \leq \max(\ell_1 + \ell_3, \ell_2 + \ell_3) \leq |v| + 1$. Thus, we are through.

■

6 Equilibrium

The following lemma deals with the relative lengths of words z having the following property: $\mathrm{Pref}_{|u|}(z) = \mathrm{Suff}_{|u|}(z) = u$ and $|z|_u = n$. The relative lengths $|z| - |z'|$ is called the equilibrium of the factors.

Theorem 7 *Let $\mathbf{x}$ be a Sturmian word. Let u be a factor of $\mathbf{x}$ and $n \geq 2$ be an integer. Given two factors z and z' of $\mathbf{x}$ such that $z = uvu$, $z' = uv'u$ and $|z|_u = |z'|_u = n$, then*

$$||z| - |z'|| \leq |u|.$$

Proof Let $\mathcal{H}_u(\mathbf{x}) = \{u_1, u_2\}$. There are two sequences $(i_k)_{1 \leq k \leq n-1}$ and $(j_k)_{1 \leq k \leq n-1}$, with $i_k = 1$ or 2 and $j_k = 1$ or 2, such that $z = u_{i_1} u_{i_2} \cdots u_{i_{n-1}} u$ and $z' = u_{j_1} u_{j_2} \cdots u_{j_{n-1}} u$.

Take

$$\begin{aligned} v &= (i_1 - 1)(i_2 - 1) \cdots (i_{n-1} - 1) \\ v' &= (j_1 - 1)(j_2 - 1) \cdots (j_{n-1} - 1). \end{aligned}$$

Both words v and v' are factors of the derived word $\mathcal{D}_u(\mathbf{x})$ which is a Sturmian word by Theorem 5. Consequently, they are balanced, i.e.

$$||v|_1 - |v'|_1| = ||v|_2 - |v'|_2| \leq 1.$$

Thus, in the one hand,

$$|\sharp\{k \mid i_k = 1\} - \sharp\{k \mid j_k = 1\}| = |\sharp\{k \mid i_k = 2\} - \sharp\{k \mid j_k = 2\}| \leq 1,$$

where $\sharp A$ denotes the cardinal of the set A. In the other hand, we have clearly

$$|z| = \sharp\{k \mid i_k = 1\} \cdot |u_1| + \sharp\{k \mid i_k = 2\} \cdot |u_2| + |u|$$

and

$$|z'| = \sharp\{k \mid j_k = 1\} \cdot |u_1| + \sharp\{k \mid j_k = 2\} \cdot |u_2| + |u|.$$

Therefore

$$||z| - |z'|| \leq ||u_1| - |u_2|| \leq |u|.$$

Which is the desired relation. ∎

We can state a new characterization of Sturmian words based on the equilibrium property. Let $\mathbf{x}$ be a recurrent word in the alphabet $\{0, 1\}$, let $k \in \mathbb{N}$ and let $\Gamma_k(\mathbf{x}) = \{z \in L(\mathbf{x}) \mid \mathrm{Pref}_1(z) = \mathrm{Suff}_1(z) = 1 \text{ and } |z|_1 = k\}$. For example, the words 1000100001 and 1011 are elements of $\Gamma_3((1000100001011)^\omega)$

Theorem 8 *Let $\mathbf{x}$ be a recurrent non periodic word in the alphabet $\{0, 1\}$. The word $\mathbf{x}$ is Sturmian if and only if, for every z and z' in $\Gamma_k(\mathbf{x})$ and for every $\in \mathbb{N}$,*

$$||z| - |z'|| \leq 1.$$

Proof Suppose that $\mathbf{x}$ is a Sturmian word, then by Theorem 7 with $u = 1$, we have the statement.

For the other implication, we reason by contradiction. We use the fact that a word $\mathbf{x}$ is Sturmian if and only if, for each $n \geq 0$, there is one and only one factor of $\mathbf{x}$ of length n having two extensions, the others having exactly one (see [10]).

Suppose that, for every z and z' in $\Gamma_k(\mathbf{x})$, we have $||z| - |z'|| \leq 1$ and that $\mathbf{x}$ is not Sturmian.

