Theoretical Informatics and Applications Will be set by the publisher
Informatique Théorique et Applications

RETURN WORDS IN STURMIAN AND EPISTURMIAN WORDS

Jacques Justin and Laurent Vuillon[1]

Abstract. Considering each occurrence of a word w in a recurrent infinite word, we define the set of return words of w to be the set of all distinct words beginning with an occurrence of w and ending exactly just before the next occurrence of w in the infinite word. We give a simpler proof of the recent result (of the second author) that an infinite word is Sturmian if and only if each of its factors has exactly two return words in it. Then, considering episturmian infinite words, which are a natural generalization of Sturmian words, we study the position of the occurrences of any factor in such infinite words and we determinate the return words. At last, we apply these results in order to get a kind of balance property of episturmian words and to calculate the recurrence function of these words.

Résumé. Si l'on considère chaque occurrence d'un mot w dans un mot infini récurrent, on définit l'ensemble des mots de retour de w comme l'ensemble de tous les mots distincts débutant avec une occurrence de w et finissant juste avant l'occurrence suivante de w. Nous donnons une nouvelle démonstration d'un résultat établi récemment par le deuxième auteur: un mot infini est sturmien si et seulement si chacun de ses facteurs a exactement deux mots de retour. Nous étudions les mots épisturmiens qui sont une généralisation naturelle des mots sturmiens. Puis nous déterminons la position d'un facteur donné et ses mots de retour dans un mot épisturmien. Enfin nous appliquons ces méthodes pour obtenir une propriété d'équilibre pour les mots épisturmiens et calculer la fonction de récurrence de ces mots infinis.

AMS Subject Classification. 68R15.

[1] LIAFA, Université Paris VII,
case 7014, 2 place Jussieu,
75251 Paris Cedex 05, France.
Phone. +33 1 44 27 70 97; Fax. +33 1 44 27 62 86;
e-mail: {Jacques.Justin, Laurent.Vuillon}@liafa.jussieu.fr

Introduction

The notion of return words is a powerful tool for the study of Symbolic Dynamical Systems, Combinatorics on Words and Number Theory. Considering each occurrence of a word w in a recurrent sequence $U = (U_n)_{n\in\mathbb{N}}$, we define the set of return words of w to be the set of all distinct words beginning with an occurrence of w and ending exactly before the next occurrence of w in the infinite sequence. This mathematical tool was introduced independently by Durand, Holton and Zamboni in order to study primitive substitutive sequences (see [8, 9, 12]). This notion is quite natural and can be seen as a symbolic version of the first return map for a dynamical system. Recently, many articles use return words. For example, Allouche, Davinson, Queffelec and Zamboni study the transcendence of Sturmian or morphic continued fractions and a main argument is to show, using return words, that arbitrarily long prefixes are "almost squares" (see [1]). Fagnot and Vuillon give a generalization of the notion of balance property for Sturmian words and the proof is based on return words and combinatorics on words (see [10]). Cassaigne also uses this tool to investigate a Rauzy conjecture (see [4]).

At last, the second author shows that a sequence is Sturmian if and only if for each word w appearing in the sequence, the number of return words of w is exactly two (see [14]). Recall that Sturmian sequences are aperiodic sequences with complexity $p(n) = n + 1$ for all n (the complexity function $p(n)$ counts the number of distinct factors of length n in the sequence) (see [3, 11]).

In this paper, we give a simpler proof of this result (Section 2) and then, in Sections 3 and 4, we study the occurrences of factors and the return words in episturmian words (episturmian words on a finite alphabet are a natural generalization of Sturmian words introduced in [7] which includes in particular Sturmian words and Arnoux-Rauzy sequences [2]). This allows (Section 5) to calculate the recurrence function, obtaining or completing known results [5, 11], and to state a kind of balance property of episturmian words which when applied to Sturmian words coincides with the well known balance property of these words.

1. Definitions and notations

1.1. Words

Given a finite *alphabet* $\mathcal{A}$, $\mathcal{A}^*$ is the set of *words* on $\mathcal{A}$ and $\mathcal{A}^+ = \mathcal{A}^* \setminus \{\varepsilon\}$ with ε the *empty word.* If $u = u(1)u(2)\cdots u(m)$ with $u(i) \in \mathcal{A}$ its *length* is $|u| = m$ and its *reversal* is $\widetilde{u} = u(m)u(m-1)\cdots u(1)$. The word u is a *palindrome* or is *palindromic* if $u = \widetilde{u}$.

Similarly $\mathcal{A}^\omega$ is the set of *infinite words* (or infinite sequences) $\mathbf{t} = t(1)t(2)\cdots$ on $\mathcal{A}$.

A finite word u is a *factor* of the finite or infinite word t if $t = t'ut''$, $t' \in \mathcal{A}^*$ and $t'' \in \mathcal{A}^* \cup \mathcal{A}^\omega$. This factor (or rather its occurrence so defined) is a *prefix* if $t' = \varepsilon$, a *suffix* if $t'' = \varepsilon$, is *interior* if $t', t'' \neq \varepsilon$. Also u is *unioccurrent* if it has exactly one occurrence in t.

The set of factors of the finite or infinite word t is $F(t)$ and $F_\ell(t) = F(t) \cap \mathcal{A}^\ell$. The set of letters occurring (resp. occurring infinitely many times) in t is $Alph(t) = F_1(\mathbf{t})$ (resp. $Ult(t)$).

1.2. Return words

Let $\mathbf{t} = t(1)t(2)\cdots, t(i) \in \mathcal{A}$ be an infinite word. Then $\mathbf{t}$ is *recurrent* if any of its factors occurs infinitely many times in it. In this case, for $u \in F(\mathbf{t})$, let $n_1 < n_2 < \cdots$ be all the integers n_i such that $u = t(n_i)\cdots t(n_i + |u| - 1)$. Then the word $t(n_i)\cdots t(n_{i+1} - 1)$ is a *return word* of u in $\mathbf{t}$. Let $\mathcal{H}_u(\mathbf{t})$ be the set of return words of u in $\mathbf{t}$. Then $\mathbf{t}$ can be factorized in a unique way as $\mathbf{t} = t(1)\cdots t(n_1 - 1)r^{(1)}r^{(2)}\cdots$ where $r^{(i)} \in \mathcal{H}_u(\mathbf{t})$. If we consider $r^{(1)}r^{(2)}\cdots$ as an infinite word on the alphabet $\mathcal{H}_u(\mathbf{t})$, this one is called the *derived word* of $\mathbf{t}$ relatively to u.

