ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta jaderná a fyzikálně inženýrská

HABILITAČNÍ PRÁCE

Combinatorics on Words and Applications

Ľubomíra Dvořáková (rozená Balková)

2014

Contents

Abstract

This habilitation thesis combines two fields of mathematics – the main one is Combinatorics on Words and combinatorial results are applied in Random Number Generation. It is written in the form of a collection of papers – all of them besides the last one have been already published in scientific journals, the last one has been submitted. The collection is accompanied with a text presenting in a nutshell and at the same time in a way comprehensive also for non-specialists the content of the papers. The following six papers have been chosen

1. *Sturmian Jungle (or Garden?) on Multiliteral Alphabets* [8];
2. *Infinite Words with Finite Defect* [9];
3. *On the Brlek–Reutenauer Conjecture* [10];
4. *Proof of the Brlek–Reutenauer Conjecture* [11];
5. *Infinite Words with Well Distributed Occurrences* [5];
6. *Pseudorandom Number Generators Based on Infinite Words* [6]

in order to draw a picture that is self-consistent and illustrates very well the steps done by the author in the last five years in her research. The first paper is purely combinatorial, it is devoted to the study of a famous class of infinite words – Sturmian words – and in particular to their generalization to multiliteral alphabets. Combinatorics on Words is treated here from many perspectives and the topic of richness in palindromes is opened already in this study. The following three papers study palindromes in infinite words, in particular, they deal with richness and defect of infinite words. Recently, this subject has been very popular in Combinatorics on Words and several research groups have contributed to its deeper exploration. One of our main contribution is the proof of a conjecture stated by Brlek and Reutenauer. In the papers 2., 3., and 4. our thorny, but successful path to its complete proof is described. The last two papers contain applications of Combinatorics on Words in Random Number Generation. We study there a new property of infinite words – baptized well distributed occurrences – that was born in connection with random number generation. It is a sufficient condition for absence of the lattice structure when pseudorandom number generators are mixed according to an infinite word. We succeeded to prove aperiodicity and absence of the lattice structure for arising generators, however, their statistical testing shows – to our pleasure – that such generators are almost as fast as the original generators and their other statistical properties resemble much more to a random sequence in comparison to the original generators. This means that such generators are suitable for use in practice.

Results of all above six papers were presented by the habilitation candidate in the form of lectures at international conferences: Combinatorial generalizations of Sturmian words from the first paper were introduced at NORCOM 2010 (the 10th Nordic Combinatorial Conference) in

Reykjavík, Island, in May 2010, and at Joint Mathematical Conference CSASC 2010, session Discrete Dynamical Systems, in Prague, in January 2010. The results on richness, defect, and the Brlek–Reutenauer conjecture (items 2., 3., 4. of the above list) were presented at Česko-slovenská MELA (Meeting on Languages) in Telč as an invited lecture in September 2012, at RuFiDim (Russian Finnish Symposium on Discrete Mathematics) in Turku, Finland, in September 2012, and at Workshop on Challenges in Combinatorics on Words in Toronto, Canada, in April 2013. The talks on pseudorandom number generators based on infinite words with well distributed occurrences were given at WORDS 2013 in Turku, Finland, in September 2013, and at Česko-slovenská MELA in Telč in September 2013. Moreover, a lecture on application of Combinatorics on Words in Random Number Generation was held in April 2014 within a research stay at the University of Oulu, Finland. Last but not least, in 2013 the candidate was laureate of the award L'Oréal UNESCO For Women in Science for the project *Application of Combinatorics on Words in Cryptology* that was closely related to the results summarized in the last two papers.

Chapter 1

Combinatorics on Words

Let us start our look inside the subject of Combinatorics on Words in a gentle manner. After this gentle introduction, we will have at hand the necessary background to start presentation of our own results from Section 1.2.2 on. For Czech readers we provided such an easy introduction in [4].

There are two moments considered as the birth of Combinatorics on Words. According to the one we choose this discipline is almost 110 or almost 75 years old.

1.1 Axel Thue

The Norwegian number theorist and logician Thue known in particular for his results in Diophantine approximations published in 1906 in an obscure Norwegian journal [36] answers to the following questions:
Question 1: "Does there exist an infinite binary word that does not contain cubes?" Let us illustrate on a concrete example the notions from his question. An infinite binary word is an infinite sequence consisting of only two symbols, say a, b. Take for simplicity a periodic sequence:

$$abbabbabbabbabbabbabbabbabb\ldots. \tag{1.1.1}$$

It is thus an infinite repetition of the chain abb, which we usually denote $(abb)^{\omega}$. Such a word contains a cube because for instance the chain abb occurs three times in a row in this word:

$$\underline{abbabbabb}abbabbabbabbabbabb\ldots.$$

Question 2: "Does there exist an infinite ternary word (consisting of three symbols) that does not contain squares?" A square is a double repetition of a chain.

