



**HAL**  
open science

## Return words in the Thue-Morse and other sequences

Lubomira Balkova, Edita Pelantová, Wolfgang Steiner

► **To cite this version:**

Lubomira Balkova, Edita Pelantová, Wolfgang Steiner. Return words in the Thue-Morse and other sequences. 2006. hal-00089863v2

**HAL Id: hal-00089863**

**<https://hal.science/hal-00089863v2>**

Preprint submitted on 4 Sep 2006 (v2), last revised 27 Sep 2007 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# RETURN WORDS IN THE THUE-MORSE AND OTHER SEQUENCES

LUBOMÍRA BALKOVÁ, EDITA PELANTOVÁ, AND WOLFGANG STEINER

ABSTRACT. We present three simple ideas which enable to determine easily the number of return words in some infinite words. Using these ideas, we give a new and very short proof of the fact that each factor of an Arnoux-Rauzy word of order  $m$  has exactly  $m$  return words. We describe the structure of return words for the Thue-Morse sequence and focus on infinite words which are associated to simple Parry numbers.

## 1. INTRODUCTION

Recently, the notion of return words has appeared in various branches of mathematics, for instance in Symbolic Dynamical Systems and Number Theory. This mathematical tool has been introduced by Durand [3] in order to obtain a nice characterization of primitive substitutive sequences. Roughly speaking, for a given factor  $w$  of the infinite word  $u$ , a return word of  $w$  is a segment between two successive occurrences of the factor  $w$ . Using return words, Vuillon [13] has found a new equivalent definition of Sturmian words. He has shown 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 generalization of Sturmian words to multiliteral alphabets is provided by Arnoux-Rauzy words. Justin and Vuillon [8] have proved that each factor of an Arnoux-Rauzy word of order  $m$  has exactly  $m$  return words.

Infinite words coding the  $m$ -interval exchange constitute different generalizations of Sturmian words to  $m$ -letter alphabets. For  $m = 3$ , it is known that every factor of these words has exactly 3 return words (Ferenczi, Holton and Zamboni [5]).

In this paper, we present three very simple ideas which enable to determine easily the number of return words in some infinite words.

- The first observation is that for the study of return words of an infinite uniformly recurrent word  $u$ , it suffices to limit the considerations to bispecial factors of  $u$ .
- We explain the role of factor graphs (also called Rauzy graphs or de Bruijn graphs) for the determination of the number of return words.
- In case of infinite words which are invariant under a substitution, we will make use of the relation between return words of a factor  $w$  and return words of its image under the substitution.

Using these ideas, we give a new and very short proof of the number of return words for Arnoux-Rauzy sequences. We describe the cardinality of the set of return words for the Thue-Morse sequence. In the main part of this paper, we focus on infinite words  $u_\beta$  which are associated to simple Parry numbers. The infinite word  $u_\beta$  codes the sequence of gaps between successive  $\beta$ -integers  $\mathbb{Z}_\beta$ , and it is the fixed point of the canonical substitution associated to  $\beta$  (see Section 2 for exact definitions). A simple Parry number is an algebraic integer  $\beta > 1$  having a finite Rényi expansion of unity  $d_\beta(1) = t_1 \dots t_m$ . In this case, the alphabet of the infinite word  $u_\beta$  contains  $m$  letters. In [6], it has been proved that  $u_\beta$  is Arnoux-Rauzy if and only if  $t_1 = \dots = t_{m-1}$  and  $t_m = 1$ . Consequently, in this case, the number of return words is equal to  $m$  for every factor of  $u_\beta$ . We show that the same statement about return words holds if  $t_1 > \max\{t_2, \dots, t_{m-1}\}$  and  $t_m = 1$ . Unlike the case  $m = 2$  (Sturmian words), Arnoux-Rauzy words with  $m \geq 3$  are not characterized by the number of return words. In the case  $t_m \geq 2$ , we focus on a simple situation where the coefficients  $t_1, \dots, t_{m-1}$  are either all the same or mutually different. Under these conditions, we show that for each factor of  $u_\beta$  the number of return words is either  $m$  or  $m + 1$  and both values occur.

## 2. PRELIMINARIES

**2.1. 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* (*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 *language*  $\mathcal{L}(u)$  is the set of all factors of the word  $u$ , and let  $\mathcal{L}_n(u)$  be the set of all factors of  $u$  of length  $n$ . 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. A concatenation of  $k$  letters  $a$  will be denoted by  $a^k$ , a concatenation of infinitely many letters  $a$  by  $a^\omega$ .

Let  $w$  be a factor of an infinite word  $u$  and let  $a, b \in \mathcal{A}$ . If  $wa$  is a factor of  $u$ , then we call  $a$  a *right extension* of  $w$ . Analogously, if  $bw$  is a factor of  $u$ , we call  $b$  a *left extension* of  $w$ . We will denote by  $\text{Rext}(w)$  the set of all right extensions of  $w$  and by  $\deg_+(w)$  the number of right extensions of  $w$ . Analogously, the set of left extensions of  $w$  will be denoted by  $\text{Lext}(w)$  and its cardinality by  $\deg_-(w)$ . A factor  $w$  is *right special* if  $\deg_+(w) > 1$ , *left special* if  $\deg_-(w) > 1$  and *bispecial* if  $w$  is right special and left special.