The set $\mathcal{H}_u(\mathbf{t})$ is finite for all $u \in F(\mathbf{t})$ if and only if $\mathbf{t}$ is *uniformly recurrent*. Lastly if r is a return word of u then the factor ru of $\mathbf{t}$ is a *complete return word* of u in $\mathbf{t}$.

1.3. Episturmian words

An infinite word $\mathbf{t} \in \mathcal{A}$ is *episturmian* if $F(\mathbf{t})$ is closed under reversal and for any $\ell \in \mathbb{N}$ there exists at most one right special word in $F_\ell(\mathbf{t})$ (a factor u is *right special* if $ux, uy \in F(\mathbf{t})$ for at least two different letters x, y), see [7, 13].

Sturmian words, which can be defined in many ways, are exactly the non-periodic episturmian words on a two-letter alphabet. Sturmian words have the remarkable balance property. A word w is *balanced* if $u, v \in F(w)$ and $|u| = |v|$ imply $||u|_x - |v|_x| \leq 1$ for any $x \in \mathcal{A}$ and with $|u|_x$ the number of x occuring in u.

As episturmian words are uniformly recurrent and as we are interested here only in factors, we limit ourselves to the consideration of standard episturmian words (an episturmian word is *standard* if all its left special factors are prefixes of it). Let $\mathbf{s}$ be a standard episturmian word and let $u_1 = \varepsilon, u_2, u_3, \cdots$ be the sequence of its palindromic prefixes. Then there exists an infinite word $\Delta(\mathbf{s}) = x_1x_2\cdots, x_i \in \mathcal{A}$ called its *directive word* such that for all $n \in \mathbb{N}_+$ (the set of positive integers),

$$u_{n+1} = (u_n x_n)^{(+)}$$

where the *right palindromic* closure $(+)$ is defined by: $w^{(+)}$ is the shortest palindrome having w as a prefix (see this construction for Sturmian words by de Luca [6]).

Example: if $\Delta(\mathbf{s}) = (abc)^\omega$ then the infinite word $\mathbf{s}$ begins by

$$\mathbf{s} = \mathbf{ab}acaba\mathbf{a}bacaba\mathbf{b}acabaabacaba\mathbf{c}\cdots$$

where the letters of the word $\Delta(\mathbf{s})$ are bolded. We have for this example $u_1 = \varepsilon$, $u_2 = a$, $u_3 = aba$ and so on.

For $a \in \mathcal{A}$ let ψ_a be the morphism given by

$$\psi_a(a) = a, \psi_a(x) = ax$$

for $x \in \mathcal{A}, x \neq a$. Let

$$\mu_n = \psi_{x_1}\psi_{x_2}\cdots\psi_{x_n}, \mu_0 = Id,$$

and

$$h_n = \mu_n(x_{n+1}).$$

Then we have the useful formula $u_{n+1} = h_{n-1}u_n$ and more generally $(u_n x)^{(+)} = \mu_{n-1}(x)u_n$ for $x \in \mathcal{A}$.

At last, there exists an infinite sequence of standard episturmian words $\mathbf{s}_0 = \mathbf{s}, \mathbf{s}_1, \mathbf{s}_2, \cdots$ such that $\mathbf{s} = \mu_n(\mathbf{s}_n)$ for $n \in \mathbb{N}$.

These notations will be kept throughout this paper.

2. A characteristic property of Sturmian words

We will say that an infinite word $\mathbf{s} \in \mathcal{A}^\omega$ has property $\mathcal{R}_n$ if for any factor w of $\mathbf{s}$ the number of return words of w is exactly n.

A letter a of the alphabet $\mathcal{A}$ will be called *separating* in $\mathbf{s} \in \mathcal{A}^\omega$ if any factor of length two of $\mathbf{s}$ contains at least one a. For example: the letter a in the infinite word $\mathbf{y} = (aaaabaab)^\omega$ is separating. Hereafter in this section the alphabet will be $\mathcal{A} = \{a, b\}$.

Lemma 2.1. *If an infinite word* $\mathbf{s}$ *has the property* $\mathcal{R}_2$ *then either* a *or* b *is separating.*

Let ψ_a be the morphism $\psi_a(a) = a$ and $\psi_a(b) = ab$. Let $\widetilde{\psi}_a$ be the morphism $\widetilde{\psi}_a(a) = a$ and $\widetilde{\psi}_a(b) = ba$.

Lemma 2.2. *Let* $\mathbf{s} \in \mathcal{A}^\omega$ *be an infinite word with the property* $\mathcal{R}_2$*. Let for instance* a *be a separating letter, then there exists an infinite word* $\mathbf{t}$ *with either* $\mathbf{s} = \psi_a(\mathbf{t})$ *or* $\mathbf{s} = \widetilde{\psi_a}(\mathbf{t})$*. Furthermore* $\mathbf{t}$ *has the property* $\mathcal{R}_2$*. Conversely, if* $\mathbf{s} = \psi_a(\mathbf{t})$ *or* $\mathbf{s} = \widetilde{\psi_a}(\mathbf{t})$ *and* $\mathbf{t} \in \mathcal{A}^\omega$ *has the property* $\mathcal{R}_2$ *then* $\mathbf{s}$ *has the property* $\mathcal{R}_2$*.*

Lemma 2.3. *If an infinite word* $\mathbf{s} \in \mathcal{A}^\omega$ *is non-periodic and if* $\mathbf{s} = \mathbf{s}_0, \mathbf{s}_1, \cdots$ *is an infinite sequence of infinite words such that either* $\mathbf{s}_{i-1} = \psi_{x_i}(\mathbf{s}_i)$ *or* $\mathbf{s}_{i-1} = \widetilde{\psi}_{x_i}(\mathbf{s}_i)$ *where* $x_i \in \mathcal{A}$ *then* $\mathbf{s}$ *is Sturmian.*

The following theorem is one half of the main result of [14]. For the second half see Remark 2.5 hereafter.