Both of these questions were answered by Thue confirmatively. Moreover he explained that he had no specific application in his mind, but he studied these questions since he found them interesting.

1.1.1 The Thue–Morse word

The Thue–Morse word provided by Thue as an example of an infinite word without cubes consisting of two symbols 0 and 1 – that answers thus positively **Question 1** – can be constructed by several equivalent ways. Let us provide three of them.

1. The Thue–Morse word $\mathbf{u}_{\rm TM} = u_0u_1u_2\ldots$ is an infinite word given by a recurrence relation:

$$u_0 = 0, \quad u_{2n} = u_n \quad \text{and} \quad u_{2n+1} = 1 - u_n \quad \text{for } n \geq 0.$$

2. Let us denote $s_2(n)$ the sum of digits in the binary expansion of $n \in \mathbb{N}$.[1] Then the Thue–Morse word $\mathbf{u}_{\mathrm{TM}} = u_0u_1u_2\ldots$ satisfies $u_n = s_2(n) \mod 2$ for all $n \in \mathbb{N}$.

3. Let us define a morphism $\varphi_{\mathrm{TM}}\colon \{0,1\}^* \to \{0,1\}^*$ by

$$\varphi_{\mathrm{TM}}(0) = 01 \quad \text{and} \quad \varphi_{\mathrm{TM}}(1) = 10,$$

   where $\{0,1\}^*$ denotes the set of all finite sequences consisting of symbols 0 and 1. Then the Thue–Morse word $\mathbf{u}_{\mathrm{TM}}$ is the so-called fixed point of φ_{TM} starting in 0. In order to get the fixed point, we apply the morphism repeatedly

$$\begin{array}{rcl} \varphi^0_{\mathrm{TM}}(0) & = & 0 \\ \varphi^1_{\mathrm{TM}}(0) & = & 01 \\ \varphi^2_{\mathrm{TM}}(0) & = & 0110 \\ \varphi^3_{\mathrm{TM}}(0) & = & 0110100110010110 \\ \varphi^4_{\mathrm{TM}}(0) & = & 01101001100101101001011001101001. \end{array} \tag{1.1.2}$$

   Since every iteration is the prefix of the next iteration and their lengths are strictly growing, it is possible to find an infinite word such that for all n the word $\varphi^n_{\mathrm{TM}}(0)$ is its prefix. This infinite word is the unique fixed point $\mathbf{u}_{\mathrm{TM}}$. We use the symbolic notation $\mathbf{u}_{\mathrm{TM}} = \lim_{n\to+\infty} \varphi^n_{\mathrm{TM}}(0)$.

Thue further constructed an infinite word $\mathbf{v}$ over $\{0,1,2\}$ that does not contain squares and responded hence positively also to **Question 2**. For $n \geq 1$ he denoted by v_n the number of ones between the n-th and $(n+1)$-th occurrence of zero in the Thue–Morse word. The desired word was then $\mathbf{v} = v_1v_2v_3\ldots$, i.e.,

$$\underbrace{\mathbf{u}_{\mathrm{TM}}}_{\mathbf{v}} = 0\underbrace{11}_{2}0\underbrace{1}_{1}0\underbrace{}_{0}0\underbrace{11}_{2}0\underbrace{}_{0}0\underbrace{1}_{1}0\underbrace{11}_{2}0\ldots$$

Since the result of Thue did not become known, Marston Morse rediscovered the Thue–Morse word in 1921 when he was studying differential geometry [33].

1.1.2 Dictionary of Combinatorics on Words

In order to describe the second moment considered as the birth of Combinatorics on Words and in order to present our results in the sequel, we have to introduce basic notions from Combinatorics on Words. By $\mathcal{A}$ we denote a finite set of symbols, usually called **letters**. The set $\mathcal{A}$ is therefore called an **alphabet**. A finite string $w = w_0w_1\ldots w_{n-1}$ of letters of $\mathcal{A}$ is said to be a **finite word**, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ε as the neutral element form a monoid $\mathcal{A}^*$. The map

$$w = w_0w_1\ldots w_{n-1} \quad \mapsto \quad \overline{w} = w_{n-1}w_{n-2}\ldots w_0$$

is a bijection on $\mathcal{A}^*$ and the word $\overline{w}$ is called the **reversal** or the **mirror image** of w. A word w which coincides with its mirror image is a **palindrome**.

