SEQUENCES WITH CONSTANT NUMBER OF RETURN WORDS

ĽUBOMÍRA BALKOVÁ, EDITA PELANTOVÁ, AND WOLFGANG STEINER

ABSTRACT. An infinite word has the property R_m if every factor has exactly m return words. Vuillon showed that R_2 characterizes Sturmian words. We prove that a word satisfies R_m if its complexity function is $(m-1)n+1$ and if it contains no weak bispecial factor. These conditions are necessary for $m=3$, whereas for $m=4$ the complexity function need not be $3n+1$. A new class of words satisfying R_m is given.

1. INTRODUCTION

Recently, return words have been intensively studied in (symbolic) dynamical systems, combinatorics on words and number theory. Roughly speaking, for a given factor w of an infinite word u, a return word of w is a word between two successive occurrences of the factor w. This can be seen as a symbolic version of the first return map in a dynamical system. This notion was introduced by Durand [5] to give a nice characterization of primitive substitutive sequences. A slightly different notion of return words was used by Ferenczi, Mauduit and Nogueira [9].

Date: October 27, 2009.

2000 *Mathematics Subject Classification.* 68R15.

Key words and phrases. return words; complexity; special factors; Sturmian words.

Sturmian words are aperiodic words over a biliteral alphabet with the lowest possible factor complexity; they were defined by Morse and Hedlund [12]. Using return words, Vuillon [14] found a new equivalent definition of Sturmian words. He showed that an infinite word u over a biliteral alphabet is Sturmian if and only if any factor of u has exactly two return words. A short proof of this fact is given in Section 5.

A natural generalization of Sturmian words to m-letter alphabets is constituted by infinite words with every factor having exactly m return words. This property is called R_m. It covers other generalizations of Sturmian words: Justin and Vuillon [11] proved that Arnoux-Rauzy words of order m satisfy R_m, Vuillon [15] proved this property for words coding regular m-interval exchange transformations.

The factor complexity, i.e., the number of different factors of length n, of the two classes of words with property R_m in the preceding paragraph is $(m-1)n+1$ for all $n \geq 0$. Vuillon [15] observed that this condition is not sufficient to describe words satisfying R_m, $m \geq 3$: the fixed point of a certain recoding of the Chacon substitution, which has complexity $2n+1$ by Ferenczi [7], has factors with more than 3 return words.

A deeper inspection of the two classes of words with property R_m shows that not only the first difference of complexity is constant, but also that the bilateral order of every factor (see Cassaigne [4] and Section 4) is zero. We show that this condition is indeed sufficient to have the property R_m, and provide a less known class of words satisfying this condition. If a word satisfies R_3, then we can show that no factor is weak bispecial, i.e., no factor has negative bilateral order. Therefore the words with R_3 are characterized by complexity $2n+1$ and the absence of weak bispecial factors.

In Section 6.1, we provide a word satisfying R_4 with an even number of factors of every positive length (containing infinitely many weak bispecial factors). Therefore words satisfying R_m do not necessarily have complexity $(m-1)n+1$, and it is an open question whether there exists a nice characterization of words satisfying R_m for $m \geq 4$.

In this article we focus only on the number of return words corresponding to a given factor of an infinite word. We do not study the ordering of return words in the infinite word, i.e., we do not study derivated sequences (see [5] for the precise definition). Let us just mention here that a derivated sequence of a word with property R_m is again a word satisfying R_m. A description of derivated sequences of Sturmian words can be found in [1].

2. Basic definitions

An *alphabet* $\mathcal{A}$ is a finite set of symbols called *letters*. A (possibly empty) concatenation of letters is a *word*. The set $\mathcal{A}^*$ of all finite words provided with the operation of concatenation is a free monoid. The *length* of a word w is denoted by $|w|$. A finite word w is called a *factor* (or *subword*) of the (finite or right infinite) word u if there exist a finite word v and a word v' such that $u = vwv'$. The word w is a *prefix* of u if v is the empty word. Analogously, w is a *suffix* of u if v' is the empty word. We say that a prefix (suffix) w of u is *proper* if $w \neq u$. A concatenation of k words w will be denoted by w^k.

The *language* $\mathcal{L}(u)$ is the set of all factors of the word u, and $\mathcal{L}_n(u)$ is the set of all factors of u of length n. Let w be a factor of an infinite word u and let $a, b \in \mathcal{A}$. If aw is a factor of u, then we call a a *left extension* of w. Analogously, we call b a *right extension* of w if $wb \in \mathcal{L}(u)$. We will denote by $\mathcal{E}_\ell(w)$ the set of all left extensions of w, and by $\mathcal{E}_r(w)$ the set of right extensions. A factor w is *left special* if $\#\mathcal{E}_\ell(w) \geq 2$, *right special* if $\#\mathcal{E}_r(w) \geq 2$ and *bispecial* if w is both left special and right special.