Theorem 2.4. *[14] If an infinite word* $\mathbf{s} \in \mathcal{A}^\omega$ *has the property* $\mathcal{R}_2$ *then it is Sturmian.*

Proof. The proof is an immediate consequence of Lemma 2.1, Lemma 2.2 and Lemma 2.3. □

Proof. Proof of Lemma 2.1. Suppose by contradiction that a and b are not separating in $\mathbf{s}$. Then aa and bb occur in $\mathbf{s}$. Write for instance

$$\mathbf{s} = a^p b^{m_1} a^{n_1} b^{m_2} a^{n_2} \cdots, p > 0, m_i, n_i > 0.$$

Then all m_i must be equal and similarly all n_i. Hence $\mathbf{s} = a^p(b^{m_1}a^{n_1})^\omega$ and $\mathbf{s}$ has not $\mathcal{R}_2$, contradiction. □

Proof. We now prove Lemma 2.2. Let $\mathbf{s}$ be an infinite word with the property $\mathcal{R}_2$. by Lemma 2.1, it has a separating letter, a for instance. Either $\mathbf{s}$ begins with a and then we write $\mathbf{s} = \psi_a(\mathbf{t})$ or $\mathbf{s}$ begins with b and we write $\mathbf{s} = \widetilde{\psi}_a(\mathbf{t})$. Let for instance $\mathbf{s} = \psi_a(\mathbf{t})$. We make a reasoning by contradiction. Suppose that $\mathbf{t}$ does not have the property $\mathcal{R}_2$.

As clearly $\mathbf{t}$ is not periodic, there exists a finite word $u \in F(\mathbf{t})$ with more than one return word in $\mathbf{t}$. As $\mathbf{t}$ has not $\mathcal{R}_2$ u has (at least) three return words in $\mathbf{t}$. If u ends with b then the occurrences of $\psi_a(u)$ in $\mathbf{s}$ are exactly the images of the occurrences of u in $\mathbf{t}$ given by the morphism. Thus $\psi_a(u)$ has three return words in $\mathbf{s}$, which leads to a contradiction.

Consequently u ends with the letter a. Consider the occurrences of ux in $\mathbf{t}$ where $x \in \mathcal{A}$ is a non specified letter. Thus all the $\psi_a(ux)$ begin with $\psi_a(u)a$. In consequence, the occurrences of $\psi_a(u)a$ in $\mathbf{s}$ are exactly the images of the occurrences of u in $\mathbf{t}$ under the morphism and then $\psi_a(u)a$ has three return words in $\mathbf{s}$. Contradiction.

Conversely, let $\mathbf{s} = \psi_a(\mathbf{t})$ and $\mathbf{t}$ has the property $\mathcal{R}_2$. Suppose that $\mathbf{s}$ does not have the property $\mathcal{R}_2$. There exists a word $u \in F(\mathbf{s})$ with at least three distinct complete return words f_1, f_2, f_3 and with minimal length.

First case, suppose that u begins with a. If u ends with b, the factorization of $\mathbf{s}$ in the code $\{a, ab\}$ shows that the occurrences of u in $\mathbf{s}$ exactly correspond to the occurrences of $v = \psi_a^{-1}(u)$ in $\mathbf{t}$. That is v has three return words and we have a contradiction. Hence we can suppose that u ends with a and write $u = u'a$. The case $u' = \varepsilon$ is clearly impossible because a which is separating has at most the return words a and ab. Then, by minimality of $|u|$, u' has exactly two return words and then $u'b$ appears in one of the f_i and $u'b \in F(\mathbf{s})$. As a is separating in $\mathbf{s}$, u' ends with a. In other words, $u = u''aa$. Thus $u = \psi_a(w)a$ with $w \in F(\mathbf{t})$. The occurrences of u in $\mathbf{s}$ exactly correspond to the occurrences of wx in $\mathbf{t}$ with non specified $x \in \mathcal{A}$. Then w has three return words in $\mathbf{t}$. Contradiction.

Second case, suppose that u begins with b. Then as a is separating the occurrences of u and au in $\mathbf{s}$ are trivially in correspondence, hence au has three return words in $\mathbf{s}$. If u ends with b, then $au = \psi_a(v)$ and, as in the first case, v has three return words in $\mathbf{t}$, a contradiction. So u ends with a and we can write $au = u'a$. If u' has three return words, as $|u'| = |u|$ we may consider u' instead of u and we are brought back to the first case.

If u' has only two return words in $\mathbf{s}$, reasoning as in the first case we get that $u'b \in F(\mathbf{s})$ whence $u' = u''a$, whence $au = \psi_a(w)a$ for some $w \in F(\mathbf{t})$. Thus w has three return words, a contradiction. □

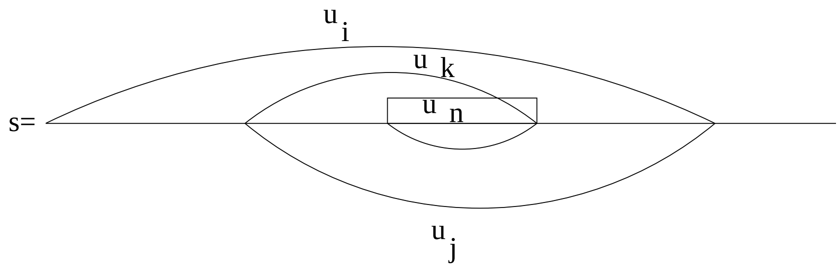

FIGURE 1. Position of u_n in $\mathbf{s}$.

Lemma 2.3 is the application to a binary alphabet of a property of episturmian words [13]. For the sake of completeness let us give an independent proof.