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

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

An infinite word  $u$  is *uniformly recurrent* if, for any  $n \in \mathbb{N}$ , there exists an  $R(n) \in \mathbb{N}$  such that any factor of  $u$  of length  $R(n)$  contains all factors of length  $n$ . It is not difficult to see that the set of return words of  $w$  is finite for any factor  $w$  if  $u$  is a uniformly recurrent word.

The variability of local configurations in  $u$  is expressed by the *factor complexity function* (or simply *complexity*)  $\mathcal{C}_u : \mathbb{N} \rightarrow \mathbb{N}$ ,  $\mathcal{C}_u(n) := \#\mathcal{L}_n(u)$ . It is apparent that the complexity difference can be calculated by

$$\Delta\mathcal{C}_u(n) := \mathcal{C}_u(n+1) - \mathcal{C}_u(n) = \sum_{w \in \mathcal{L}_n(u)} (\deg_-(w) - 1) = \sum_{w \in \mathcal{L}_n(u)} (\deg_+(w) - 1),$$

where it is sufficient to sum over left special and right special words respectively. It is well known that a word  $u$  is aperiodic if and only if  $\mathcal{C}_u(n) \geq n + 1$  for all  $n \in \mathbb{N}$ . Infinite aperiodic words with the minimal complexity  $\mathcal{C}_u(n) = n + 1$  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 [2].

A mapping  $\varphi$  on the free monoid  $\mathcal{A}^*$  is called a *morphism* if  $\varphi(vw) = \varphi(v)\varphi(w)$  for all  $v, w \in \mathcal{A}^*$ . Obviously, for defining the morphism, it suffices to give  $\varphi(a)$  for all  $a \in \mathcal{A}$ . The action of a morphism can be naturally extended on right-sided infinite words by

$$\varphi(u_0u_1u_2\dots) := \varphi(u_0)\varphi(u_1)\varphi(u_2)\dots$$

A non-erasing morphism  $\varphi$ , for which there exists a letter  $a \in \mathcal{A}$  such that  $\varphi(a) = aw$  for some non-empty word  $w \in \mathcal{A}^*$ , is called a *substitution*. An infinite word  $u$  such that  $\varphi(u) = u$  is called a *fixed point* of the substitution  $\varphi$ . Obviously, every substitution has at least one fixed point, namely  $\lim_{n \rightarrow \infty} \varphi^n(a)$ .

**2.2. Infinite words associated with  $\beta$ -integers.** Let  $\beta > 1$  be a real number. Then every positive real number  $x$  can be represented as

$$x = \sum_{j=-\infty}^k x_j \beta^j = x_k \dots x_1 x_0 \bullet x_{-1} x_{-2} \dots \quad \text{with } x_j \in \mathbb{N}.$$

If moreover  $x - \sum_{j=\ell}^k x_j \beta^j < \beta^\ell$  for all  $\ell \leq k$ , then the expression  $x_k \dots x_1 x_0 \bullet x_{-1} x_{-2} \dots$  is called (greedy)  *$\beta$ -expansion* of  $x$ , which we denote by  $\langle x \rangle_\beta$ . For  $x \in [0, 1)$ , this expansion can be found using the transformation  $T_\beta(x) := \{\beta x\}$ , where  $\{\cdot\}$  denotes the fractional part. Then  $\langle x \rangle_\beta = \bullet x_1 x_2 x_3 \dots$  with  $x_j = \lfloor \beta T_\beta^{j-1}(x) \rfloor$ , where  $\lfloor \cdot \rfloor$  denotes the (lower) integer part.

For  $x = 1$ , this algorithm does not provide the  $\beta$ -expansion of  $x$  (which is  $1.00\dots$ ), but we obtain an important sequence, the *Rényi expansion of 1* in base  $\beta$ , which is thus defined as

$$d_\beta(1) = t_1 t_2 t_3 \dots, \quad \text{where } t_j := \lfloor \beta T_\beta^{j-1}(1) \rfloor.$$

If  $t_j = 0$  for all  $j > m$  and  $t_m \neq 0$ , we write  $d_\beta(1) = t_1 \dots t_m$ . Such  $\beta$  is called a *simple Parry number*. From now on, let us limit our considerations to this case. A sequence  $t_1 \dots t_m$  is a Rényi expansion of 1 for some  $\beta > 1$  if and only if the sequence satisfies

$$t_j \dots t_m <_{\text{lex}} t_1 \dots t_m \quad \text{for all } j \in \{2, \dots, m\}$$

where  $v <_{\text{lex}} w$  means that  $v$  is lexicographically less than  $w$  (Parry [9]).