Let w be a factor of an infinite word $u = u_0u_1 \cdots$ (with $u_j \in \mathcal{A}$), $|w| = \ell$. An integer j is called an *occurrence* of w in u if $u_ju_{j+1} \cdots u_{j+\ell-1} = w$. Let j, k, $j < k$, be successive occurrences of w. Then $u_ju_{j+1} \cdots u_{k-1}$ is a *return word* of w. The set of all return words of w is denoted by $\mathcal{R}(w)$,

$$\mathcal{R}(w) = \{u_ju_{j+1} \dots u_{k-1} \mid j, k \text{ being successive occurrences of } w \text{ in } u\}.$$

If v is a return word of w, then vw is called *complete return word.*

An infinite word is *recurrent* if any of its factors occurs infinitely often or, equivalently, if any of its factors occurs at least twice. It is *uniformly recurrent* if, for any $n \in \mathbb{N}$, every sufficiently long factor contains all factors of length n. It is not difficult to see that a recurrent word on a finite alphabet is uniformly recurrent if and only if the set of return words of any factor is finite.

The variability of local configurations in u is expressed by the *factor complexity function* (or simply *complexity*) $C(n) = \#\mathcal{L}_n(u)$. It is well known that a word u is aperiodic if and only if $C(n) \geq n + 1$ for all $n \in \mathbb{N}$ (see [12]). Infinite aperiodic words with the minimal complexity $C(n) = n + 1$ for all $n \in \mathbb{N}$ are called *Sturmian words.* These words have been studied extensively, and several equivalent definitions of Sturmian words can be found in Berstel [3].

3. Simple facts for return words

3.1. **Restriction to bispecial factors.** If a factor w is not right special, i.e., if it has a unique right extension $b \in \mathcal{A}$, then the sets of occurrences of w and wb coincide, and

$$\mathcal{R}(w) = \mathcal{R}(wb).$$

If a factor w has a unique left extension $a \in \mathcal{A}$, then $j \geq 1$ is an occurrence of w in the infinite word u if and only if $j - 1$ is an occurrence of bw. This statement does not hold for $j = 0$. Nevertheless, if u is a recurrent infinite word, then the set of return words of w stays the same no matter whether we include the return word corresponding to the prefix w of u or not. Consequently, we have

$$\mathcal{R}(aw) = a\mathcal{R}(w)a^{-1} = \{ava^{-1} \mid v \in \mathcal{R}(w)\},$$

where ava^{-1} means that the word v is prolonged to the left by the letter a and it is shortened from the right by erasing the letter a (which is always a suffix of v for $v \in \mathcal{R}(w)$).

For an aperiodic uniformly recurrent infinite word u, each factor w can be extended to the left and to the right to a bispecial factor. To describe the cardinality and the structure of $\mathcal{R}(w)$ for arbitrary w, it suffices therefore to consider bispecial factors w.

3.2. **Tree of return words.** It is convenient to consider a tree constructed in the following way: Label the root with a factor w, and attach $\#\mathcal{E}_r(w)$ children, with labels wb, $b \in \mathcal{E}_r(w)$. Repeat this recursively with every node labeled by v, except if w is a suffix of v. If u is uniformly recurrent, then this algorithm stops, and it is easy to see that the labels of the leaves of this tree are exactly the complete return words of w. Therefore we have

$$\#\mathcal{R}(w) = \#\{\text{leaves}\} = 1 + \sum_{\text{non-leaves } v} (\#\mathcal{E}_r(v) - 1). \tag{1}$$

In particular, if w is the unique right special factor of its length, then $\#\mathcal{R}(w) = \#\mathcal{E}_r(w)$.

Figure 1. Tree of return words of 01 in the Thue-Morse sequence, and a more compact representation by a trie.

A similar construction can be done with left extensions, yielding similar formulae. Since we can restrict our attention to bispecial factors w by Section 3.1, we obtain the following proposition.

Proposition 3.1. *Let u be a recurrent word and $m \in \mathbb{N}$. Suppose that for every $n \in \mathbb{N}$ at least one of the following conditions is satisfied:*

- *There is a unique left special factor $w \in \mathcal{L}_n(u)$, and $\#\mathcal{E}_\ell(w) = m$.*
- *There is a unique right special factor $w \in \mathcal{L}_n(u)$, and $\#\mathcal{E}_r(w) = m$.*

Then u satisfies R_m, i.e., every factor has exactly m return words.

Recall that Arnoux-Rauzy words of order m are defined as uniformly recurrent infinite words which have for every $n \in \mathbb{N}$ exactly one right special factor w of length n with $\#\mathcal{E}_r(w) = m$ and exactly one left special factor w of length n with $\#\mathcal{E}_\ell(w) = m$. They are also called strict episturmian words. It is easy to see that Sturmian words are recurrent, and we obtain the following corollary to Proposition 3.1.

Corollary 3.2. *Arnoux-Rauzy words of order m satisfy R_m, in particular Sturmian words satisfy R_2.*

4. Sufficient conditions for property R_m

This section is devoted to sufficient conditions for a word u having the property R_m, but we mention first two evident necessary conditions.