Proof. Proof of Lemma 2.3. If the property is false then $\mathbf{s}$ is not Sturmian hence it has a prefix u which is not balanced. Choose such a sequence $\mathbf{s}, \mathbf{s}_1, \mathbf{s}_2, \cdots$ with $|u|$ minimal. Suppose for instance $\mathbf{s} = \widetilde{\psi}_a(\mathbf{s}_1)$. Then $ux = \widetilde{\psi}_a(v)$ for some prefix v of $\mathbf{s}_1$ and $x \in \mathcal{A} \cup \{\varepsilon\}$. If $|v| < |u|$ then v is balanced, whence as $\widetilde{\psi}_a$ is a Sturmian morphism, ux is balanced, contradiction. If $|v| \geq |u|$ as $|ux| = |v| + |v|_b$ we have $|v|_b \leq 1$ thus $|u|_b \leq 1$ and u is balanced, contradiction. □

Remark 2.5. The converse of Theorem 2.4, that is: any Sturmian word has property $\mathcal{R}_2$, proved in [14], could also be proved using arguments similar to the previous ones. It also immediately follows from Corollary 4.5 hereafter.

3. Occurrences of factors in the standard episturmian words

With notations as in Section 1.3, $\mathbf{s}$ is a standard episturmian word with directive word $\Delta(\mathbf{s}) = x_1 x_2 \cdots, x_i \in \mathcal{A}$. The palindromic prefixes of $\mathbf{s}$ are $u_1 = \varepsilon, \cdots, u_{i+1} = (u_i x_i)^{(+)}$. Recall that for $a \in \mathcal{A}$, $\psi_a(a) = a$ and $\psi_a(x) = ax$ if $x \neq a$, we note the morphism $\mu_n = \psi_{x_1}\psi_{x_2}\cdots\psi_{x_n}$ and the image of x_{n+1} by this morphism $h_n = \mu_n(x_{n+1})$ with $h_0 = x_1$ and $\mu_0 = Id$. By Section 1.3 $u_{n+1} = h_{n-1}u_n$ and the h_n are prefixes of $\mathbf{s}$.

Theorem 3.1. *For a given n, vu_n is a prefix of $\mathbf{s}$ if and only if*

$$v = h_{m_1} h_{m_2} \cdots h_{m_p} \ (1)$$

with $m_1 > m_2 > \cdots > m_p \geq n-1$ (this sequence could be empty that is $v = \varepsilon$).

The disposition of the occurrences of u_n given by this theorem can be illustrated by Figure 1.

Proof. ($\Leftarrow$) By induction on the length p of the product in (1). The property is trivial for $p = 0$. We suppose that it is true for $p-1$. With $v' = h_{m_1} h_{m_2} \cdots h_{m_{p-1}}$,

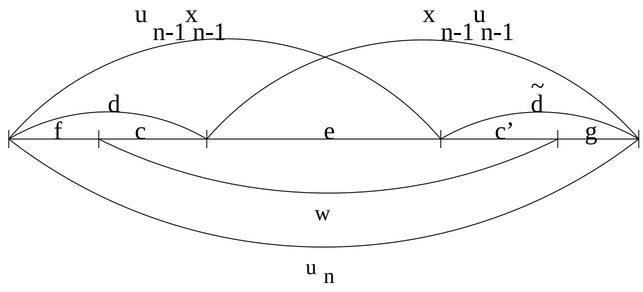

FIGURE 2. Decomposition of u_n.

we have that $v'u_{n'}$ is a prefix of $\mathbf{s}$ if $m_{p-1} \geq n' - 1$, in particular we can take $n' = m_p + 2$. But $u_{m_p+2} = h_{m_p} u_{m_p+1}$ and then $h_{m_p} u_n$ is a prefix of u_{m_p+2}. Thus we get that vu_n is a prefix of $v'u_{m_p+2}$, hence of $\mathbf{s}$.

($\Rightarrow$) We proceed by induction on n. The property is true for $n = 1$ i.e. $u_n = \varepsilon$, because any prefix of $\mathbf{s}$ can be written in the form (1) with $m_p \geq 0$ (as can easily be seen using $u_{i+1} = h_{i-1}u_i$). Suppose the property is true for $n-1$. If a is the first letter of $\mathbf{s}$ then $\mathbf{s} = \psi_a(\mathbf{s}_1)$ where $\mathbf{s}_1$ is a standard episturmian word. If we denote by u'_i, and h'_i the u_i and h_i of $\mathbf{s}_1$, we have

$$u_n = \psi_a(u'_{n-1})a \text{ and } h_n = \psi_a(h'_{n-1}).$$

As $v'u'_{n-1}$ is a prefix of $\mathbf{s}_1$ with $v' = h'_{m_1-1} \cdots h'_{m_p-1}$ it follows that vu_n is a prefix of $\mathbf{s}$ with v given by (1). □

Now let w be some factor of $\mathbf{s}$.

Lemma 3.2. *Let n be the minimal integer such that w is a factor of u_n. Then w is unioccurrent in u_n (i.e. there exists a unique pair of words $f, g \in \mathcal{A}^*$ such that $u_n = fwg$).*

Proof. We have $u_n = (u_{n-1}x_{n-1})^{(+)} = de\widetilde{d}$ where $d, e \in \mathcal{A}^*$ and $de = u_{n-1}x_{n-1}$ and e is the longest palindromic suffix of $u_{n-1}x_{n-1}$ (see Figure 2). Moreover by Lemma 1 of [7] $u_{n-1}x_{n-1}$ has a palindromic suffix unioccurrent in it and it is easily seen that this suffix is e. Consider the rightmost occurrence of w in u_n, defined by $u_n = fwg$, $f, g \in \mathcal{A}^*$. As $|fw| > |u_{n-1}|$ and $|wg| > |u_{n-1}|$, we have $w = cec'$ and $u_{n-1}x_{n-1} = de = fce$ for some $c, c' \in \mathcal{A}^*$. If there is in u_n another occurrence of w then $f'ce$ is a prefix of u_{n-1} for some word f' strictly shorter than f. Thus e has two occurrences in $u_{n-1}x_{n-1}$, a contradiction. □

Theorem 3.3. *If $\mathbf{s}$ is standard episturmian and if n is minimal such that w occurs in u_n then there exists a bijection between the occurrences of w and those of u_n in $\mathbf{s}$. More precisely, the occurrences of w are given by the prefixes vfw of $\mathbf{s}$ where v is given by (1) and f by Lemma 3.2.*

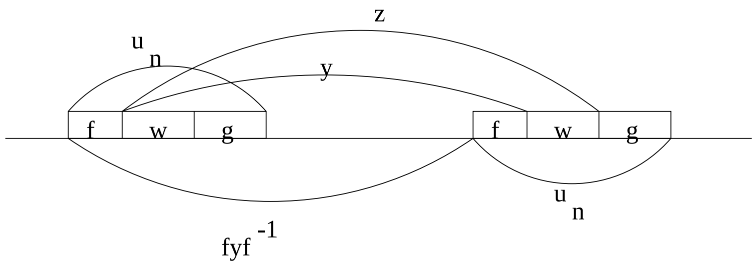

FIGURE 3. Return words on u_n and y.