Under an **infinite word** we understand an infinite string $\mathbf{u} = u_0u_1u_2\ldots$ of letters from $\mathcal{A}$. A finite word w is a **factor** of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \varepsilon$, then w is said to be a **prefix** of v. If $s = \varepsilon$, then w is a **suffix** of v.

[1] $\mathbb{N}$ stands everywhere for the set $\{0,1,2,\ldots\}$.

The **language** $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. Factors of $\mathbf{u}$ of length n form the set denoted by $\mathcal{L}_n(\mathbf{u})$. Clearly, $\mathcal{L}(\mathbf{u}) = \cup_{n\in\mathbb{N}}\mathcal{L}_n(\mathbf{u})$. We say that the language $\mathcal{L}(\mathbf{u})$ is **closed under reversal** if $\mathcal{L}(\mathbf{u})$ contains with every factor w also its reversal $\overline{w}$.

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_i u_{i+1} u_{i+2}\dots$. Such an index i is called an **occurrence** of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ occurs infinitely many times in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be **recurrent**. If the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent (a proof can be found in [24]). The infinite word $\mathbf{u}$ is said to be **uniformly recurrent** if $\mathbf{u}$ is recurrent and for any factor w of $\mathbf{u}$ the distances between successive occurrences of w form a bounded sequence.

The **(factor) complexity** of an infinite word $\mathbf{u}$ is the map $\mathcal{C}\colon \mathbb{N} \mapsto \mathbb{N}$ defined by $\mathcal{C}(n) = \#\mathcal{L}_n(\mathbf{u})$. To determine the increment of complexity of an infinite word $\mathbf{u}$, one has to count the possible **extensions** of factors of length n. A **left extension** of $w \in \mathcal{L}(\mathbf{u})$ is any letter $a \in \mathcal{A}$ such that $aw \in \mathcal{L}(\mathbf{u})$. The set of all left extensions of a factor w will be denoted by $\mathrm{Lext}(w)$. We will mostly deal with recurrent infinite words $\mathbf{u}$. In this case, any factor of $\mathbf{u}$ has at least one left extension. A factor w is called **left special** (or LS for short) if w has at least two left extensions. Clearly, any prefix of a LS factor is LS as well. Similarly, one can define a **right extension**, a **right special** (or RS) factor, $\mathrm{Rext}(w)$. We say that a factor w of $\mathbf{u}$ is a **bispecial** (or BS) factor if it is both RS and LS. Using the introduced terminology, the increment or the **first difference of complexity** $\Delta\mathcal{C}(n) = \mathcal{C}(n+1) - \mathcal{C}(n)$ is given by

$$\Delta\mathcal{C}(n) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Rext}(w) - 1\big) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Lext}(w) - 1\big)\,. \tag{1.1.3}$$

A non-zero contribution to $\Delta\mathcal{C}(n)$ in the left-hand sum is given only by RS factors $w \in \mathcal{L}_n(\mathbf{u})$, and for recurrent words, a non-zero contribution to $\Delta\mathcal{C}(n)$ in the right-hand sum is provided only by LS factors $w \in \mathcal{L}_n(\mathbf{u})$. If we denote $\mathrm{Bext}(w) = \{awb \in \mathcal{L}(\mathbf{u}) \mid a, b \in \mathcal{A}\}$, then the **second difference of complexity** $\Delta^2\mathcal{C}(n) = \Delta\mathcal{C}(n+1) - \Delta\mathcal{C}(n) = \mathcal{C}(n+2) - 2\mathcal{C}(n+1) + \mathcal{C}(n)$ is given by

$$\Delta^2\mathcal{C}(n) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1\big)\,. \tag{1.1.4}$$

Denote by $\mathrm{b}(w)$ the quantity

$$\mathrm{b}(w) = \#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1.$$

The number $\mathrm{b}(w)$ is called the **bilateral order** of the factor w and was introduced in [18]. It is readily seen that if w is not a BS factor, then $\mathrm{b}(w) = 0$. Bispecial factors are distinguished according to their bilateral order in the following way:

- if $\mathrm{b}(w) > 0$, then w is a **strong** BS factor;
- if $\mathrm{b}(w) < 0$, then w is a **weak** BS factor;
- if $\mathrm{b}(w) = 0$, then w is an **ordinary** BS factor.