The alphabet $\mathcal{A}$ of u must have m letters since the occurrences of the empty word are all integers $n \geq 0$, and its return words are therefore all letters u_n. Furthermore, u must be uniformly recurrent since every factor has a return word and only finitely many of them.

An important role in our further considerations is played by weak bispecial factors.

Definition 4.1. *A factor w of a recurrent word is* weak bispecial *if $B(w) < 0$, where*

$$B(w) = \#\{awb \in \mathcal{L}(u) \mid a, b \in \mathcal{A}\} - \#\mathcal{E}_\ell(w) - \#\mathcal{E}_r(w) + 1$$

is the bilateral order *of w.*

Since $\#\{awb \in \mathcal{L}(u) \mid a, b \in \mathcal{A}\} = \sum_{a \in \mathcal{E}_\ell(w)} \#\mathcal{E}_r(aw) = \sum_{b \in \mathcal{E}_r(w)} \#\mathcal{E}_\ell(wb)$, the inequality $B(w) < 0$ is equivalent to

$$\sum_{a \in \mathcal{E}_\ell(w)} (\#\mathcal{E}_r(aw) - 1) < \#\mathcal{E}_r(w) - 1$$

and to

$$\sum_{b \in \mathcal{E}_r(w)} (\#\mathcal{E}_\ell(wb) - 1) < \#\mathcal{E}_\ell(w) - 1.$$

The bilateral order was defined by Cassaigne [4] in order to calculate the second complexity difference. If we set $\Delta C(n) = C(n+1) - C(n)$, then we have

$$\Delta C(n) = \sum_{w \in \mathcal{L}_n(u)} \big(\#\mathcal{E}_\ell(w) - 1\big) = \sum_{w \in \mathcal{L}_n(u)} \big(\#\mathcal{E}_r(w) - 1\big)$$

and therefore

$$\Delta C(n+1) - \Delta C(n) = \sum_{w\in\mathcal{L}_n(u)} \sum_{a\in\mathcal{E}_\ell(w)} (\#\mathcal{E}_r(aw) - 1) - \sum_{w\in\mathcal{L}_n(u)} (\#\mathcal{E}_r(w) - 1)$$

$$= \sum_{w\in\mathcal{L}_n(u)} \big(\#\{awb \in \mathcal{L}(u) \mid a, b \in \mathcal{A}\} - \#\mathcal{E}_\ell(w) - \#\mathcal{E}_r(w) + 1\big) = \sum_{w\in\mathcal{L}_n(u)} B(w).$$

If $B(w) = 0$ for all factors w, then the first complexity difference is constant. If no factor is weak bispecial, then $\Delta C(n)$ is non-decreasing. Since $\Delta C(0) = \#\mathcal{A} - 1$ and $\#A = m$, we obtain the following lemma.

Lemma 4.2. *If u satisfies R_m and no factor is weak bispecial, then $\Delta C(n) \geq m-1$ for all $n \geq 0$.*

The number of return words can be bounded by the following lemmas.

Lemma 4.3. *If u is a uniformly recurrent word with no weak bispecial factor, then*

$$\#\mathcal{R}(w) \geq 1 + \Delta C(|w|)$$

for every factor $w \in \mathcal{L}(u)$.

Proof. Let $w \in \mathcal{L}(u)$ and denote by $v_1, v_2, \ldots, v_r$ the right special factors of length $|w|$. Since no factor is weak bispecial and u is uniformly recurrent, every v_j can be extended to the left without decreasing the total amount of "right branching" until w is reached. More precisely, we have (mutually different) right special factors $v_j^{(1)}, v_j^{(2)}, \ldots, v_j^{(s_j)}$ with suffix v_j, prefix w and no other occurrence of w such that $\#\mathcal{E}_r(v_j) - 1 \leq \sum_{i=1}^{s_j}(\#\mathcal{E}_r(v_j^{(i)}) - 1)$. Since all $v_j^{(i)}$ are nodes in the tree of return words and $v_j^{(i)} \neq v_{j'}^{(i')}$ if $(j,i) \neq (j',i')$, we can use (1) and obtain

$$\#\mathcal{R}(w) \geq 1+\sum_{j=1}^{r}\sum_{i=1}^{s_j}(\#\mathcal{E}_r(v_j^{(i)})-1) \geq 1+\sum_{j=1}^{r}(\#\mathcal{E}_r(v_j)-1) = 1+\Delta C(|w|).$$

□

Lemma 4.4. *If u has no weak bispecial factor and $\Delta C(n) < m$ for all $n \geq 0$, then*

$$\#\mathcal{R}(w) \leq m$$

for every factor $w \in \mathcal{L}(u)$.