First case: there exists n_0 such that each factor of $\mathbf{x}$ of length n_0 has a unique right extension. Then the word is periodic, which is in contradiction with the fact that $\mathbf{x}$ is non periodic.

Second case: there exists n_0 such that two factors of $\mathbf{x}$ of length n_0 has two right extensions. Let n_0 be the smallest one having this property and v, w be the factors of $\mathbf{x}$ of length n_0 such that $v0$, $v1$, $w0$ and $w1$ are also factors of $\mathbf{x}$ and $|v0| = |v1| = |w0| = |w1| = n_0 + 1$ and for $n < n_0$, there exists for each n a unique word of length n with two right extensions. Thus we have $v = av'$ and $w = bw'$ where a and b are letters of the alphabet. As v' and w' are factors of length $n_0 - 1$ with two right extensions, then $v' = w'$. In other word, $0v'0, 0v'1, 1v'0, 1v'1$ are factors of $\mathbf{x}$. Let m be the number of 1's in $1v'1$, by definition $1v'1$ is an element of $\Gamma_m(\mathbf{x})$. Furthermore, the factor $0v'0$ can be extended to the right and to the left. In general form, we can find p and q positive integers such that $10^p0v'00^q1$ is an element of $\Gamma_m(\mathbf{x})$. Thus $|10^p0v'00^q1| - |1v'1| = 2 + p + q > 1$ and there is a contradiction because we find two elements of $\Gamma_m(\mathbf{x})$ with equilibrium greater than 1. ∎

7 Generalized Balance

Let w, w' and u be factors of $\mathbf{x}$ such that $|w| = |w'|$. We denote

$$\Delta_u(w, w') = ||w|_u - |w'|_u|$$

the *balance* of u upon w and w' and

$$\delta(u) = \max\{\Delta_u(v, v') \mid v, v' \in L(\mathbf{x}), |v| = |v'|\}$$

the maximal balance of u.

Proposition 9 *Let $\mathbf{x}$ be a Sturmian word and u be a factor of $\mathbf{x}$. We have the following cases.*

i. if u is non-overlapping, then $\delta(u) \leq 2$;

ii. if v is the period of u, $u = v^r$ and $v^{r+1} \notin L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then

(a) if $|v| = 1$, then $\delta(u) \leq 2$;

(b) otherwise, $\delta(u) \leq 3$;

iii. if v is the period of u, $u = v^r$ and $v^{r+1} \in L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then

(a) if $|v| = 1$, then $\delta(u) \leq \max(2, r)$;

(b) if $|v| = 2$, then $\delta(u) \leq r + 1$;

(c) otherwise, $\delta(u) < r + 2$.

Remark Most of these bounds are reached. For each example, it is easy to verify that w and w' are factors of the same Sturmian word.

i. Let $w = a^k ba^k b$, $w' = a^{k-1}ba^{k-1}ba^2$ and $u = a^k b$. We have $\Delta_u(w, w') = 2$.

ii. (a) Let $w = a^k ba^k$, $w' = a^{k-1}ba^{k-1}ba$ and $u = a^k$. We have $\Delta_u(w, w') = 2$.

iii. (a) if $|u| = 1$, let $w = a^{2k-1}$, $w' = a^{k-1}ba^{k-1}$ and $u = a^k$. We have $\Delta_u(w, w') = k$.

(b) Let $w = (ab)^{2n+1}b(ab)^{2n+1}$, $w' = b(ab)^n b(ab)^{2n} b(ab)^n ab$ and $u = (ab)^{n+1}$. Then we have $\Delta_u(w, w') = n + 1$.

(c) Let $w = (a^k b)^{2n+1} a^{k-1} b (a^k b)^{2n+1}$, $w' = a^{k-1} b (a^k b)^n a^{k-1} b (a^k b)^{2n} a^{k-1} b (a^k b)^n a^2$ and $u = (a^k b)^n a^k$, with $r = n + \frac{k}{k+1}$, $k > 2$. $\Delta_u(w, w') = n+2 = \lfloor r+2 \rfloor$.