Proof. Let $u_n = fwg$ be as in the preceding Lemma. If $g \neq \varepsilon$ then write $g'g'' = g$ with $0 \leq |g'| < |g|$. By construction, wg' is not right special otherwise it would be a suffix of u_n (which is right and left special) and w would have two occurrences in u_n in contradiction with the Lemma. Then any occurrence of w in $\mathbf{s}$ is followed by g. In addition to that, if $f \neq \varepsilon$ we write $f = f'f''$ with $0 \leq |f''| < |f|$ then $f''wg$ is not left special (because it would then be a prefix of $\mathbf{s}$ shorter than u_n and w would have two occurrences in u_n, in contradiction with Lemma 3.2). So each occurrence of wg in $\mathbf{s}$ is preceded by f, that is each occurrence of w is contained in an occurrence fwg of u_n. □

4. RETURN WORDS IN EPISTURMIAN WORDS

In order to study the return words of the factor w of $\mathbf{s}$ it is sufficient, by the preceding theorem, to study the return words of the corresponding u_n. More precisely we have the following trivial corollary.

Corollary 4.1. *If $w \in F(\mathbf{s})$ and u_n, f, g are those of Lemma 3.2, then y is a return word of w if and only if fyf^{-1} is a return word of u_n, and z is a complete return word of w if and only if fzg is a complete return word of u_n.*

Proof. The proof is easy (see Figure 3 in general two occurrences of u_n overlap but for clarity we draw a figure with two distinct occurrences of u_n.)). □

Proposition 4.2. *For each letter x such that $u_n x \in F(\mathbf{s})$ there exists a unique complete return word of u_n beginning with $u_n x$.*

Proof. The existence is obvious. For the unicity, suppose that there exist two complete return words beginning with $u_n x$. Clearly, no one is a prefix of the other. We can write them $u_n w w_1$ and $u_n w w_2$ where w begins with x and w_1, w_2 begin with different letters. Then $u_n w$ is right special and then it has u_n for suffix. As $w_1 \neq \varepsilon$ there exists an interior occurrence of u_n in $u_n w w_1$, this leads to a contradiction. □

Remark 4.3. By Theorem 6 of [7] $u_n x \in F(\mathbf{s})$ if and only if $x \in Alph(x_n x_{n+1} \cdots)$.

The next theorem gives a precise description of the return words of u_{n+1}.

Theorem 4.4. *The return words of the palindromic prefix u_{n+1} are the $\mu_n(x)$ where $x \in Alph(x_{n+1}x_{n+2}\cdots)$ and the corresponding complete return words of u_{n+1} are the $(u_{n+1}x)^{(+)}$. Furthermore, the derived word relative to the factors u_{n+1} is $\mathbf{s}_n = \mu_n^{-1}(\mathbf{s})$.*

Proof. Clearly, the property is true for $n = 0$ as the return words of ε are the $\mu_0(x) = x \in Alph(\mathbf{s})$. We note $u'_1 = \varepsilon$, $u'_2 = x_2, \cdots$ the palindromic prefixes of $\mathbf{s}_1 = \psi_{x_1}^{-1}(\mathbf{s})$. If $f'u'_n$ is a prefix of $\mathbf{s}_1$ then $\psi_{x_1}(f')u_{n+1}$ is a prefix of $\mathbf{s}$ because $\psi_{x_1}(u'_n)x_1 = u_{n+1}$. Conversely if fu_{n+1} is a prefix of $\mathbf{s}$ then $f = \psi_{x_1}(f')$ for some $f' \in \mathcal{A}^*$ and $f'u'_n$ is a prefix of $\mathbf{s}_1$. Thus g' is a return word of u'_n in $\mathbf{s}_1$ if and only if $\psi_{x_1}(g')$ is a return word of u_{n+1} in $\mathbf{s}$. Assuming by induction on n that $g' = \psi_{x_2}\psi_{x_3}\cdots\psi_{x_n}(x)$ with $x \in Alph(x_{n+1}x_{n+2}\cdots)$, we get $\psi_{x_1}(g') = \mu_n(x)$.

Moreover $\mu_n(x)u_{n+1}$ is a complete return word of u_{n+1}, but by a formula given in Section 1.3 it is $(u_{n+1}x)^{(+)}$.

At last if $\mathbf{s}_n = y_1y_2\cdots, y_i \in \mathcal{A}$ then $\mu_n(y_1)\mu_n(y_2)\cdots$ gives the factorization of $\mathbf{s} = \mu_n(\mathbf{s}_n)$ in return words of u_{n+1}. Thus $\mathbf{s}_n$ is the derived word relative to u_{n+1}. □

Now following [7, 13] let us say that the standard episturmian word $\mathbf{s} \in \mathcal{A}^\omega$ (or any infinite word with the same factors as $\mathbf{s}$) is $\mathcal{A}-$strict if its directive word Δ satisfies $Ult(\Delta) = Alph(\Delta) = \mathcal{A}$.