Proof. Let $v_1, v_2, \dots, v_r$ denote the right special factors which are labels of non-leave nodes in the tree of return words of w, and $n = \max_{1\leq j\leq r} |v_j|$. Since no bispecial factor is weak, every v_j can be extended to the left to factors of length n without decreasing the total amount of "right branching". More precisely, we have (mutually different) right special factors $v_j^{(1)}, v_j^{(2)}, \dots, v_j^{(s_j)}$ of length n with suffix v_j such that $\#\mathcal{E}_r(v_j) - 1 \leq \sum_{i=1}^{s_j}(\#\mathcal{E}_r(v_j^{(i)}) - 1)$. Since w occurs in v_j only as prefix, no v_j can be a proper suffix of $v_{j'}$. Hence we have $v_j^{(i)} \neq v_{j'}^{(i')}$ if $(j,i) \neq (j',i')$ and

$$\begin{aligned}\#\mathcal{R}(w) &= 1 + \sum_{j=1}^{r}\big(\#\mathcal{E}_r(v_j) - 1\big) \leq 1 + \sum_{j=1}^{r}\sum_{i=1}^{s_j}\big(\#\mathcal{E}_r(v_j^{(i)}) - 1\big)\\ &\leq 1 + \Delta C(n) \leq m.\end{aligned}$$

□

For words with no weak bispecial factors, these three lemmas give a very simple characterization of the property R_m.

Theorem 4.5. *If u is a uniformly recurrent word with no weak bispecial factor, then it satisfies R_m if and only if $C(n) = (m-1)n + 1$ for all $n \geq 0$.*

5. Properties R_2 and R_3

For $m = 2$ and $m = 3$, we can completely characterize the words with property R_m.

Definition 5.1. *Let v be a return word of $w \in \mathcal{L}(u)$. We say that the return word v starts with b if wb is a prefix of the complete return word vw and that it ends with a if aw is a suffix of vw.*

A right special factor w is called *maximal right special* if w is not a proper suffix of any right special factor, i.e., $\sum_{a\in\mathcal{E}_\ell(w)}(\#\mathcal{E}_r(aw)-1)=0$. Any maximal right special factor is therefore weak bispecial.

Lemma 5.2. *If $w \in \mathcal{L}(u)$ is a maximal right special factor such that for any $b \in \mathcal{E}_r(w)$ there exists a unique $v \in \mathcal{R}(w)$ starting with b, then u is eventually periodic.*

Proof. Denote the return words of w by $v_1, v_2, \ldots, v_r$, where, w.l.o.g., v_j starts with b_j, ends with a_j and b_{j+1} is the only letter in $\mathcal{E}_r(a_jw)$ for $1 \le j < r$. Then b_1 is the only letter in $\mathcal{E}_r(a_rw)$ and $u = p(v_1v_2\cdots v_r)^\infty$ for some prefix p. □

Corollary 5.3. *If u satisfies R_2, then it has no maximal right special factor.*

Proof. Assume that w is a maximal right special factor. Then the two return words of w have different starting letters, hence u is eventually periodic by Lemma 5.2 and $\#\mathcal{R}(wa) = 1$. □

On a binary alphabet, the notions "weak bispecial" and "maximal right special" coincide. Therefore Corollaries 3.2, 5.3 and Lemma 4.3 provide a short proof of the following theorem.

Theorem 5.4 (Vuillon [14]). *An infinite word u satisfies R_2 if and only if it is Sturmian.*

For words with property R_3, we need the following lemma.

Lemma 5.5. *Let w be a weak bispecial factor with a unique $a \in \mathcal{E}_\ell(w)$ such that more than one return word of w starts with a letter in $\mathcal{E}_r(aw)$, then $\#\mathcal{R}(aw) < \#\mathcal{R}(w)$.*

Proof. Any return word of aw is of the form $av_1v_2\cdots v_r a^{-1}$ for some $r \geq 1$ and $v_j \in \mathcal{R}(w)$, $1 \leq j \leq r$. If v_1 ends with a, then $r = 1$. If v_1 ends with $a' \neq a$, then the assumption of the lemma implies that there is a unique return word of w starting with a letter in $\mathcal{E}_r(a'w)$ (and $\#\mathcal{E}_r(a'w) = 1$). Therefore v_2 and inductively the sequence of words $v_2, \ldots, v_r$ are completely determined by the choice of v_1. This implies that $\#\mathcal{R}(aw)$ equals the number of return words of w starting with a letter in $\#\mathcal{E}_r(aw)$. Since w is weak bispecial, we have $\#\mathcal{E}_r(aw) < \#\mathcal{E}_r(w)$ and thus $\#\mathcal{R}(aw) < \#\mathcal{R}(w)$. □

Remark. There are two cases for Lemma 5.5: Either aw is right special or there is more than one return word of w starting with the unique right extension of aw.

Corollary 5.6. *If u satisfies R_3, then it has no weak bispecial factor.*

Proof. Assume that w is a weak bispecial factor. Since u is uniformly recurrent the problem is symmetric, and we may assume, w.l.o.g., $\#\mathcal{E}_\ell(w) \leq \#\mathcal{E}_r(w)$.

If $\#\mathcal{E}_r(w) = 3$, then every return word of w starts with a different letter in $\mathcal{E}_r(w)$. Since at most for one $a \in \mathcal{E}_\ell(w)$, the factor aw is right special, we obtain a contradiction to R_3 by Lemma 5.2 or 5.5.