We will moreover need the notion of **palindromic extension**. The set of palindromic extensions of a palindrome $w \in \mathcal{L}(\mathbf{u})$ is defined by $\mathrm{Pext}(w) = \{awa \in \mathcal{L}(\mathbf{u}) \mid a \in \mathcal{A}\}$. The number of palindromes of a fixed length occurring in an infinite word is measured by the so called **palindromic complexity** $\mathcal{P}$, a map which assigns to any non-negative integer n the number

$$\mathcal{P}(n) = \#\{w \in \mathcal{L}_n(\mathbf{u}) \mid w \text{ is a palindrome}\}\,.$$

Let j, k, $j < k$, be two successive occurrences of a factor w in $\mathbf{u}$. Then $u_j u_{j+1} \dots u_{k-1}$ is called a **return word** of w. If v is a return word of w, then the word vw is called a **complete return word** of w. It is obvious that an infinite recurrent word is uniformly recurrent if and only if the set of return words of any of its factors is finite.

We will often work with **morphisms**, i.e., mappings $\varphi\colon \mathcal{A}^* \to \mathcal{A}^*$ satisfying for any $v, w \in \mathcal{A}^*$ that $\varphi(vw) = \varphi(v)\varphi(w)$. A morphism is thus uniquely given if we define images of letters $\varphi(a)$ for all $a \in \mathcal{A}$. A morphism can be naturally extended to infinite words:

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

If an infinite word $\mathbf{u}$ satisfies $\varphi(\mathbf{u}) = \mathbf{u}$, we call $\mathbf{u}$ a **fixed point** of the morphism φ.

Example 1. Let us illustrate the introduced notions on the infinite word $\mathbf{u} = (abb)^\omega$. Its alphabet is $\mathcal{A} = \{a, b\}$. The infinite word $\mathbf{u}$ is uniformly recurrent and its language is closed under reversal. The word $babb$ is a factor of length 4 of $\mathbf{u}$. The word $abbabba$ is a prefix of length 7 of $\mathbf{u}$. The only LS factors are ε and b. Consequently $\Delta\mathcal{C}(n) = 0$ for all $n \geq 2$. It is readily seen that the set of all factors of length 2 of $\mathbf{u}$ equals $\mathcal{L}_2(\mathbf{u}) = \{ab, bb, ba\}$. Therefore $\mathcal{C}(n) = 3$ for all $n \geq 2$. The only BS factors are ε and b with $\mathrm{b}(\varepsilon) = 0$ and $\mathrm{b}(b) = -1$. It is not difficult to see that there is one palindrome of any even length, i.e., $\mathcal{P}(2n) = 1$ for every $n \in \mathbb{N}$, and there are two palindromes of length one – the letters a, b – and one palindrome of every odd length larger than one, i.e., $\mathcal{P}(2n+3) = 1$ for every $n \in \mathbb{N}$. Since $u_0 u_1 u_2 u_3 u_4 u_5 u_6 u_7 u_8 u_9 \dots = ab\underline{\underline{babba}bba}\dots$, the index $i = 2$ is an occurrence of the factor $babba$ and bab is a return word of $babba$ and $babbabba$ is a complete return word of $babba$. If we define a morphism φ on $\{a, b\}^*$ by $\varphi(a) = abb$ and $\varphi(b) = abb$, then $\mathbf{u}$ is evidently a fixed point of φ.

1.2 Gustav A. Hedlund and Marston Morse

Even more often one considers for the birth of Combinatorics on Words the famous paper [34] by Hedlund and already mentioned Morse from 1940. When studying differential equations of Sturm–Liouville type, they discovered a certain class of infinite words and named them in honour of the French mathematician J. C. F. Sturm. Hedlund and Morse noticed that not every map $f\colon \mathbb{N} \to \mathbb{N}$ is the factor complexity of an infinite word. Infinite words are **eventually periodic**, i.e., they are of the form wv^ω (where v, w are finite words over the corresponding alphabet and ω denotes an infinite repetition) if and only if their factor complexity is eventually constant, i.e., there exists a constant K such that $\mathcal{C}(n) = K$ for sufficiently large n. The words that are not eventually periodic are called **aperiodic**. For aperiodic words Hedlund and Morse showed that for all $n \in \mathbb{N}$ their factor complexity satisfies

$$\mathcal{C}(n) \geq n + 1.$$

1.2.1 Sturmian words

Sturmian words are aperiodic words with the lowest possible complexity.