The $\mathcal{A}-$strict episturmian words are exactly the (generalized) Arnoux-Rauzy sequences on $\mathcal{A}$ whose study was begun in [2] and which can be defined as the recurrent infinite words having exactly one right- and one left-special factor of each length and with complexity function $p(n) = (\operatorname{Card}(\mathcal{A}) - 1)n + 1$. Then we have

Corollary 4.5. *For any $\mathcal{A}-$strict episturmian words (or Arnoux-Rauzy sequences on $\mathcal{A}$) each factor has exactly Card($\mathcal{A}$) return words.*

Proof. By Theorem 4.4 and as the episturmian word is $\mathcal{A}-$strict the return words of u_{n+1} are the $\mu_n(x), x \in \mathcal{A}$ whence the result. □

5. Applications

5.1. A kind of balance property

With $\mathbf{s} \in \mathcal{A}^\omega$ standard episturmian and notations as above we have

Theorem 5.1. *If $c \in \mathcal{A}$ then the factors of $\mathbf{s}$ not containing c are factors of an episturmian word on $\mathcal{A}_1 = \mathcal{A} \setminus \{c\}$.*

Proof. Suppose first that $\mathbf{s}$ is $\mathcal{A}-$strict that is $Ult(\Delta) = Alph(\Delta) = \mathcal{A}$, with $\Delta = x_1x_2\cdots$ the directive word of $\mathbf{s}$. Let x_n be the leftmost occurrence of c in Δ. Then c belongs to $u_{n+1} = u_ncu_n$ but not to u_n. By Theorem 4.4 the return words of u_{n+1} are the $\mu_n(x), x \in \mathcal{A}$. If $x = c = x_n$ then by the same Theorem the complete return word of u_{n+1} is

$$\mu_n(c)u_{n+1} = (u_{n+1}c)^{(+)} = u_ncu_ncu_n$$

whence $\mu_n(c) = u_n c$. If $x \neq c$ then

$$\mu_n(x) = \mu_{n-1}(c)\mu_{n-1}(x) = \mu_n(c)\mu_{n-1}(x) = u_n c \mu_{n-1}(x).$$

Now, consider a standard episturmian word $\mathbf{s}'$ with directive word Δ' obtained by deleting all c in Δ and denote by u'_i, μ'_i the u_i, μ_i of $\mathbf{s}'$. As $x_1 x_2 \cdots x_{n-1}$ is a prefix of Δ' we have $u'_i = u_i$ for $1 \leq i \leq n$ and $\mu_{n-1}(x) = \mu'_{n-1}(x)$ for $x \in \mathcal{A}$. Thus $\mu_n(x) = u_n c \mu'_{n-1}(x)$ for $x \neq c$. By corollary 4.1, the return words of c in $\mathbf{s}$ are cu_n and $c\mu'_{n-1}(x)u_n$, for $x \in \mathcal{A}_1 = \mathcal{A} \setminus \{c\}$.

Therefore the factors of $\mathbf{s}$ not containing c are factors of the $\mu'_{n-1}(x)u_n$ for $x \in \mathcal{A}_1$ and by Theorem 4.4 these words are the complete return words of u_n in $\mathbf{s}'$.

At last, if $\mathbf{s}$ is not $\mathcal{A}-$strict, as the return words of u_{n+1} in $\mathbf{s}$ are some of the $\mu_n(x), x \in \mathcal{A}$, it suffices to replace $\mathbf{s}$ by an $\mathcal{A}-$strict standard episturmian word whose directive word begins with $x_1 x_2 \cdots x_n$. $\square$

Theorem 5.2. *If $\mathbf{s} \in \mathcal{A}^\omega$ is standard episturmian, let $\{d, e\}$ be a two-letter subset of $\mathcal{A}$. Then for any $u, v \in F(\mathbf{s}) \cap \{d, e\}^*$ with $|u| = |v|$, we have $||u|_d - |v|_d| \leq 1$.*

Proof. Assume without loss of generality that $\mathbf{s}$ is $\mathcal{A}-$strict. If $\mathrm{Card}(\mathcal{A}) = 2$ there is nothing to prove as $\mathbf{s}$ is Sturmian. Otherwise, let c be a letter in $\mathcal{A} \setminus \{d, e\}$. Let $\mathcal{A}_1 = \mathcal{A} \setminus \{c\}$. The words in $F(\mathbf{s}) \cap \mathcal{A}_1^*$ are by Theorem 5.1 factors of a standard $\mathcal{A}_1-$strict episturmian word $\mathbf{s}'$. Deleting in the same way a letter $c' \in \mathcal{A}_1 \setminus \{d, e\}$ we get an $\mathcal{A}_2-$strict standard episturmian $\mathbf{s}''$, with $\mathcal{A}_2 = \mathcal{A}_1 \setminus \{c\}$. Continuing, we arrive at a Sturmian word on $\{d, e\}$ and this one has the balance property. $\square$

Remark 5.3. The property stated in the Theorem 5.2 is not characteristic as trivial examples show.

5.2. Recurrence function

With $\mathbf{s}$ standard episturmian, $\mathcal{A} = Alph(\mathbf{s}), \Delta(\mathbf{s})$ and the other notations as above, given any $w \in F(\mathbf{s})$, we define $W(w)$ to be the smallest integer such that every $v \in F(\mathbf{s})$ with $|v| = W(w)$ contains at least one occurrence of w (this integer exits because $\mathbf{s}$ is uniformly recurrent). The *recurrence function* $R(\ell)$ is then given by

$$R(\ell) = \sup\{W(w) | w \in F_\ell(\mathbf{s})\} \ (2).$$

This is the minimal length $R(\ell)$ such that each block of $\mathbf{s}$ of that length contains each factor of length ℓ.

Lemma 5.4. *Let r be the longest complete return word of w in s. Then $W(w) = |r| - 1$.*

Proof. Let $v \in F(\mathbf{s})$ with $|v| = |r| - 1$. If w does not occur in v then there exists a complete return word of w of the form $xvy, x, y \in \mathcal{A}^+$. As $|xvy| > |r| - 1$ we have a contradiction. Thus we have $W(w) \leq |r| - 1$.

Now the complete return word r can be written $xr'y, x, y \in \mathcal{A}$. Clearly w does not occur in r'. As $|r'| = |r| - 2$ the proof is complete. $\square$

Now let r_n (resp. r'_n) denote the longest(resp. longest complete) return word of u_n in $\mathbf{s}$. For $w \in F(\mathbf{s}) \setminus \{\varepsilon\}$, define n_w by $w \in F(u_{n_w+1}) \setminus F(u_{n_w})$, that is $n_w + 1$ is the minimal integer such that w is a factor of u_{n_w+1}.