In this article, we study infinite words associated with non-negative  $\beta$ -integers  $\mathbb{Z}_\beta^+$ , which are the numbers with vanishing  $\beta$ -fractional part, formally

$$\mathbb{Z}_\beta^+ := \{x \geq 0 \mid \langle x \rangle_\beta = x_k x_{k+1} \dots x_0 \bullet 00 \dots \text{ for some } k \geq 0\}.$$

If  $\beta$  is an integer, then clearly  $\mathbb{Z}_\beta^+ = \mathbb{N}$  and the distance between neighboring elements of  $\mathbb{Z}_\beta$  for a fixed  $\beta$  is always 1. The situation changes if  $\beta \notin \mathbb{N}$ . In this case, the number of different distances between neighboring elements of  $\mathbb{Z}_\beta$  is at least 2. Thurston [12] showed that for a simple Parry number  $\beta$  with  $d_\beta(1) = t_1 \dots t_m$ , distances occurring between neighbors of  $\mathbb{Z}_\beta$  form the set  $\{\Delta_k \mid 0 \leq k < m\}$ , where  $\Delta_k := \bullet t_{k+1} \dots t_m$ .

If we assign the letter  $k$  to the gap  $\Delta_k$  for all  $k \in \{0, \dots, m-1\}$  and if we write down the order of distances in  $\mathbb{Z}_\beta^+$  on the real line, we naturally obtain an infinite word  $u_\beta$ . It can be shown that the word  $u_\beta$  is the unique fixed point of the canonical substitution  $\varphi$  (see e.g. [4]),

$$\begin{aligned} \varphi : \quad 0 &\mapsto 0^{t_1} 1 \\ &\quad 1 \mapsto 0^{t_2} 2 \\ &\quad \quad \quad \vdots \\ &\quad m-2 \mapsto 0^{t_{m-1}} (m-1) \\ &\quad m-1 \mapsto 0^{t_m}. \end{aligned}$$

Obviously, this substitution is primitive, i.e., there exists an integer exponent  $k$  such that for each pair of letters  $a, b \in \mathcal{A}$ , the letter  $a$  appears in the word  $\varphi^k(b)$ . Queffélec [11] showed that a fixed point of a primitive substitution is a uniformly recurrent infinite word. This implies that the infinite word  $u_\beta$  associated to a simple Parry number  $\beta$  is uniformly recurrent and, therefore, the set of return words  $M(w)$  of any factor  $w$  of  $u_\beta$  is finite.

### 3. FACTOR GRAPHS AND RETURN WORDS

In order to study return words  $M(w)$  of factors  $w$  of an infinite word  $u$ , it is possible to limit our considerations to bispecial factors. Namely, if a factor  $w$  is not right special, i.e., if it has a unique right extension  $a \in \mathcal{A}$ , then the sets of occurrences of  $w$  and  $wa$  coincide, and

$$M(w) = M(wa).$$

If a factor  $w$  has a unique left extension  $b \in \mathcal{A}$ , 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 uniformly recurrent infinite word, then the set  $M(w)$  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

$$M(bw) = bM(w)b^{-1} = \{bvb^{-1} \mid v \in M(w)\},$$

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

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 of  $M(w)$ , it suffices therefore to consider bispecial factors  $w$ .

Return words and the role of special factors can be well visualized by means of the *factor graphs* (or *de Bruijn graphs* or *Rauzy graphs*). The factor graph  $\Gamma_n$  is an oriented graph with set

of vertices  $\mathcal{L}_n(u)$  and set of edges  $\mathcal{L}_{n+1}(u)$ . An edge  $w_0w_1\dots w_n \in \mathcal{L}_{n+1}(u)$  goes from the vertex  $w_0\dots w_{n-1}$  to the vertex  $w_1\dots w_n$ . The factor  $w = w_0\dots w_{n-1}$  is left special if and only if at least two edges ends in the vertex  $w$  and  $w$  is right special if and only if at least two edges begins in  $w$ . Each return word of a factor  $w$  of the length  $n$  is visualized as an oriented walk in  $\Gamma_n$ , which begins and ends in the vertex  $w$ , and such that the vertex  $w$  is not entered in the course of this walk.

**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  $\deg_-(w) = m$ .*
- *There is a unique right special factor  $w \in \mathcal{L}_n(u)$ , and  $\deg_+(w) = m$ .*

*Then every factor has exactly  $m$  return words.*

*Proof.* As we have explained at the beginning of the section, it is enough to prove that every bispecial factor  $w$  has exactly  $m$  return words. Let  $n$  be the length of  $w$ . Suppose first that  $w$  is the unique right special factor in  $\mathcal{L}_n(u)$ . Hence  $w$  has outdegree  $m$  in  $\Gamma_n$  and all other vertices have outdegree 1. Therefore the choice of the edge starting the walk at  $w$  determines the walk uniquely. Consequently, there are exactly  $m$  ways to come back to  $w$ , i.e.,  $w$  has exactly  $m$  return words. If  $w$  is the unique left special factor in  $\mathcal{L}_n(u)$ , similar arguments can be used.  $\square$

Let us remind that Arnoux-Rauzy words of order  $m$  are defined as infinite words which have for every  $n \in \mathbb{N}$  exactly one right special factor  $w$  of length  $n$  with  $\deg_+(w) = m$  and exactly one left special factor  $w$  of length  $n$  with  $\deg_-(w) = m$ . Therefore we obtain the following corollary.