The proof of Proposition 9 is an immediate consequence of the combination of Proposition 10 and Proposition 11 mentioned below. The proof of these two latter propositions will be given in the appendix.

Proposition 10 *Let $\mathbf{x}$ be a Sturmian word and u a factor of $\mathbf{x}$. Let $\mathcal{H}_u(\mathbf{x}) = \{u_1, u_2\}$, with $|u_1| \le |u_2|$. We have the following:*

i. if u is non-overlapping, then $|u_1| \ge |u|$;

ii. if v is the period of u, $u = v^r$ and $v^{r+1} \notin L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then $|u_1| \ge \max\{|v|+1, (r-1)|v|+1\}$;

iii. if v is the period of u, $u = v^r$ and $v^{r+1} \in L(\mathbf{x})$, with $r > 1 \in \mathbb{Q}$, then $u_1 = v$.

Proposition 11 *Let $\mathbf{x}$ be a Sturmian word and u be a factor of $\mathbf{x}$. Let $\mathcal{H}_u(\mathbf{x}) = \{u_1, u_2\}$, with $|u_1| \le |u_2|$. We have the following inequality*

$$\delta(u) \le \max\left(2, \frac{|u|-2}{|u_1|} + 2\right).$$

Proof of Proposition 9 We distinguish the same cases as in the proposition's statement. Let $\mathcal{H}_u(\mathbf{x}) = \{u_1, u_2\}$, with $|u_1| \le |u_2|$. Let us denote $e = \frac{|u|-2}{|u_1|} + 2$. We have then $\delta(u) \le \max(2, e)$.

i. By Proposition 10, we have $|u_1| \ge |u|$. Thus $e \le \frac{|u|-2}{|u|} + 2 = 3 - \frac{2}{|u|} < 3$. Therefore $\delta(u) < 3$, and, since $\delta(u)$ is an integer, we get $\delta(u) \le 2$.

ii. By Proposition 10, we have $|u_1| \ge \max\{|v|+1, (r-1)|v|+1\}$. We have also $|u| = r|v|$. Let $e_1 = \frac{r|v|-2}{|v|+1} + 2$ and $e_2 = \frac{r|v|-2}{(r-1)|v|+1} + 2$. We have obviously $e \le \min(e_1, e_2)$.

   (a) We have $|v| = 1$, thus $e_2 = \frac{r-2}{r} + 2 = 3 - \frac{2}{r} < 3$. Therefore $e < 3$ and $\delta(u) < 3$, that is $\delta(u) \le 2$.

   (b) Here we have two sub-cases.

   - If $r \ge 2$, then $e_2 = \frac{(r-1)|v|+1}{(r-1)|v|+1} + \frac{|v|-3}{(r-1)|v|+1} + 2$. As $r \ge 2$, we have $\frac{|v|-3}{(r-1)|v|+1} < 1$, then $e_2 < 4$. Therefore, $\delta(u) < 4$ and then also $\delta(u) \le 3$.
   - If $1 < r < 2$, then $e_1 < \frac{2|v|-2}{|v|+1} + 2 = \frac{2|v|+2}{|v|+1} - \frac{4}{|v|+1} + 2 < 4$. As in the precedent case, we get $\delta(u) \le 3$.

iii. Since v is the period of $|u|$ and $vu \in L(\mathbf{x})$, we have $|u_1| = |v|$. Then $e = \frac{r|v|-2}{|v|} + 2 = r + 2 - \frac{2}{|v|}$

   (a) Since $|v| = 1$, we get $e = r$, and thus $\delta(u) \le \max(2, r)$.

   (b) Since here $|v| = 2$, we get $e = r+1$, and thus $\delta(u) \le r+1$. We need not a maximum here, because $r+1 > 2$.

   (c) The general case gives the inequality $e < r+2$, and $\delta(u) < r+2$ too.

∎

We are know ready to state the main theorem.