Lemma 5.5. *If w is a factor of $\mathbf{s}$ then*

$$W(w) = |r_{n_w+1}| + |w| - 1.$$

Proof. By Lemma 3.2, we can write in a unique way $u_{n_w+1} = fwg$, $f, g \in \mathcal{A}^*$. By Corollary 4.1 the longest complete return word of w is $f^{-1} r'_{n_w+1} g^{-1}$. In consequence by Lemma 5.4 we have

$$W(w) = |r'_{n_w+1}| - |f| - |g| - 1 = |r_{n_w+1}| + |w| - 1.$$

□

Then by equation (2) we get

$$R(\ell) = \sup\{|r_{n_w+1}| \, |w \in F_\ell(s)\} + \ell - 1 \ (3).$$

In order to get a more explicit form of $R(\ell)$, let us calculate r_n for $n > 0$. For this, we give first two definitions about positions of letters in the directive word $\Delta(\mathbf{s}) = x_1 x_2 \cdots$. For $i \in \mathbb{N}_+$, let $S(i)$ be the smallest $j > i$ such that $x_j = x_i$, if it exists, $S(i)$ undefined otherwise, and let $P(i)$ be the largest $j < i$ such that $x_j = x_i$ if it exists, $P(i)$ undefined otherwise.

Lemma 5.6. *a) $|r_n|$ is a monotone increasing function of n.*

b) If some $x \in \mathcal{A}$ does not occur in u_n then $|r_n| = |u_n| + 1$. Otherwise $|r_n| = |u_n| - |u_p|$ with $p = \inf\{P(i) | i \geq n\}$.

Proof. By Theorem 4.4 $r_{n+1} = \mu_n(x)$ and $r_n = \mu_{n-1}(y)$ for some $x \in \mathcal{B} = Alph(x_{n+1} x_{n+2} \cdots)$ and $y \in \mathcal{B} \cup \{x_n\}$. Suppose by contradiction that $|r_n| > |r_{n+1}|$. By the maximality of $|r_{n+1}|$ we have $x \neq x_n$ unless $\mathcal{B} = \{x_n\}$ which would give $y = x_n$ and $r_{n+1} = r_n$, a contradiction. Thus $r_{n+1} = \mu_{n-1}(x_n x)$. If $y = x_n$ then clearly $|r_n| < |r_{n+1}|$. Otherwise $y \in \mathcal{B}$ and the maximality of $|r_{n+1}|$ implies $|\mu_{n-1}(x)| \geq |\mu_{n-1}(y)|$ whence $|r_n| < |r_{n+1}|$.

b) If $x \in \mathcal{A}$ does not occur in u_n then $(u_n x)^{(+)} = u_n x u_n$ is a longest complete return word of u_n hence $|r_n| = |u_n x| = |u_n| + 1$. If the letter x occurs in u_n then $(u_n x)^{(+)} = v u_p \widetilde{v}$ with $v u_p = u_n$ and u_p the longest palindromic prefix of u_n followed by x in u_n. Thus we have $|(u_n x)^{(+)}| = 2|u_n| - |u_p|$. The longest complete return word of u_n is obtained when $p = \inf\{P(i) | i \geq n\}$ and then $|r_n| = |u_n| - |u_p|$. □

Now let

$$D(\ell) = \sup\{n_w | w \in F_\ell(s)\}.$$

Then by part a) of Lemma 5.6 and formula (3), we get

$$R(\ell) = |r_{D(\ell)+1}| + \ell - 1 \ (4)$$

At last for obtaining $D(\ell)$, remark that if $u_{n+1} = vu_p\widetilde{v}$ with $vu_p = u_p\widetilde{v} = u_n$ then $x_p = x_n$ and $n = S(p)$. Let t be the minimal integer such that $Alph(x_1x_2\cdots x_t) = \mathcal{A}$. If $w \in F_\ell(s)$ then either $u_{n_w+1} = u_{n_w}xu_{n_w}$ for some $x \in \mathcal{A}$ not occurring in u_{n_w}, whence $n_w \le t$, or $u_{n_w+1} = vu_p\widetilde{v}$ with $n_w = S(p)$ and $w = fx_pu_px_pg$ for some $f, g \in \mathcal{A}^*$, whence $\ell \ge |u_p| + 2$.

Conversely, for any $x \in \mathcal{A}$ there exist factors of s of length $\ell \ge 1$ containing x and for any p such that $|u_p| + 2 \le \ell$ and that $S(p)$ exists, there exists $w \in F_\ell(s)$ containing $x_pu_px_p$.

Consequently for $\ell \ge 1$

$$D(\ell) = \sup(\{S(p) \mid |u_p| + 2 \le \ell\} \cup \{t\}) \ (5)$$

This achieves the determination of $D(\ell)$. Clearly D is a monotone increasing function. If $\{n_1, n_2, \cdots\}$, $n_i < n_{i+1}$, is the image of D, writing $D^{-1}(n_i) = [b_i, b_{i+1}[$, we have in conclusion

Theorem 5.7. *The recurrence function of the episturmian word* $\mathbf{s}$ *is given by*

$$R(\ell) = |r_{n_i+1}| + \ell - 1 \text{ for } \ell \in [b_i, b_{i+1}[$$

where all notations are as above.

Corollary 5.8. *The growth of* $R(\ell)$ *is linearly bounded if and only if* $S(p) - p$ *is bounded for* $p \in \mathbb{N}_+$.

Proof. If the $S(p) - p$ are bounded by M then for $\ell = |u_q| + 2$, $D(\ell) \le q + M$ whence

$$|u_{D(\ell)+1}| + 1 \le 2^{M+1}(|u_q| + 1)$$

whence by formula (4) and Lemma 5.6

$$R(|u_q| + 2) < (2^{M+1} + 1)(|u_q| + 2).$$

The proof of the only if part requires a lemma

Lemma 5.9. *If* $x_{n+1} \neq x_n$ *then* $|h_n| > |u_n|$.