If $\#\mathcal{E}_r(w) = 2$, then $\mathcal{E}_r(aw) = \{b\}$ and $\mathcal{E}_r(a'w) = \{b'\}$. Since, w.l.o.g., two return words of w start with b and one starts with b', we obtain a contradiction to R_3 by Lemma 5.5. □

By combining Corollary 5.6 and Theorem 4.5, we obtain the following theorem.

Theorem 5.7. *A uniformly recurrent word u satisfies R_3 if and only if $C(n) = 2n+1$ for all $n \geq 0$ and u has no weak bispecial factor.*

Remarks.

- The theorem remains true if "weak bispecial" is replaced by "maximal right special": If $\Delta C(n) = 2$ for all $n \geq 0$, then every factor w with $\#\mathcal{E}_r(w) = 3$ is the unique right special factor of its length, and it cannot be weak bispecial. If $\#\mathcal{E}_r(w) = 2$, then the two notions coincide.
- By symmetry, "weak bispecial" can be replaced by "maximal left special".
- The condition on weak bispecial factors cannot be omitted. Ferenczi [7] showed that the fixed point $\sigma^\infty(1)$ of the substitution given by $\sigma : 1 \mapsto 12, 2 \mapsto 312, 3 \mapsto 3312$, a recoding of the Chacon substitution, has complexity $2n+1$ and it contains weak bispecial factors.

6. Property R_4

6.1. **A word with complexity $\neq 3n+1$.** The following proposition shows that $C(n)$ need not be $(m-1)n+1$ for all $n \geq 0$ if u satisfies R_m.

Proposition 6.1. *Define the substitution σ by*

$$\begin{aligned} \sigma : 1 &\mapsto 13231 & 4 &\mapsto 42324 \\ 2 &\mapsto 13231424131 & 3 &\mapsto 42324131424 \end{aligned}$$

Then the fixed point $\sigma^\infty(1)$ satisfies R_4.

Proof. By Section 3.1, it is sufficient to consider bispecial factors of $u = \sigma^\infty(1)$. The factors of length 2 are $\mathcal{L}_2(u) = \{13, 14, 23, 24, 31, 32, 41, 42\}$. For the bispecial factors $1, 2, 23, 2413$, the return words can be determined easily:

$$\mathcal{R}(1) = \{13, 1323, 1424, 142324\}$$

$$\mathcal{R}(2) = \{23, 2314, 2413, 241314\}$$

$$\mathcal{R}(23) = \{2314, 2314241314, 232413, 232413142413\}$$

$$\mathcal{R}(2413) = \{241314, 24131423, 24132314, 2413231423\}$$

The language of u is closed under the morphism φ defined by $\varphi : 1 \leftrightarrow 4$, $2 \leftrightarrow 3$, since $\sigma\varphi(w) = \varphi\sigma(w)$ for all factors w. Therefore we have $\mathcal{R}(\varphi(w)) = \varphi(\mathcal{R}(w))$.

The only factors of the form $a1b$, $a, b \in \mathcal{A}$ are 314 and 413, hence 1 is a weak bispecial factor, and $1, 4$ are the only bispecial factors with prefix or suffix 1 or 4. Similarly, 23 and 32 are weak bispecial factors and no other bispecial factor has prefix or suffix 23 or 32.

The return words of the weak bispecial factor 2413142 are factors of $\sigma(v)$, with a factor v of length $|v| \geq 2$ having prefix 2 or 3, suffix 2 or 3 and no other occurrence of 2 and 3. Since the only possibilites for v are $23, 2413, 32, 3142$, we obtain

$$\mathcal{R}(2413142) = \{24131423, 241314232413231423,$$
$$24131424132314, 2413142413231423241323 14\}.$$

All remaining bispecial factors w have prefix 24132 or 31423 and suffix 23142 or 32413, and therefore a decomposition $w = t\,\sigma(v)\,t'$ with $t \in \{24, 31\}$, $t' \in \{1323142, 4232413\}$ and a unique bispecial factor v. If v is empty, then we have w.l.o.g. $w = 241323142$ and

$$\mathcal{R}(w) = \{2413231423, 2413231423241314, \\ 2413231424131423, 2413231424131423241314242\}.$$

If v is not empty, then the uniqueness of v implies that the set of complete return words of w is $t\,\sigma(\mathcal{R}(v)v)\,t'$. Since v is shorter than w, we obtain inductively that all bispecial factors have exactly 4 return words. □

6.2. **Weak bispecial factors.** The preceding example shows that weak bispecial factors cannot be excluded in words u satisfying R_4. Nevertheless, we can show that the existence of a weak bispecial factor imposes strong restrictions on the structure of the word u.