Definition 1. An infinite word $\mathbf{u}$ is called **Sturmian** if for all $n \in \mathbb{N}$ it holds

$$\mathcal{C}(n) = n + 1.$$

They have been intensively studied from the very beginning. Besides their low factor complexity the reason of their popularity is the fact that the famous Fibonacci word belongs to this class.

Example 2. The fixed point of the morphism $\varphi_{\mathrm{F}} : \{0,1\}^* \to \{0,1\}^*$ defined by $\varphi_{\mathrm{F}}(0) = 01,\ \varphi_{\mathrm{F}}(1) = 0$ is called the **Fibonacci word $\mathbf{u}_{\mathrm{F}}$**. Let us write down a prefix of the Fibonacci word $\mathbf{u}_{\mathrm{F}} = \lim_{n\to+\infty} \varphi_{\mathrm{F}}^n(0) = 0100101001001010010100100101001001\ldots$

The Fibonacci word is closely connected to the Fibonacci numbers. Let us recall that the Fibonacci numbers were introduced by Leonardo of Pisa, known as Fibonacci, in a mathematical game dealing with rabbits: An adult couple (denote it 0) has always after one month a pair of young (denote it 1) and that pair of young grows up after one month. Fibonacci was interested in how large the population of rabbits would be after n months provided the rabbits are immortal.

It is easy to see that the response is provided by the Fibonacci word. If we denote the length of the n-iteration $F_n = |\varphi_{\mathrm{F}}^n(0)|$, it follows that the number of rabbits after n months is equal to F_n. It is not difficult to verify that $F_0 = 1,\ F_1 = 2$ and $F_{n+1} = F_n + F_{n-1}$.

Thanks to a long and fruitful study of Sturmian words, a lot of properties and equivalent definitions of these words are known nowadays.

If $\mathbf{u}$ is a Sturmian word, then $\mathbf{u}$ has the following properties:

- $\mathbf{u}$ is binary;
- $\mathbf{u}$ is aperiodic;
- $\mathcal{L}(\mathbf{u})$ is closed under reversal;
- $\mathcal{L}(\mathbf{u})$ contains infinitely many palindromes;
- $\mathbf{u}$ is uniformly recurrent;
- $\mathcal{L}(\mathbf{u})$ contains no weak bispecial factors.

The following theorem summarizes several well-known combinatorial characterizations of Sturmian words.

Theorem 1. *Let $\mathbf{u}$ be an infinite word. The properties listed below are equivalent:*

(i) $\mathbf{u}$ *is Sturmian, i.e.,* $\mathcal{C}(n) = n + 1$ *for all* $n \in \mathbb{N}$*;*

(ii) $\mathbf{u}$ *is binary and contains a unique left special factor of every length;*

(iii) $\mathbf{u}$ *is binary, aperiodic and every bispecial factor is ordinary;*

(iv) any factor of $\mathbf{u}$ *has exactly two return words;*

(v) $\mathbf{u}$ *contains one palindrome of every even length and two palindromes of every odd length;*

(vi) $\mathbf{u}$ *is binary and every palindrome has a unique palindromic extension.*

Equivalence of the first three statements is evident using (1.1.3), resp. (1.1.4). The characterization by return words is due to Vuillon [37]. The two equivalent properties concerning palindromes have been proved by Droubay and Pirillo [21]. Notice that the fifth property can be equivalently rewritten as

$$\mathcal{P}(n) + \mathcal{P}(n+1) = 3 \quad \text{for all } n \in \mathbb{N},$$

and also as

$$\mathcal{P}(n+2) = \mathcal{P}(n) \quad \text{for all } n \in \mathbb{N}.$$

Let us recall that $\mathcal{P}(0) = 1$ since the empty word is considered to be a palindrome.

1.2.2 Combinatorial Generalizations of Sturmian Words

In the paper **Sturmian Jungle (or Garden?) on Multiliteral Alphabets** [8] (see page 28 for the whole text), we were interested in the generalization of equivalent combinatorial characterizations of Sturmian words to multiliteral alphabets. Let us describe here briefly results of this paper that is a part of the habilitation. We will write down and baptize the generalizations of properties from Theorem 1. Let $\mathbf{u}$ be an infinite word over the alphabet $\mathcal{A}$. Denote $k = \#\mathcal{A}$.

(i) Property $\mathcal{C}$:

the factor complexity of $\mathbf{u}$ satisfies $\mathcal{C}(n) = (k-1)n + 1$ for all $n \in \mathbb{N}$.