**Corollary 3.2.** *Any factor of an Arnoux-Rauzy word of order  $m$  has exactly  $m$  return words. In particular, any factor of a Sturmian word has exactly two return words.*

Let us apply this observation to words  $u_\beta$ .

**Proposition 3.3.** *Let  $d_\beta(1) = t_1\dots t_m$  be a Rényi expansion of unity. If*

- $t_m = 1$  and
- $t_1 = t_2 = \dots = t_{m-1}$  or  $t_1 > \max\{t_2, \dots, t_{m-1}\}$ ,

*then each factor  $w \in \mathcal{L}(u_\beta)$  has exactly  $m$  return words.*

*Proof.* In [6], it has been proved that each prefix of  $u_\beta$  is a left special factor with exactly  $m$  left extensions and that  $u_\beta$  does not have any other left special factors. This proves the statement.  $\square$

*Remark.* If  $d_\beta(1)$  satisfies  $t_1 = t_2 = \dots = t_{m-1}$ , then, for every  $n \in \mathbb{N}$ , there exists moreover exactly one right special factor of length  $n$ . Thus,  $u_\beta$  is Arnoux-Rauzy. Otherwise, we have some  $n \in \mathbb{N}$  such that there exists more than one right special factor of length  $n$ .

#### 4. RETURN WORDS IN THE THUE-MORSE SEQUENCE

The Thue-Morse sequence is a very well known, intensively studied binary sequence (see e.g. Allouche and Shallit [1]), which is a fixed point of the substitution

$$0 \mapsto \varphi(0) = 01 \quad \text{and} \quad 1 \mapsto \varphi(1) = 10.$$

This substitution has two fixed points. Since the substitution is primitive, both its fixed points have the same language and are uniformly recurrent. We will denote by  $u_{TM}$  one of the fixed points.

**Observation 4.1.** *The form of the substitution implies some simple facts about the language of  $u_{TM}$ .*

- (1)  $000, 111 \notin \mathcal{L}(u_{TM})$ ,
- (2)  $01010, 10101 \notin \mathcal{L}(u_{TM})$ ,
- (3)  $a_1\dots a_n \in \mathcal{L}(u_{TM}) \implies (1-a_1)\dots(1-a_n) \in \mathcal{L}(u_{TM})$ .

We have chosen the Thue-Morse sequence for studying return words, since it is easy to describe the bispecial factors of  $u_{TM}$ . Clearly, 0, 01 and 010 and their complements 1, 10 and 101 are bispecial factors. The structure of longer bispecial factors is also simple.

**Proposition 4.2.** *Let  $w$  be a bispecial factor of  $u_{TM}$  with length  $|w| \geq 4$ . Then there exists a bispecial factor  $v$  such that  $\varphi(v) = w$ .*

*Proof.* According to Claim 1 in Observation 4.1, any bispecial factor with at least 2 letters starts with 01 or 10 and ends with 01 or 10. According to Claim 2, the bispecial factors of length 4 are  $1001 = \varphi(10)$  or  $0110 = \varphi(01)$ . The preimage of  $w$  with  $|w| \geq 5$  is uniquely determined since any factor of length at least 5 contains two consecutive 0's or 1's (Claim 2 in the same observation), and  $0w$ ,  $1w$  as well as  $w0$  and  $w1$  belong to the language. Clearly, the preimage is bispecial as well.  $\square$

Now we show that for a broad class of substitutions there exists a straightforward relation between return words of a factor  $w$  and its image  $\varphi(w)$ .

**Definition 4.3.** *Let  $w$  be a factor of a fixed point  $u$  of a substitution  $\varphi$ . We say that a word  $v_0v_1 \dots v_{n-1}v_n \in \mathcal{L}_{n+1}(u)$  is a minimal preimage of  $w$  if*

- $w$  is a factor of  $\varphi(v_0v_1 \dots v_{n-1}v_n)$ ,
- $w$  is neither a factor of  $\varphi(v_1 \dots v_{n-1}v_n)$  nor of  $\varphi(v_0v_1 \dots v_{n-1})$ .

Clearly, any factor  $\varphi(w)$  has at least one minimal preimage, namely the factor  $w$ . In the Thue-Morse sequence, the factor  $\varphi(0) = 01$  has two minimal preimages, namely 0 and 11.