Theorem 12 *Let* $\mathbf{x}$ *be a Sturmian word. Let* u*,* w *and* w' *be three factors of* $\mathbf{x}$*. We have*

$$|w| = |w'| \implies ||w|_u - |w'|_u| \le |u|.$$

Proof Remark that the main theorem is true for $|u| = 1$. We suppose that $|u| \ge 2$. In this proof, we use the results and the cases of the Proposition 9.

i. If u is non-overlapping, then, by Proposition 9, $||w|_u - |w'|_u| \le \delta(u) \le 2 \le |u|$. The statement is true for non-overlapping case.

ii. (a) If $|v| = 1$, then, by hypothesis, $|u| \ge 2$. Thus by Proposition 9 we have $\delta(u) \le 2 \le |u|$. This gives the statement.
  (b) If $|v| \ge 2$ then $|u| = r|v| \ge 3$. Thus we have $\delta(u) \le 3 \le |u|$.

iii. (a) If $|v| = 1$ then $|u| = r$. In consequence, by Proposition 9, $\delta(u) \le r = |u|$.
  (b) If $|v| = 2$ then $|u| = 2r$. Thus we have $\delta(u) \le r + 1 = \frac{|u|}{2} + 1$. It is sufficient to prove that $\frac{|u|}{2} + 1 \le |u|$. As $|u| \ge 2$ the statement is true in the case $|v| = 2$
  (c) If $|v| > 2$, as by Proposition 9, $\delta(u) \le r + 2$, then $\frac{\delta(u)}{r} \le \frac{\lceil r \rceil}{r} + \frac{2}{r} \le 3$. This gives the bound $\delta(u) \le 3r$. In consequence, if $|v| \ge 3$, then $\delta(u) \le 3r \le |v|r = |u|$.

∎

Remark The reciprocal of Theorem 12 is obvious since, if we take $|u| = 1$, we get the classical definition of Sturmian words by balance.

Nevertheless, it would be interesting to study the words such that $E(u) \le 2$ for every factor u of length 2.

Proposition 9 also permits us to write the following corollary.

Corollary 13 *Let* $\mathbf{x}$ *be a Sturmian word. Let* α *be the slope of* $\mathbf{x}$ *and let* $\alpha = [0, a_1, a_2, \ldots, a_n, \ldots]$ *be its continued fraction expansion. If the partial quotients* $(a_n)_{n\in\mathbb{N}}$ *are bounded then* $(\delta(u))_{u\in L(\mathbf{x})}$ *is bounded too.*

Proof By Proposition 2, there exists an integer d such that for any factor $v \in L(\mathbf{x})$, we have $v^{d+1} \notin L(\mathbf{x})$. Let u be a factor of $\mathbf{x}$. By Proposition 9, if $\delta(u) > 3$, then we have $u = v^r$ and $v^{r+1} \in L(\mathbf{x})$, and then $\delta(u) \le r + 2$. But in this case, we have $r < d$ and then $\delta(u)$ is bounded by $d + 1$. Thus, for any factor $u \in \mathbf{x}$, we have $\delta(u) \le \max(3, d + 1)$. ∎

8 Appendix

Here we will give the proofs of Propositions 10 and 11. Some extra definitions will be useful.

We denote $\mathrm{occ}_i(u, w)$ the i^{th} occurrence of u in w, i.e. if $w = \alpha u \beta$ such that $|\alpha u|_u = i$, then $\mathrm{occ}_i(u, w) = |\alpha|$.

By extension, assuming that we consider a fixed infinite word, we define

$$\mathrm{occ}_0(u,w) = -\min\{|\alpha| \mid \alpha \in A^+, \exists \beta \in A^*, \alpha w \in L(\mathbf{x}), \alpha w = u\beta\}.$$

Remark Such an α always exists because of the uniform recurrence of $\mathbf{x}$. Similarly, if w has k occurrences of u, we define

$$\mathrm{occ}_{k+1}(u,w) = \min\{|\beta| \mid \beta \in A^*, \exists \alpha \in A^+, w\alpha \in L(\mathbf{x}), w\alpha = \beta u\}.$$

For example, if $\mathbf{x}$ is the Fibonacci word,

$$\mathbf{x} = 0100101001001010010100\cdots,$$

then we have $\mathrm{occ}_0(01, 101001) = -1$, $\mathrm{occ}_1(01, 101001) = 1$, $\mathrm{occ}_2(01, 101001) = 4$ and $\mathrm{occ}_3(01, 101001) = 6$.