Lemma 6.2. *Let w be a weak bispecial factor of a word u satisfying R_4. Then there exist factors $w_1, w_2 \in \mathcal{A}w \cup w\mathcal{A}$ and v_1, v_2, v_3, v_4 such that*

$$\mathcal{R}(w_1) = \{v_1v_3, v_1v_4, v_2v_3, v_2v_4\} \text{ and } \mathcal{R}(w_2) = \{v_3v_1, v_3v_2, v_4v_1, v_4v_2\}. \tag{2}$$

Proof. Let w be a weak bispecial factor. In the proof, we will use substantially the relation

$$\sum_{a\in\mathcal{E}_\ell(w)} (\#\mathcal{E}_r(aw) - 1) < \#\mathcal{E}_r(w) - 1 \tag{3}$$

and the consequence of Lemma 5.5 that there must be at least two letters $a \in \mathcal{E}_\ell(w)$ such that at least two return words of w start with a letter in $\mathcal{E}_r(aw)$.

Note that the property R_m forces $\#\mathcal{E}_r(w) \leq m$ and $\#\mathcal{E}_\ell(w) \leq m$. Since the problem is symmetric, assume w.l.o.g. $4 \geq \#\mathcal{E}_r(w) \geq \#\mathcal{E}_\ell(w) \geq 2$. We have three different situations:

- $\#\mathcal{E}_r(w) = 2$: Let $\mathcal{E}_\ell(w) = \{a_1, a_2\}$. According to (3), we have $\#\mathcal{E}_r(a_1w) = \#\mathcal{E}_r(a_2w) = 1$. Let b_1 be the unique letter in $\mathcal{E}_r(a_1w)$ and b_2 be the unique right extension of a_2w. By Lemma 5.5, there exist two return words of w starting with b_1 and two return words of w starting with b_2. Set $w_1 = wb_1$, $w_2 = wb_2$.
- $\#\mathcal{E}_r(w) = 3$: There exists a unique letter $b_1 \in \mathcal{E}_r(w)$ such that two return words of w start with b_1. As w is weak bispecial, the inequality (3) gives
$$\sum_{a\in\mathcal{E}_\ell(w)} (\#\mathcal{E}_r(aw) - 1) \le 1\,.$$
If all aw have a unique right extension, then the letter $a_1 \in \mathcal{E}_\ell(wb_1)$ is the unique letter for which at least two return words start with a letter in $\mathcal{E}_r(aw)$, which is not possible by Lemma 5.5.

  Therefore there exists a unique $a_1 \in \mathcal{E}_\ell(w)$ with $\#\mathcal{E}_r(a_1w) = 2$, and $\#\mathcal{E}_r(aw) = 1$ for all $a \in \mathcal{E}_\ell(w) \setminus \{a_1\}$. According to Lemma 5.5, there exists a letter $a_2 \neq a_1$ such that at least two return words start with a letter in $\mathcal{E}_\ell(a_2w)$. This implies $b_1 \in \mathcal{E}_r(a_2w)$ and thus $b_1 \notin \mathcal{E}_r(a_1w)$. Set $w_1 = a_1w$, $w_2 = wb_1$.
- $\#\mathcal{E}_r(w) = 4$: For every $b \in \mathcal{E}_r(w)$, there is a unique return word of w starting with b. By Lemma 5.5, we have $a_1, a_2 \in \mathcal{E}_\ell(w)$ with $\#\mathcal{E}_r(a_iw) \ge 2$. The inequality (3) for this case implies that $\#\mathcal{E}_r(a_iw) = 2$. Set $w_1 = a_1w$, $w_2 = a_2w$.

Consider "complete return words of the set $\{w_1, w_2\}$": words which have either w_1 or w_2 as prefix, either w_1 or w_2 as suffix, and no other occurrence of w_1 and w_2. By the definitions of w_1 and w_2, there are exactly two such words $v_1w_{i_1}, v_2w_{i_2}$ with prefix w_1 and two words $v_3w_{i_3}, v_4w_{i_4}$ with prefix w_2.

If $i_1 = i_2 = 2$ and $i_3 = i_4 = 1$, then R_4 implies that (2) holds.

If $i_1 = i_2 = 1$, then w_1 has only the two return words v_1, v_2. If $i_2 = i_3 = i_4 = 1$, then the return words of w_1 are v_1v_3, v_1v_4, v_2. Similarly, $i_3 = i_4 = 2$ and $i_1 = i_2 = i_3 = 2$ are not possible.

The only remaining case is $i_1 = i_4 = 1$, $i_2 = i_3 = 2$. Then the return words of w_1 are v_1 and $v_2v_3^{r_i}v_4$, $i \in \{1,2,3\}$, $0 \le r_1 < r_2 < r_3$. The return words of w_2 are v_3 and $v_4v_1^{s_i}v_2$, $i \in \{1,2,3\}$, $0 \le s_1 < s_2 < s_3$.

The return words of v_2w_2 are therefore of the form $v_2v_3^{r_i}v_4v_1^{s_j}$. Let S_1 be the set of these 4 pairs (r_i, s_j). Similarly, let S_2 be the set of the 4 pairs (s_i, r_j) such that $v_4v_1^{s_i}v_2v_3^{r_j}$ is a return word of v_4w_1.