**Lemma 4.4.** *Let an infinite word  $u$  be a fixed point of a substitution  $\varphi$  and  $w$  be a factor of  $u$ . If the only minimal preimage of  $\varphi(w)$  is the factor  $w$ , then*

$$M(\varphi(w)) = \varphi(M(w)).$$

*Proof.* Let  $v$  be a return word of  $\varphi(w)$ , i.e.,  $v\varphi(w)$  is a factor of  $u = \varphi(u)$ ,  $\varphi(w)$  occurs in  $v\varphi(w)$  exactly two times and  $\varphi(w)$  is a prefix of  $v\varphi(w)$ . Since  $w$  is the only minimal preimage of  $\varphi(w)$ ,  $v = \varphi(v')$  for some factor  $v'$  and the factor  $w$  occurs in  $v'w$  twice. Therefore  $v'$  is a return word of  $w$ .  $\square$

**Theorem 4.5.** *Every factor of the Thue-Morse sequence has either 3 or 4 return words.*

*Proof.* For the description of the cardinality of  $M(w)$ , it is sufficient to consider bispecial factors  $w$ . Any bispecial factor  $w$  of  $u_{TM}$  with length at least 4 is of the form  $w = \varphi(v)$  and contains either 00 or 11. Therefore the bispecial factor  $\varphi(v)$  has the unique minimal preimage  $v$ . To prove the theorem, we have, according to Lemma 4.4, to determine the cardinality of the set of return words for bispecial factors shorter than 4, i.e., for the factors 0, 01 and 010 (the return words for complement of  $w$  are complements of return words of  $w$ ). Since

$$\begin{aligned} M(0) &= \{0, 01, 011\}, \\ M(01) &= \{01, 010, 011, 0110\}, \\ M(010) &= \{010, 0100110, 01011010, 010110011\}, \end{aligned}$$

the theorem is proved.  $\square$

## 5. RETURN WORDS AND $\beta$ -INTEGERS

As we know from Section 3, the return words of all factors of a uniformly recurrent word can be obtained from the return words of the bispecial factors. Hence, let us focus on bispecial factors.

**5.1. Bispecial factors of  $u_\beta$ .** In order to describe bispecial factors of  $u_\beta$ , where  $\beta$  is a simple Parry number, we will make use of the description of some special types of factors of  $u_\beta$  from [6].

**Definition 5.1.** *Let  $\iota_k := \min\{i \geq 1 \mid t_{k-i} \neq 0\}$  for  $k \in \{2, \dots, m\}$ . ( $\iota_k < k$  since  $t_1 \geq 1$ .)*

**Lemma 5.2.** [6, Lemma 4.5] *All factors of  $u_\beta$  of the form  $a0^rb$ , where  $a, b$  are non-zero letters and  $r \in \mathbb{N}$ , are the following ones:*

$$\begin{aligned} \iota_k 0^{t_k} k &\quad \text{for } k = 2, \dots, m-1, \\ k 0^{t_1} 1 &\quad \text{for } k = 1, \dots, m-1, \\ \iota_m 0^{t_1+t_m} 1. & \end{aligned}$$

**Corollary 5.3.**  $0^r$  is a bispecial factor if and only if  $r < t_1 + t_m$ .

- If  $r \leq t_1$ , then  $\text{Rext}(0^r) = \{0, 1\} \cup \{k \mid 2 \leq k < m, r \leq t_k\}$ .
- If  $t_1 < r < t_1 + t_m$ , then  $\text{Rext}(0^r) = \{0, 1\}$ .

**Lemma 5.4.** If  $v$  is a bispecial factor which contains a non-zero letter, then there exists a unique bispecial factor  $w$  and a unique  $s \in \{t_1, \dots, t_{m-1}\}$  such that  $v = \varphi(w)0^s$ .

*Proof.* Let  $v'$  be such that  $v = v'0^s$  and  $v'$  ends with a non-zero letter. From Lemma 5.2, it is clear that  $s \in \{t_1, \dots, t_{m-1}\}$  and  $v'$  has the prefix  $0^{t_1}1$ . Thus,  $v$  is either of the form

- $v = 0^{t_1}10^s = \varphi(0)0^s$  or
- $v = 0^{t_1}1v''0^s$ .

In the latter case, there exists  $w \in \mathcal{L}(u_\beta)$  such that  $\varphi(w) = 0^{t_1}1v''$ , and  $w$  is obviously bispecial.  $\square$

**Observation 5.5.** If  $(m-1)a \in \mathcal{L}(u_\beta)$  for a letter  $a$ , then  $a = 0$ .

*Proof.* This follows from Lemma 5.2 since  $t_1 \geq 1$  and  $t_k < k$  for every  $k$ .  $\square$

**Lemma 5.6.** If  $w$  is a bispecial factor, then there exists  $s \in \{t_1, \dots, t_{m-1}\}$  such that  $\varphi(w)0^s$  is bispecial. Moreover, if  $\deg_+(w) = 2$ , then this  $s$  is unique and  $\deg_+(\varphi(w)0^s) = 2$ .