We can remark that:

Lemma 14 *Let $\mathbf{x}$ be a Sturmian word and u be a factor of $\mathbf{x}$ with $\mathcal{H} = \{u_1, u_2\}$. Let w be another factor of $\mathbf{x}$ and k be an integer such that $0 \le k \le |w|_u$.*

We have either $\mathrm{occ}_{k+1}(u,w) - \mathrm{occ}_k(u,w) = |u_1|$ or $\mathrm{occ}_{k+1}(u,w) - \mathrm{occ}_k(u,w) = |u_2|$.

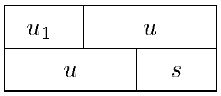
Proof It is an immediate consequence of the definition of return words. ■

The following lemma will be useful too. It is a classical result of combinatorics (see [17]).

Lemma 15 *Let v be a word such that there exist two non-empty words p and s such that $v = ps = sp$. Then v has a integral period strictly smaller than $|v|$.*

Proof of Proposition 10 We will give a different proof for each case.

i. Suppose $|u_1| < |u|$, then, by definition of u_1, there exists a word s such that $u_1 u = us$ with $|s| = |u_1| < |u|$. Thus u will be overlapping which is absurd.



ii. If $|u_1| \ge |u|$, the inequality is satisfied. Suppose, now, that $|u_1| < |u|$. Then, u is clearly overlapping, consequently, as remarked in Section 2, u_1 is a period of u. As v is the period of u, we have $|u_1| \ge |v|$. Moreover, since $v^{r+1} \notin L(\mathbf{x})$, we cannot have the equality, thus $|u_1| > |v|$.

Now, let suppose that $|v| < |u_1| \le (r-1)|v|$. Since $u_1 u$ has u as prefix and since $|u_1| > |u|$, the word v is a period of u_1, i.e. there exists $t \in \mathbb{Q}, 1 < t \le r-1$ such that $u_1 = v^t$. Either $t \in \mathbb{N}$ (Figure 2), and then $u_1 u = v^{t+r} \in L(\mathbf{x})$ which is in contradiction with the hypothesis, or $t \notin \mathbb{N}$ (Figure 3), and then, take $s = \{t\} \cdot |v|$, where $\{t\}$ denotes the fractional part of t, we have $v = v[s, |v|) \cdot v[0, s)$. Then, by Lemma 15, v is periodic (with an integral power), which implies that v is not the smallest period of u. This is also a contradiction to the hypothesis.

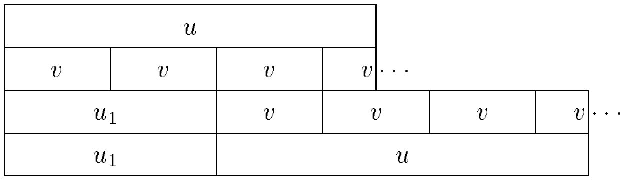


Figure 2: case $t \in \mathbb{N}$

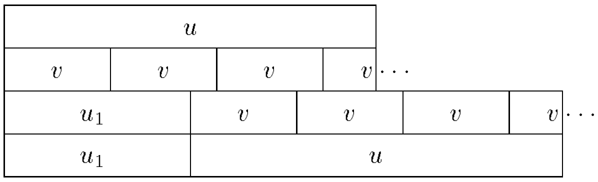
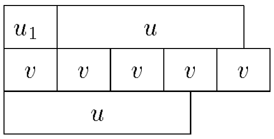


Figure 3: case $t \notin \mathbb{N}$

iii. We reason as above: we have that $|u_1| \geq |v|$. Since $vu = v^{r+1} \in L(\mathbf{x})$ and since we have $\mathrm{Pref}_{|u|}(vu) = u$, we can conclude that $u_1 = u$.