(ii) Property $\mathcal{L}R$:

$\mathbf{u}$ contains one left special and one right special factor of every length.

(iii) Property $\mathcal{B}O$:

all bispecial factors of $\mathbf{u}$ are ordinary.

(iv) Property $\mathcal{R}$:

any factor of $\mathbf{u}$ has exactly k return words.

(v) Property $\mathcal{P}$:

the palindromic complexity of $\mathbf{u}$ satisfies $\mathcal{P}(n) + \mathcal{P}(n+1) = k + 1$ for all $n \in \mathbb{N}$.

(vi) Property $\mathcal{PE}$:

every palindrome has a unique palindromic extension in $\mathbf{u}$.

The most studied generalizations of Sturmian words are Arnoux–Rauzy words (AR words) and words coding k-interval exchange transformation (k-iet words). **Arnoux–Rauzy words** are defined as words with the language closed under reversal and having for every length n exactly one left special factor of length n, and moreover, this factor has all possible left extensions. We will not recall the definition of k-iet words here because they will not be treated anywhere else in the text. We will only recall that AR words fulfill all Properties: $\mathcal{C}, \mathcal{L}R, \mathcal{B}O, \mathcal{R}, \mathcal{P}, \mathcal{PE}$ and k-iet words satisfy Properties: $\mathcal{C}, \mathcal{B}O, \mathcal{R}$. If moreover the permutation defining the k-iet word is symmetric, then these words have Properties $\mathcal{P}$ and $\mathcal{PE}$. Property $\mathcal{L}R$ does not hold for k-iet words.

No two of the above properties are equivalent on the set of infinite words over a multiliteral alphabet. We have shown the non-equivalence by counterexamples. However some properties imply others, or it can be shown that a couple of properties are equivalent on a certain class of infinite words. Before listing them, let us point out that there are three types of contributions of the paper [8]:

- a detailed summary of known relations between Properties (i) – (vi);
- discovery of new relations between Properties (i) – (vi);
- a rich summary of new and known examples of concrete infinite words illustrating relations between Properties (i) – (vi).

1. $\mathcal{LR}$ implies none of Properties $\mathcal{C}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$ (easy to find an example).
2. none of Properties $\mathcal{C}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$ imply $\mathcal{LR}$ (known example – k-iet words).

3. $\mathcal{R} \not\Rightarrow \mathcal{P}$ (example taken from our paper [13]).

4. $\mathcal{C} \not\Rightarrow \mathcal{R}$ (known example).

5. Let $\mathbf{u}$ be a uniformly recurrent word. Then $\mathcal{BO} \Rightarrow \mathcal{R}$ (consequence of our results from [13]).

6. Let $\mathbf{u}$ be a uniformly recurrent ternary word. Then $\mathcal{BO} \Leftrightarrow \mathcal{R}$ (consequence of the previous statement and of our results from [13]).

7. Let $\mathbf{u}$ be a uniformly recurrent, but not ternary word. Then $\mathcal{R} \not\Rightarrow \mathcal{BO}$ (example taken from our paper [13]).

8. $\mathcal{PE} \not\Rightarrow \mathcal{C}$ (known example).

9. Let $\mathbf{u}$ be a ternary word with the language closed under reversal. Then $\mathcal{C} \Rightarrow \mathcal{P}$ (consequence of our results from [12]).

10. Let $\mathbf{u}$ be a ternary word with the language closed under reversal. Then $\mathcal{BO} \Rightarrow \mathcal{PE}$ (consequence of our results from [12]).

11. Let $\mathbf{u}$ be a ternary word with the language closed under reversal. Then $\mathcal{PE} \not\Rightarrow \mathcal{C}$ (example taken from our paper [12]).

12. $\mathcal{PE} \Rightarrow \mathcal{P}$ (by definition).

13. $\mathcal{P} \not\Rightarrow \mathcal{PE}$ (example taken from our paper [12]).

14. Let $\mathbf{u}$ be a ternary word with the language closed under reversal. Then $\mathcal{R} \Rightarrow \mathcal{PE}$ (consequence of our results from [12] and [13]).

15. Let $\mathbf{u}$ be a ternary word with the language closed under reversal. Then $\mathcal{PE} \not\Rightarrow \mathcal{R}$ (example taken from our paper [12]).

16. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Assume $\mathcal{C}$. Then $\mathcal{BO} \Leftrightarrow \mathcal{PE}$ (original result).