*Proof.* If  $w$  is bispecial, then it has at least two different right extensions, say  $a < b$ , and:

- If  $b < m-1$ , then  $\varphi(w)0^{t_{a+1}}(a+1)$  and  $\varphi(w)0^{t_{b+1}}(b+1)$  are factors of  $u_\beta$ . Hence  $\varphi(w)0^s$  with  $s = \min\{t_{a+1}, t_{b+1}\}$  is bispecial.
- If  $b = m-1$ , then Observation 5.5 implies that  $\varphi(w(m-1)0) = \varphi(w)0^{t_m}0^{t_1}1 \in \mathcal{L}(u_\beta)$ , and  $\varphi(w)0^{t_{a+1}}$  is bispecial since  $0, a+1 \in \text{Rext}(\varphi(w)0^{t_{a+1}})$ .

If the bispecial factor  $w$  has only two right extensions, then this choice of  $s$  is clearly unique.  $\square$

The two next corollaries are direct consequences of Lemma 5.6 and its proof.

**Corollary 5.7.** Let  $w$  be a bispecial factor satisfying that both  $\varphi(w)0^p$  and  $\varphi(w)0^q$  are bispecial factors of  $u_\beta$  for some  $p \neq q$ , then  $\deg_+(w) \geq 3$ .

**Corollary 5.8.** Let  $w$  be a bispecial factor which contains a non-zero letter, and let  $a \neq b$  be non-zero letters satisfying  $a, b \in \text{Rext}(w)$ . Then  $t_a = t_b$ .

**5.2. Return words of  $u_\beta$ .** Once, bispecial factors of  $u_\beta$  described, we can investigate their return words. Lemmas 5.4 and 5.6 show that occurrences of  $w$  in  $u_\beta$  can be described by means of their “preimages” or “images” and this fact is useful for description of their return words.

**Proposition 5.9.** Let  $w$  be a bispecial factor of  $u_\beta$  such that there exists a unique  $s \in \{t_1, \dots, t_{m-1}\}$  with the property that  $\varphi(w)0^s$  is a bispecial factor of  $u_\beta$ . Then

$$M(\varphi(w)0^s) = \varphi(M(w)).$$

*Proof.* It is an obvious consequence of the form of the substitution that if the letter  $m-1$  is not a suffix of  $w$ , then  $w$  is the only minimal preimage of  $\varphi(w)$ . According to the Observation 5.5, no bispecial factor ends with  $m-1$ . Therefore Lemma 4.4 implies  $M(\varphi(w)) = \varphi(M(w))$ . Let us remind that if a factor  $v$  is not right special, then  $M(v) = M(va)$ , where  $a$  is the unique right extension of  $v$ . According to our assumption for  $k < s$ , the factor  $\varphi(w)0^k$  is not right special and  $\varphi(w)0^{k+1}$  is its unique right extension. Hence we may apply the previous rule  $s$  times to obtain the statement.  $\square$

The next lemma will turn out to be useful for determining return words of prefixes of  $0^\omega$ .

**Lemma 5.10.** Let  $d_\beta(1) = t_1 \dots t_m$ , where at most one coefficient is 0. Let  $j \in \{0, \dots, m-1\}$  and  $w$  be a return word of  $j$ . Then we have either

- $wj = j\varphi^j(0)$  or
- $wj = j\varphi^j(k0)$  for some  $k \geq 1$  with  $t_k \neq 0$  or
- $wj = j\varphi^j(1k0)$  for some  $k \geq 1$  with  $t_k = 0$ .

*Proof.* For  $j = 0$ , we have to look for factors of the form  $0v0$ , where  $v$  does not contain 0. By Lemma 5.2, the only factors of this form are  $00$ ,  $0k0$  (for  $t_k \neq 0$ ) and  $01k0$  (for  $t_k = 0$ ), as  $t_k = 1$  by the assumption that at most one  $t_k$  is 0. Therefore, the statement is true for  $j = 0$  and we can proceed by induction on  $j$ . For a return word  $w$  of  $j+1$ , it follows from the form of the substitution that  $w(j+1) = (j+1)\varphi(w'j)$ , and  $jw'$  is a return word of  $j$ . The induction assumption implies that  $w(j+1) = (j+1)\varphi(w'j) = (j+1)\varphi^{j+1}(x0)$  with  $x$  being the empty word,  $x = k$  and  $x = 1k$  respectively.  $\square$

The cardinality of the set of return words of a factor of the infinite word  $u_\beta$  is described partially by Proposition 3.3 in case  $t_m = 1$ . The two next theorems deal also with the case  $t_m \geq 2$ .