■

Proof of Proposition 11 Let w and w' be two factors of $\mathbf{x}$, such that $|w| = |w'|$. We can suppose that $|w|_u - |w'|_u \geq 0$.

We are going to restrain the study to a subset of $\{(w, w') \mid w, w' \in L(\mathbf{x}), |w| = |w'|\}$, the underlying idea being that we only need to consider "the worst cases".

■

Step 1 *We can suppose that $|w'|_u \geq 1$.*

Proof Let us suppose that $|w'|_u = 0$. Let $k = |w|_u$. We are going to prove that $k = \Delta_u(w, w') \leq \max(2, \frac{|u|-2}{|u_1|} + 2)$.

Let w'_1 be the longest word in $L(\mathbf{x})$ such that $|w'_1|_u = 0$. Obviously, we have $w'_1 = \mathrm{Suff}_{|u_2|-1}(u_2) \cdot \mathrm{Pref}_{|u|-1}(u)$. Now let w_1 be the shortest word such that $|w_1|_u = k$. Assuming that $u_1^{k-1}u$ is in $L(\mathbf{x})$, we have $w_1 = u_1^{k-1}u$, otherwise $|w_1|$ is obviously larger.

Therefore, we must have $|w_1| \leq |w| = |w'| \leq |w'_1|$. Which leads to the inequality

$$(k-1)|u_1| + |u| \leq |u_2| + |u| - 2,$$

or equivalently

$$|u_2| \geq (k-1)|u_1| + 2.$$

By using Proposition 6, i.e. $|u_2| \leq |u_1| + |u|$, we can conclude that

$$|u| \geq (k-2)|u_1| + 2.$$

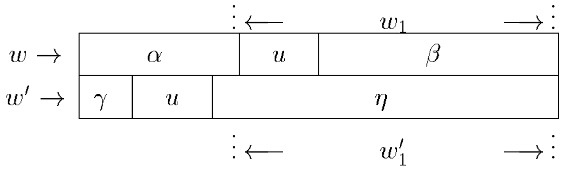
Now, we can see that $\frac{|u|-2}{|u_1|} + 2 \geq k = \Delta_u(w, w')$. We are through. ■

Step 2 *We can suppose that* $|w|_u \geq 4$.

Proof If $|w|_u \leq 3$, then we have, using Step 1, $\Delta_u(w, w') \leq 3 - 1 = 2$. Which is in accordance with the Proposition. ∎

Step 3 *We can suppose that* $\mathrm{occ}_1(u, w) = 0$.

Proof Suppose that $w = \alpha u \beta$, with $|\alpha| = \mathrm{occ}_1(u, w) \neq 0$. Let $\ell = |w| - |\alpha|$. If we take $w_1 = \mathrm{Suff}_\ell(w)$ and $w'_1 = \mathrm{Suff}_\ell(w')$, then $|w_1|_u = |w|_u$ and $|w'_1|_u \leq |w'|_u$ and then $\Delta_u(w_1, w'_1) \geq \Delta_u(w, w')$. So we have no need of considering the couple (w, w').

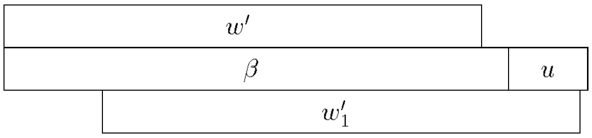


∎

Symmetrically, we suppose that $\mathrm{occ}_1(\tilde{u}, \tilde{w}) = 0$ (Here, $\tilde{u}$ is the reversal of u). So that we have $w = u_1 \alpha u$ or $w = u_2 \alpha u$, with $\alpha \in A^*$. ($w = u$ is excluded because $|w|_u \geq 4$.)

Step 4 *We can suppose that we have* $w' = \beta\eta$ *with* $\eta = \mathrm{Pref}_{|u|-1}(u)$.