Proof. Suppose first that, for some n, $u_n = h_n$ and $x_{n+1} \neq x_n$. We have $x_{n+1} = x_1$. Also $u_{n+2} = h_nu_{n+1} = h_nh_{n-1}u_n = u_nh_{n-1}u_n$. Hence h_{n-1} is a palindrome, thus its last letter x_n is x_1, in consequence $x_n = x_{n+1}$, contradiction.Thus $u_n \neq h_n$ whenever $x_{n+1} \neq x_n$.

Now suppose by contradiction $|h_n| < |u_n|$. Let u'_i, h'_i be the u_i and h_i of $\mathbf{s}_1$. Then $u_n = \psi_{x_1}(u'_{n-1})x_1$ and $h_n = \psi_{x_1}(h'_{n-1})$. As $|h_n| < |u_n|$ it follows that h'_{n-1} is a prefix of u'_{n-1} whence,as these words are different by the just above property, $|h'_{n-1}| < |u'_{n-1}|$. Passing in the same way to $\mathbf{s}_2, \cdots, \mathbf{s}_{n-1}$, we get that with evident notations $h_1^{(n-1)}$ is a prefix of $u_1^{(n-1)}$ and this is false as $u_1^{(n-1)} = \varepsilon$. □

End of the proof. Suppose $S(q) - q > M$ for arbitrarily large M. For $\ell = |u_q| + 2, D(\ell) \ge S(q)$. By $u_{D(\ell)+1} = h_{D(\ell)-1}h_{D(\ell)-2}\cdots h_qu_{q+1}$, we get $|u_{D(\ell)+1}| > M|h_q| + |u_q| > (M+1)|u_q|$, whence easily $R(\ell)/\ell$ is not bounded. □

5.3. Examples

Example 5.10. Let $\mathbf{s}$ be standard Sturmian with directive word $\Delta(\mathbf{s}) = a^{e_1} b^{e_2} a^{e_3} \cdots, e_i > 0$. It is well known that the continued fraction expansion of the slope $\alpha < 1/2$ of $\mathbf{s}$ is $[0, e_1+1, e_2, \cdots]$. Denote by $q_0 = 1$, $q_1 = e_1 + 1, \cdots q_{j+1} = e_{j+1}q_j + q_{j-1}, \cdots$ the denominators of the convergents.

Let, for $j \geq 1, L_j = e_1 + e_2 + \cdots + e_j$. Then $x_{n+1} \neq x_n$ if and only if n is some L_j. We deduce $S(L_j) = L_{j+1} + 1$ and $P(L_{j+1}+1) = L_j$. It follows that, for $|u_{L_j}| + 2 \leq \ell < |u_{L_{j+1}}| + 2$, we have by equation (5) $D(\ell) = L_{j+1} + 1$. Then using Lemma 5.6 with $n = D(\ell) + 1$, we get $|r_n| = |u_n| - |u_p|$ where $p = \inf\{P(i) | i \geq D(\ell) + 1\} = L_{j+1}$. Thus by equation (4), we have

$$R(\ell) = |u_{L_{j+1}+2}| - |u_{L_{j+1}}| + \ell - 1.$$

It is easily seen that $u_{L_{j+1}+2} = h_{L_{j+1}} h_{L_{j+1}-1} u_{L_{j+1}} = h_{L_{j+1}} h_{L_j} u_{L_{j+1}}$. It can also be shown that the h_{L_j} satisfy the same recurrence relation as the q_j, whence $h_{L_j} = q_j$. Moreover, by a known property of Sturmian words, $|u_{L_j}| = q_j - 2$ whence at last the known formula

$$R(\ell) = q_{j+1} + q_j + \ell - 1 \text{ for } q_j \leq \ell < q_{j+1}.$$

Example 5.11. In the general case $\Delta(\mathbf{s}) = y_1^{e_1} y_2^{e_2} \cdots, e_i > 0, y_i \in \mathcal{A}, y_{i+1} \neq y_i$. When the sequence $y_1 y_2 \cdots$ is periodic, $R(\ell)$ is given by rather simple formula recalling the Sturmian case. Let us consider only here the simplest case: $\mathbf{s} = abacaba \cdots$ is the Rauzy word, also called Tribonacci word, having directive word $(abc)^\omega$. Clearly $S(i) = i + 3$ and $P(i+3) = i$ whence easily

$$R(\ell) = |u_{j+4}| - |u_{j+1}| + \ell - 1 = |h_{j+3}| + \ell - 1$$

for $|u_j| + 2 \leq \ell < |u_{j+1}| + 2$.

References

[1] J.-P. Allouche, J. L. Davison, M. Queffélec and L. Q. Zamboni *Transcendence of Sturmian or morphic continued fractions* preprint 1999.

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

[3] J. Berstel and P. Séébold , *Sturmian words,* In M. Lothaire, Algebraic combinatorics on Words. (1999). To appear.

[4] J. Cassaigne, *Ideas for a proof of Rauzy's conjecture on the recurrence functions of infinite words,* talk in Rouen "words 1999".

[5] N. Chekova, P. Hubert and A .Messaoudi, *Propriétés combinatoires,ergodiques et arithmétiques de la substitution de Tribonacci,* prepublication 98-24 IML.

[6] A. de Luca, *Sturmian words: Structure, Combinatorics and their Arithmetics*, Theoret. Comput. Sci. 183 (1997) 45-82.

[7] X. Droubay, J. Justin and G. Pirillo *Episturmian words and some constructions of Rauzy and de Luca,* Theoret. Comput. Sci.(2000), to appear.

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

[9] F. Durand, *Contributions à l'étude des suites et systèmes dynamiques substitutifs,* Ph.D. Thesis, Université de la Méditerranée (Aix-Marseille II) 1996.

[10] I. Fagnot and L. Vuillon,*Generalized balances in Sturmian words,* prepublication LIAFA 2000/02.

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

[12] C. Holton, L. Q. Zamboni *Geometric realizations of substitutions*, Bull. Soc. Math. France **126** (1998), no. 2, 149–179.

[13] J. Justin and G. Pirillo *Episturmian words and episturmian morphisms,* prepublication LIAFA 2000/23.

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

Communicated by (The editor will be set by the publisher).
April 21, 2000.