**Theorem 5.11.** *Let  $u_\beta$  be the infinite word associated with the (confluent) simple Parry number  $\beta$  satisfying  $d_\beta(1) = t_1 \dots t_{m-1}t_m = t \dots t t_m$ .*

- *If  $t_m = 1$ , then every factor  $w \in \mathcal{L}(u_\beta)$  has exactly  $m$  return words.*
- *If  $t_m \geq 2$ , then every factor  $w \in \mathcal{L}(u_\beta)$  has either  $m$  or  $m+1$  return words (and both values occur).*

*Proof.* The condition  $t_1 = \dots = t_{m-1} = t$  and Lemma 5.6 imply that for each bispecial factor  $w$  there exists a unique  $s$  such that  $\varphi(w)0^s$  is a bispecial factor ( $s = t$ ). Using Proposition 5.9, we have  $\#M(\varphi(w)0^s) = \#M(w)$ . Lemma 5.4 states that to determine the cardinality of the set of all return words, it suffices to study the cardinality of the set  $M(0^r)$ , where  $0^r$  is a bispecial factor of  $u_\beta$ , i.e.,  $1 \leq r < t + t_m$  by Corollary 5.3. In the next considerations, we use mainly Lemma 5.2. For  $1 \leq r \leq t$ , the block  $0^r$  is followed either by

- 0, then the return word of  $0^r$  is 0, or by
- $k0^t$  with  $k \in \{1, 2, \dots, m-1\}$ , hence we obtain the return word  $0^r k$ .

Consequently, one gets

$$M(0^r) = \{0, 0^r 1, 0^r 2, \dots, 0^r(m-1)\} \quad \text{and} \quad \#M(0^r) = m.$$

For  $t < r < t + t_m$  (this case appears only for  $t_m \geq 2$ ), the block  $0^r$  is followed either by

- 0, then the return word of  $0^r$  is 0, or by
- 1, then  $0^r 1$  is suffix of  $0^{t_m+t} 1 = \varphi((m-1)0)$ . If we denote the return word of  $0^r$  by  $w$ , then, obviously,  $w$  is followed by the very next occurrence of  $0^{t_m+t} 1$ . Thus, the factor  $0^{t_m+t-r} w 0^{t_m}$  can be written as  $\varphi((m-1)v(m-1))$ , where  $v$  does not contain  $m-1$ . Therefore  $(m-1)v$  is a return word of  $m-1$ , and by Lemma 5.10, either  $v(m-1) = \varphi^{m-1}(0)$  or  $v(m-1) = \varphi^{m-1}(k0)$  for some  $k \geq 1$ .

Consequently, we have

$$M(0^r) = \{0, 0^{r-t} \varphi^m(0) 0^{-t_m}\} \cup \{0^{r-t} \varphi^m(k0) 0^{-t_m} \mid 1 \leq k < m\}$$

and  $\#M(0^r) = m+1$ .  $\square$

**Theorem 5.12.** *Let  $u_\beta$  be the infinite word associated with the simple Parry number  $\beta$  satisfying  $d_\beta(1) = t_1 \dots t_{m-1}t_m$  such that  $t_i \neq t_j$  for all  $i, j$ ,  $1 \leq i < j < m$ .*

- *If  $t_m = 1$ , then every factor  $w \in \mathcal{L}(u_\beta)$  has exactly  $m$  return words.*
- *If  $t_m \geq 2$ , then every factor  $w \in \mathcal{L}(u_\beta)$  has either  $m$  or  $m+1$  return words (and both values occur).*

*Proof.* By Lemma 5.4, every bispecial factor is either of the form  $0^r$  or  $\varphi(w)0^s$ . If  $s$  is unique, then the problem is reduced to the determination of return words of  $w$ , by Proposition 5.9. Suppose that there exist  $p > q$  such that  $\varphi(w)0^p$  and  $\varphi(w)0^q$  are bispecial. Then  $w$  has at least two non-zero right extensions, by Corollary 5.7. Thus, by Corollary 5.8 and the assumption of the theorem,  $w$  does not contain a non-zero letter. Consequently, it suffices to describe the number of return words for bispecial factors of the form  $0^r$ ,  $1 \leq r < t_1 + t_m$ , and for  $\varphi(0^r)0^s$ . In the latter case, it suffices to consider  $r < t_1$ , since  $0^r$ ,  $t_1 \leq r < t_1 + t_m$ , has just the right extensions 0, 1 and, according to Lemma 5.6, there exists a unique  $s$  such that  $\varphi(0^r)0^s$  is bispecial.



**Return words of  $0^r$ ,  $1 \leq r < t_1 + t_m$ .**

If  $1 \leq r \leq t_1$ , then  $0^{r+1} \in \mathcal{L}(u_\beta)$ , thus  $0 \in M(0^r)$ . As an obvious consequence of the form of substitution, it follows that  $k0^r$ ,  $k \geq 1$ , is always and solely suffix of  $\varphi^k(0)0^r$ . Since  $r \leq t_1$ , the factor  $\varphi^k(0)$  contains  $0^r$  for each  $k \geq 1$ . Therefore, there is exactly one return word of  $0^r$  ending by  $k$  for every  $k \geq 1$ .