17. Let $\mathbf{u}$ be a uniformly recurrent word. Assume $\mathcal{C}$. Then $\mathcal{PE} \Rightarrow \mathcal{R}$ (consequence of previous statements).

18. Let $\mathbf{u}$ be a uniformly recurrent word. Assume $\mathcal{C}$. Then $\mathcal{R} \not\Rightarrow \mathcal{P}$ (example taken from our paper [13]).

19. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then $\mathcal{R} \not\Rightarrow \mathcal{P}$ (example taken from our paper [13]).

20. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Assume $\mathcal{C}$. Then $\mathcal{P} \Leftrightarrow$ every non-palindromic BS factor w of $\mathbf{u}$ satisfies $\mathrm{b}(w) = 0$ and every palindromic BS factor w of $\mathbf{u}$ satisfies $\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1$ (original result).

1.3 Palindromes in Nature

We know already that a palindrome is a word that stays the same when read backwards. Nobody can be surprised that in natural languages there are no especially long palindromes. The longest palindromic words in Czech are those ones of the type "nepochopen" (not understood), "nepotopen" (not sunk), "nezasazen" (not planted), "nezařazen" (not filed). In English the longest palindromic word is "tattarrattat". Its victory is however doubtful because it is not a common word but an invention by James Joyce who used in his novel Ulysses [29] this neologism to denote strong knocking on the door:

"I was just beginning to yawn with nerves thinking he was trying to make a fool of me when I knew his tattarrattat at the door."

The sentences might be more interesting from the palindromic point of view. They give rise to palindromes if we release gaps between words and eventually diacritics. In Czech the best known palindromic sentences are the following ones:

"Bažantu padá za záda putna žab." (A bucket of frogs is falling behind the pheasant's back.)
"Jelenovi pivo nelej." (Do not pour beer to deer.)
"Kobyla má malý bok." (Mares have small sides.)

Of course, palindromic numbers might be of interest, too. In particular when there is at least a partial explanation for their reason. For instance there is a palindrome consisting of odd ciphers related to the foundation of the Charles Bridge 135797531. The Museum of the Charles Bridge in Prague uses it as a part of its logo. According to the historian of astronomy Zdeněk Horský the foundation stone might have been laid on July, 9, 1357 at 5:31. There was a favorable constellation of Sun and Saturn at that moment. The palindrome thus consists of the following items: year – day – month – hour – minutes.

1.4 Palindromes in Infinite Words

A much more interesting situation comes to light in the world of infinite words. Palindromes of any length can occur here. However, any anarchy is not reigning here neither.

The palindromic complexity is bounded by the first difference of factor complexity.

Proposition 1 ([3]). *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$\mathcal{P}(n) + \mathcal{P}(n+1) \leq \Delta\mathcal{C}(n) + 2 \quad \textit{for all } n \in \mathbb{N}. \tag{1.4.1}$$

Moreover, an infinite word can contain in any of its factors w at most $|w| + 1$ distinct palindromes (including the empty word). This upper bound on the number of palindromes occurring in a finite word given by Droubay, Justin, and Pirillo [20] initiated many interesting investigations on palindromes in infinite words. A finite word w containing the utmost number $|w| + 1$ of palindromes is called **rich**. An infinite word is said to be **rich** if all its factors are rich. (We keep here the terminology introduced by Glen et al. [24] in 2007, which seems to us to be prevalent nowadays. However, Brlek et al. [15] baptized such words **full** already in 2004.)

There exist several equivalent characterization of rich words. The most recent one using bilateral orders of factors was stated in the paper **Sturmian Jungle (or Garden?) on Multiliteral Alphabets** [8] that is a part of this habilitation.

1.5 Defect and Richness

Brlek et al. [15] suggested to study the **defect** $D(w)$ of a finite word w defined as the difference between the upper bound $|w|+1$ and the actual number of palindromes contained in w. The defect of an infinite word is then defined as the maximal defect of a factor of the infinite word. In this convention, rich words are precisely the words with zero defect.

Since we focus on palindromes in infinite words, richness, and defect, we will introduce several notions and known results related to this topic. Let us remark that not only all prefixes of rich words are rich, but also all factors are rich. A result from [20] provides us with a handful tool which helps to evaluate the defect of a factor.