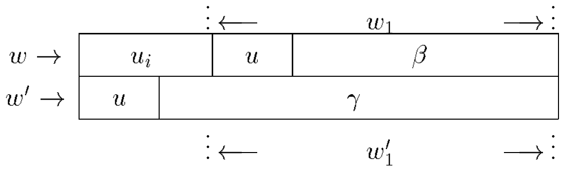
Proof If the condition is not right, we have $w' = \mathrm{Pref}_{|w'|}(\beta u)$, with $|\beta|$ the first occurrence of u in βu which is not in w'. If we take $w'_1 = \mathrm{Pref}_{|w'|}\left(\mathrm{Suff}_{|w'|+1}(\beta u)\right)$, then $|w'_1|_u \leq |w'|_u$, because we are sure to have not added any occurrence of u, but we can have lost some in the beginning of βu. Then $\Delta_u(w, w'_1) \geq \Delta_u(w, w')$, and so we can eliminate the couple (w, w').



∎

Step 5 *We can suppose that* $\mathrm{occ}_1(u, w') > O$.

Proof Suppose we have $\mathrm{occ}_1(u, w') = 0$. We have then $w = u_i u \beta$, with $i = 1$ or $i = 2$. We have also $w' = u\gamma$. Let $\ell = |w| - |u_i|$. We take $w_1 = \mathrm{Suff}_\ell(w)$ and $w'_1 = \mathrm{Suff}_\ell(w')$. We have $|w_1|_u = |w|_u - 1$ and $|w'_1|_u \leq |w'|_u - 1$, so $\Delta_u(w_1, w'_1) \geq \Delta_u(w, w')$. Then studying the couple (w, w') is useless.



∎

Summary: We can restrain the study to couples (w, w') such that

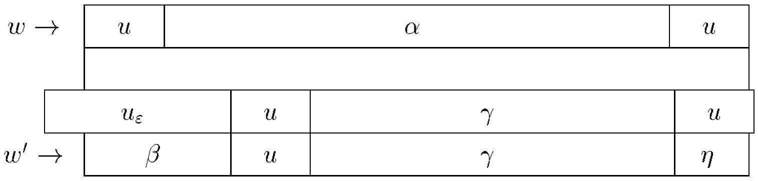
i. $|w| = |w'|$;

ii. $|w|_u - |w'|_u \geq 2$;

iii. $w = u\alpha u$;

iv. $w' = \beta u \gamma \eta$, with $\beta = \mathrm{Suff}_\ell(u_\varepsilon)$, for some $\varepsilon = 1$ or 2 and $0 < \ell < |u_\varepsilon|$, and $\eta = \mathrm{Pref}_{|u|-1}(u)$.

That is to say, we have the following sketch:



Proof of Proposition 11 (continued): As mentioned above, we restrain the study to couples (w, w') with the properties of the summary. So, we can write w and w' as follows:

- $w = u_{i_1} u_{i_2} \cdots u_{i_p} u$, with either $u_{i_k} = u_1$ or $u_{i_k} = u_2$.
- $w' = \beta u_{j_1} u_{j_2} \cdots u_{j_q} \eta$ with either $u_{i_k} = u_1$ or $u_{i_k} = u_2$, with $\beta = \mathrm{Suff}_\ell(u_\varepsilon)$, for some $0 < \ell < |u_2|$, and $\eta = \mathrm{Pref}_{|u|-1}(u)$.

We have then $|w|_u = p + 1$ and $|w'|_u = q$, therefore $\Delta_u(w, w') = p - q + 1$. As we suppose $\Delta_u(w, w') \geq 2$, we have $p \geq 1 + q$.

If $p = 1 + q$, then we have $\Delta_u(w, w') = 2$ is in accordance with the desired result.