If  $t_1 < r < t_1 + t_m$ , then it suffices again to apply Lemma 5.10 and an analogous method as in the second part of the proof of Theorem 5.11, and one obtains

$$M(0^r) = \{0\} \cup \{0^{r-t_1}\varphi^m(0)0^{-t_m}\} \cup \{0^{r-t_1}\varphi^m(k0)0^{-t_m} \mid 1 \leq k < m, t_k \neq 0\} \\ \cup \{0^{r-t_1}\varphi^m(1k0)0^{-t_m} \mid 1 \leq k < m, t_k = 0\}.$$

Consequently,  $\#M(0^r) = m$  for  $1 \leq r \leq t_1$  and  $\#M(0^r) = m + 1$  for  $t_1 < r < t_1 + t_m$ .

**Return words of  $\varphi(0^r)0^s$ ,  $1 \leq r < t_1$ ,  $s \in \{t_1, \dots, t_{m-1}\}$ .**

We show that, for every letter  $k$ , there exists exactly one return word ending with  $k$  and thus  $\#M(\varphi(0^r)0^s) = m$ . Let us distinguish two cases:

- For  $k \geq 1$ , it is an immediate consequence of the form of the substitution that each factor of  $u_\beta$  having suffix  $k\varphi(0^r)0^s$  has as well the factor  $\varphi^k(0)\varphi(0^r)0^s$  as suffix. Moreover, the factor  $\varphi^k(0)\varphi(0^r)0^s$  has  $\varphi(0^r)0^s$  as prefix. The only return word ending with  $k$  is therefore the shortest suffix  $v$  of  $\varphi^k(0)$  such that  $v\varphi(0^r)0^s$  has  $\varphi(0^r)0^s$  as prefix.
- For  $k = 0$ , the factor  $0\varphi(0^r)0^s$  has the suffix  $\varphi^m(0)\varphi(0^r)0^s$ . The only return word ending with  $0$  is therefore the shortest suffix of  $\varphi^m(0)$  starting with  $\varphi(0^r)0^s$ .

□

## ACKNOWLEDGEMENTS

The authors acknowledge financial support by the Czech Science Foundation GA ČR 201/05/0169 and by the grant LC06002 of the Ministry of Education, Youth and Sports of the Czech Republic.

## REFERENCES

- [1] J.-P. Allouche, J. Shallit, *The ubiquitous Prouhet-Thue-Morse sequence*, Sequences and their applications (Singapore, 1998), 1–16, Springer Ser. Discrete Math. Theor. Comput. Sci., Springer, London, (1999)
- [2] J. Berstel, *Recent results on extension of Sturmian words*, Int. J. Algebra Comput. **12** (2002), 371–385
- [3] F. Durand, *A characterization of substitutive sequences using return words*, Discrete Math. **179** (1998), 89–101
- [4] S. Fabre, *Substitutions et  $\beta$ -systèmes de numération*, Theoret. Comput. Sci. **137** (1995), 219–236
- [5] S. Ferenczi, Ch. Holton, L. Q. Zamoni, *Structure of three interval exchange transformations. II. A combinatorial description of the trajectories.*, J. Anal. Math. **89** (2003), 239–276
- [6] Ch. Frougny, Z. Masáková, E. Pelantová, *Complexity of infinite words associated with beta-expansions*, RAIRO Theor. Inform. Appl. **38** (2004), 163–185; Corrigendum, RAIRO Theor. Inform. Appl. **38** (2004), 269–271
- [7] J. Justin, G. Pirillo, *On a combinatorial property of Sturmian words*, Theor. Comput. Sci. **154(2)** (1996), 387–394
- [8] J. Justin, L. Vuillon, *Return words in Sturmian and episturmian words*, Theor. Inform. and App. **34** (2000), 343–356
- [9] W. Parry, *On the beta-expansions of real numbers*, Acta Math. Acad. Sci. Hungar. **11** (1960), 401–416
- [10] A. Rényi, *Representations for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hungar. **8** (1957), 477–493
- [11] M. Queffélec, *Substitution Dynamical Systems – Spectral Analysis*, in: Lecture Notes in Math. **1294**, Springer, Berlin, 1987
- [12] W. P. Thurston, *Groups, tilings, and finite state automata*, AMS Colloquium lectures, 1989
- [13] L. Vuillon, *A characterization of Sturmian words by return words*, Eur. J. Comb. **22(2)** (2001), 263–275

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:* [1.balkova@centrum.cz](mailto:1.balkova@centrum.cz), [Pelantova@km1.fjfi.cvut.cz](mailto:Pelantova@km1.fjfi.cvut.cz)

CNRS – LIAFA, UNIVERSITÉ DENIS DIDEROT, 2 PLACE JUSSIEU, CASE 7014, 75251 PARIS CEDEX 05, FRANCE

*E-mail address:* [steiner@liafa.jussieu.fr](mailto:steiner@liafa.jussieu.fr)