We show that there must be some $i \in \{1,2,3\}$ such that $(r_i, s_2) \in S_1$ and $(r_i, s_3) \in S_1$, by considering the return words of $v_1^{s_2}w_1$ and of $v_1^{s_2}v_2w_2$. The return words of $v_1^{s_2}v_2w_2$ are of the form $v_1^{s_2}v_2tv_3^{r_i}v_4v_1^{s_j-s_2}$ with $t \in (v_3^*v_4v_1^{s_1}v_2)^*$, $i \in \{1,2,3\}$ and $j \in \{2,3\}$. For these t and r_i, $v_1^{s_2}v_2tv_3^{r_i}v_4$ is a return word of $v_1^{s_2}w_1$. If there was no r_i with $(r_i, s_2) \in S_1$ and $(r_i, s_3) \in S_1$, then these words would provide 4 different return words of $v_1^{s_2}w_1$, wich contradicts R_4 since v_1 is another return word.

Similarly, we must have some $i \in \{1,2,3\}$ such that $(s_i, r_2) \in S_2$ and $(s_i, r_3) \in S_2$. By considering the return words of $v_1^{s_2}w_1$ and $v_4v_1^{s_2}w_1$, we obtain as well the existence of some $i \in \{1,2,3\}$ such that $(r_2, s_i) \in S_1$ and $(r_3, s_i) \in S_1$. Finally, we must also have some $i \in \{1,2,3\}$ such that $(s_2, r_i) \in S_2$ and $(s_3, r_i) \in S_2$.

Putting everything together, we have two possibilities for S_1. Either it contains (r_1, s_1) and no other pair (r_i, s_j) with $i = 1$ or $j = 1$, or $S_1 = \{(r_1, s_2), (r_1, s_3), (r_2, s_1), (r_3, s_1)\}$. Similarly, S_2 contains (s_1, r_1) and no other pair (s_1, r_j) or (s_i, r_1), or $S_2 = \{(s_1, r_2), (s_1, r_3), (s_2, r_1), (s_3, r_1)\}$.

If $(r_1, s_1) \in S_1$ and $(s_1, r_1) \in S_2$, then $v_2 v_3^{r_1} v_4 w_1$ has only one return word, $v_2 v_3^{r_1} v_4 v_1^{s_1}$. If $(r_1, s_1) \notin S_1$ and $(s_1, r_1) \notin S_2$, then $v_2 v_3^{r_1} v_4 w_1$ has only two return words, $v_2 v_3^{r_1} v_4 v_1^{s_2}$ and $v_2 v_3^{r_1} v_4 v_1^{s_3}$. If $(r_1, s_1) \in S_1$ and $(s_1, r_1) \notin S_2$, then the return words of $v_2 v_3^{r_1} v_4 w_1$ are of the form $v_2 v_3^{r_1} v_4 v_1^{s_1} v_2 v_3^{r_i} v_4 v_1^{s_j}$ with $(r_i, s_j) \in S_1 \backslash \{(r_1, s_1)\}$, thus there are only three words. Similarly, $v_4 v_1^{s_1} v_2 w_2$ has only three return words if $(r_1, s_1) \notin S_1$ and $(s_1, r_1) \in S_2$.

This shows that $i_1 = i_4 = 1$, $i_2 = i_3 = 2$ is impossible, and the lemma is proved. □

7. Words associated with β-integers

In this section, we describe a new class of infinite words with the property R_m. The language of these words is not necessarily closed under reversal.

Consider the fixed point $u = \sigma^\infty(0)$ of the primitive substitution

$$\begin{array}{rrcl} \sigma: & 0 & \mapsto & 0^{t_1}1 \\ & 1 & \mapsto & 0^{t_2}2 \\ & & \vdots & \\ & m-2 & \mapsto & 0^{t_{m-1}}(m-1) \\ & m-1 & \mapsto & 0^{t_m} \end{array} \tag{4}$$

for some integers $m \geq 2$, $t_1, t_m \geq 1$ and $t_2, \ldots, t_{m-1} \geq 0$. The incidence matrix of σ is a companion matrix of the polynomial $x^m - t_1 x^{m-1} - \cdots - t_m$. Let $\beta > 1$ be the dominant root of this polynomial (the Perron-Frobenius eigenvalue of the matrix). If

$$t_j \cdots t_m \prec t_1 \cdots t_m \quad \text{for all } j \in \{2, \ldots, m\},$$

where $\preceq$ denotes the lexicographic ordering, then β is a simple Parry number (or simple β-number) and σ is a *β-substitution*, see e.g. Fabre [6]. It is easy to see that u codes in this case the sequence of distances between consecutive nonnegative β-integers

$$\mathbb{Z}_\beta^+ = \left\{ \sum_{j=0}^{J} x_j \beta^j \,\middle|\, J \geq 0, x_j \in \mathbb{Z}, x_j \geq 0, x_j \cdots x_0 \prec t_1 \cdots t_m \text{ for } 0 \leq j \leq J \right\}$$

and a letter k corresponds to the distance $t_{k+1}/\beta + \cdots + t_m/\beta^{m-k}$. (0 corresponds to distance 1.)