Elsewhere, we suppose that $p > 1 + q$. We have then

$$\begin{aligned} |w| &= |u_{i_1} u_{i_2} \cdots u_{i_{q+1}}| + \sum_{k=q+2}^{p} |u_{i_k}| + |u| \\ |w'| &= |u_\varepsilon u_{j_1} u_{j_2} \cdots u_{j_q}| + (|u| - 1) - (|u_\varepsilon| - |\beta|) \end{aligned}$$

Thus, using the equality $|w| = |w'|$, we have

$$\sum_{k=q+2}^{p} |u_{i_k}| = |u_\varepsilon u_{j_1} u_{j_2} \cdots u_{j_q}| - |u_{i_1} u_{i_2} \cdots u_{i_{q+1}}| + (|u| - 1) - (|u_\varepsilon| - |\beta|) - |u|.$$

By Proposition 6, we have

$$\big| |u_{i_1} u_{i_2} \cdots u_{i_{q+1}}| - |u_\varepsilon u_{j_1} u_{j_2} \cdots u_{j_q}| \big| \leq |u|.$$

We have also $\sum_{k=q+2}^{p} |u_{i_k}| \geq (p - q - 1)|u_1|$ and $|u_\varepsilon| - |\beta| \geq 1$

We can therefore conclude that $(p - q - 1)|u_1| \leq |u| - 2$ i.e. $(p - q - 1) \leq \frac{|u|-2}{|u_1|}$ and thus

$$\Delta_u(w, w') = p - q + 1 \leq \frac{|u| - 2}{|u_1|} + 2.$$

■

Acknowledgments: *We wish to express our thanks to Jean Berstel for giving us the subject of this paper and for many stimulating conversations also. We also thank the referee for pointing out the papers of Ostrowski and Kesten and for useful corrections. We are indebted to Peter Habermehl for his careful reading of our English. The first author would also like to thank her new laboratory (Institut Gaspard Monge) for having welcomed her.*

References

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

[2] J. Berstel and P. Séébold, Sturmian words, in M. Lothaire, *Algebraic combinatorics on Words,* (2001), to appear.

[3] V. Berthé, Fréquences des facteurs des suites sturmiennes, *Theoret. Comput. Sci.* **165** (1996) 295–309.

[4] V. Berthé and R. Tijdeman, Balance properties of multidimensional words, preprint IML 2000/05.

[5] J. Cassaigne, Complexité et facteurs spéciaux, Journées Montoises (Mons, 1994). *Bull. Belg. Math. Soc. Simon Stevin* **4** (1997) no. 1, 67–88.

[6] J. Cassaigne, S. Ferenczi and L. Zamboni, Imbalances in Arnoux-Rauzy Sequences. *Ann. Inst. Fourier.* To appear.

[7] F. Durand, A characterization of substitutive sequences using return words, *Discrete. Math.* **179** (1998) 89–101.

[8] B. Gaujal, Optimal allocation sequences of two processes sharing a resource, *Discrete Event Dynamic Systems* **7** (1997) 327–354.

[9] R. L. Graham, Covering the positive integers by disjoint sets of the form $[n\alpha + \beta] : n = 1, 2, \ldots$, *J. Combin. Theory Ser.* A **15** (1973) 354–358.

[10] G. A. Hedlund, M. Morse, Symbolic dynamics II: Sturmian trajectories, *Amer. J. Math.* **62** (1940) 1–42.

[11] C. Holton, L. Q. Zamboni, Descendants of primitive substitutions, *Theory Comput. Systems* **32** (1999) 133–157.

[12] P. Hubert, Well balanced sequences, *Theoret. Comput. Sci.* to appear.

[13] H. Kesten, On a conjecture of Erdös and Szüsz related to uniform distribution mod 1, *Acta Arithmetica* **XII** (1966), 193–212.

[14] J. Mairesse and L. Vuillon, Asymptotic Behavior in a Heap Model with Two Pieces, LIAFA Report 98/09 (1998).

[15] Von A. Ostrowski, Bemerkungen zur Theorie der Diophantischen Approximationen, *Abh. Math. Semin. Hamburg Univ.* **1** (1922), 77–98.

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

[17] D. Perrin, Words, in M. Lothaire, *Combinatorics on words,* Addison-Wesley, 1985.

[18] R. Tijdeman, Exact covers of balanced sequences and Fraenkel's conjecture, *Proc. Number Theory Conference Graz in 1998,* De Gruyter.

[19] L. Vuillon, A characterization of Sturmian words by return words, *European J. Combin.* to appear.