Remark. The most prominent example of a β-substitution is the Fibonacci substitution ($m = 2$, $t_1 = t_2 = 1$), where β is the golden mean.

It is not difficult to show that all prefixes of u are left special factors, with all m letters being left extensions (see e.g. Frougny, Masáková and Pelantová [10]). For every factor w, the tree of return words constructed by the left extensions (see Section 3.2) contains therefore a node with m children, the shortest prefix of u having w as suffix. The word u is uniformly recurrent since all fixed points of primitive substitutions have this property (Queffélec [13]). Therefore every factor w has at least m return words. If there exists a left special factor which is not a prefix of u, then this factor has more than m return words. By Proposition 3.1, we obtain the following proposition.

Proposition 7.1. *If $u = \sigma^\infty(0)$ for some substitution σ of the form (4), then it satisfies R_m if and only if $C(n) = (m-1)n + 1$ for all $n \geq 0$.*

Bernat, Masáková and Pelantová [2] characterized the fixed points of β-substitutions satisfying $\Delta C(n) = m - 1$ for all $n \geq 0$. The techniques of their proof can also be used to construct non-prefix left special factors if σ is a substitution of the form (4) which is not a β-substitution, and their conditions can be reformulated as in the following corollary.

Corollary 7.2. *If $u = \sigma^\infty(0)$ for some substitution σ of the form (4), then it has the property R_m if and only if*

- *$t_m = 1$ and*
- *$t_j \cdots t_{m-1} t_1 \cdots t_{j-1} \preceq t_1 \cdots t_{m-1}$ for all $j \in \{2, \dots, m-1\}$.*

Note that the language of u is closed under reversal if and only if $t_1 = t_2 = \cdots = t_{m-1}$. Then u is an Arnoux-Rauzy word of order m.

Acknowledgements

The authors acknowledge financial support by the Czech Science Foundation GA ČR 201/05/0169, by the grant LC06002 of the Ministry of Education, Youth and Sports of the Czech Republic, and by the project ANR-06-JCJC-0073 of the French Agence Nationale de la Recherche.

References

[1] I. M. Araújo, V. Bruyère, *Words derivated from Sturmian words*, Theor. Comput. Sci. **340** (2005), 204–219.

[2] J. Bernat, Z. Masáková, E. Pelantová, *On a class of infinite words with affine factor complexity*, to appear in Theor. Comput. Sci.

[3] J. Berstel, *Recent results on extensions of Sturmian words*, Int. J. Algebra Comput. **12** (2002), 371–385.

[4] J. Cassaigne, *Complexité et facteurs spéciaux*, Bull. Belg. Math. Soc. Simon Stevin **4** (1997), 67–88.

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

[6] S. Fabre, *Substitutions et β-systèmes de numération*, Theor. Comput. Sci. **137** (1995), 219–236

[7] S. Ferenczi, *Les transformations de Chacon: combinatoire, structure géométrique, lien avec les systèmes de complexité* $2n+1$, Bull. Soc. Math. Fr. **123** (1995), 271–292.

[8] S. Ferenczi, Ch. Holton, L. Q. Zamboni, *Structure of three interval exchange transformations. II. A combinatorial description of the trajectories.*, J. Anal. Math. **89** (2003), 239–276.

[9] S. Ferenczi, C. Mauduit, A. Nogueira, *Substitution dynamical systems: algebraic characterization of eigenvalues*, Ann. Sci. Éc. Norm. Supér. **29** (1996), 519–533.

[10] Ch. Frougny, Z. Masáková, E. Pelantová, *Complexity of infinite words associated with beta-expansions*, Theor. Inform. Appl. **38** (2004), 163–185; Corrigendum, Theor. Inform. Appl. **38** (2004), 269–271.

[11] J. Justin, L. Vuillon, *Return words in Sturmian and episturmian words*, Theor. Inform. Appl. **34** (2000), 343–356.

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

[13] M. Queffélec, *Substitution Dynamical Systems – Spectral Analysis*, Lecture Notes in Math. **1294**, Springer, Berlin, 1987.

[14] L. Vuillon, *A characterization of Sturmian words by return words*, Eur. J. Comb. **22** (2001), 263–275.

[15] L. Vuillon, *On the number of return words in infinite words with complexity* $2n+1$, LIAFA Research Report 2000/15.

Doppler Institute for Mathematical Physics and Applied Mathematics, and Department of Mathematics, FNSPE, Czech Technical University, Trojanova 13, 120 00 Praha 2, Czech Republic

E-mail address: l.balkova@centrum.cz, Pelantova@km1.fjfi.cvut.cz

LIAFA, CNRS, Université Paris Diderot – Paris 7, Case 7014, 75205 Paris Cedex 13, France

E-mail address: steiner@liafa.jussieu.fr