Proposition 2 ([20]). *A finite or infinite word* $\mathbf{u}$ *is rich if and only if the longest palindromic suffix of* w *occurs exactly once in* w *for any prefix* w *of* $\mathbf{u}$*.*

The longest palindromic suffix of a factor w will occur often in our considerations, therefore we will denote it by $lps(w)$. In accordance with the terminology introduced in [20], the factor with a unique occurrence in another factor is called **unioccurrent**. The proof of the above proposition is based on the fact that there exists a bijection between the set of palindromes contained in w and the first prefixes of w ending in the corresponding palindromes. It follows that the other prefixes cause the increment of the defect.

Corollary 1. *The defect* $D(w)$ *of a finite word* w *is equal to the number of prefixes* w' *of* w*, for which the* $lps(w')$ *is not unioccurrent in* w'*.*

This corollary implies that $D(v) \geq D(w)$ whenever w is a factor of v. It enables to give a reasonable definition of the defect of an infinite word (see [15]).

Definition 2. The defect of an infinite word $\mathbf{u}$ is the number (finite or infinite)

$$D(\mathbf{u}) = \sup\{D(w) \bigm| w \text{ is a prefix of } \mathbf{u}\}\,.$$

Let us point out several facts concerning defects that are easy to prove:

1. If we consider all factors of a finite or an infinite word $\mathbf{u}$, we obtain the same defect, i.e.,

$$D(\mathbf{u}) = \sup\{D(w) \bigm| w \in \mathcal{L}(\mathbf{u})\}\,.$$

2. Any infinite word with finite defect contains infinitely many palindromes.

3. Infinite words with zero defect correspond exactly to rich words.

The authors of [20] who were the first ones to tackle the problem of richness showed that Sturmian and episturmian words are rich. (Let us recall that **episturmian words** are defined as infinite words with the language closed under reversal and having at most one left special factor of every length.) In [15], an insight into the richness of periodic words can be found. Further on, let us summarize equivalent characterizations of rich words.

Theorem 2. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then the following statements are equivalent:*

1. *The word* $\mathbf{u}$ *is rich.*

2. *For any prefix* w *of* $\mathbf{u}$ *the* $lps(w)$ *is unioccurrent in* w*.*

3. *All complete return words of any palindrome in* $\mathbf{u}$ *are palindromes.*

4. *The equality*

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$

   holds for all $n \in \mathbb{N}$.

5. *Any bispecial factor* w *of* $\mathbf{u}$ *satisfies:*
   - *if* w *is non-palindromic, then* $\mathrm{b}(w) = 0$*,*
   - *if* w *is a palindrome, then* $\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1$.

Glen et al. [24] have proved the characterization based on the notion of complete return words, Bucci et al. [17] have characterized richness using the palindromic and factor complexity, and most recently, we have found a new characterization of rich words considering bilateral orders of factors [8].

Our aim in the sequel is to introduce the papers [9, 10, 11] devoted to the study of defect. All of them are embodied in this habilitation. Periodic words with finite defect have been studied in [15] and in [24]. It holds that the defect of an infinite periodic word with the minimal period w is finite if and only if $w = pq$, where both p and q are palindromes. In [24] words with finite defect have been baptized **almost rich**.

1.5.1 Almost Rich Words

The following characterization of infinite words with finite defect – called as well almost rich words – follows from observations made in [24].

Theorem 3. *Let* $\mathbf{u}$ *be a uniformly recurrent word containing infinitely many palindromes. Then the following statements are equivalent:*

1. *The word* $\mathbf{u}$ *is almost rich.*

2. *There exists an integer* H *such that for any prefix* w *of* $\mathbf{u}$ *with* $|w| \geq H$*, the* $lps(w)$ *is unioccurrent in* w.

3. *There exists an integer* K *such that all complete return words of any palindrome in* $\mathbf{u}$ *of length at least* K *are palindromes.*

It is easy to see that the second statement of Theorem 3 can be equivalently rewritten as: There exists an integer H such that for any factor w of $\mathbf{u}$ with $|w| \geq H$, the $lps(w)$ is unioccurrent in w. Let us stress that if we put in the previous theorem $D(\mathbf{u}) = K = H = 0$, all statements become known results for rich words, see Theorem 2.

Example 3. Let us provide an example of a uniformly recurrent word $\mathbf{u}$ with finite defect and let us find for $\mathbf{u}$ the lowest values of constants K and H from Theorem 3. Take the Fibonacci word $\mathbf{u}_{\mathrm{F}}$, i.e., the fixed point of φ_{F}: $0 \to 01,\ 1 \to 0$. Define $\mathbf{u}$ as its morphic image $\sigma(\mathbf{u}_{\mathrm{F}})$, where σ: $0 \to cabcbac,\ 1 \to d$, i.e.,

$$\mathbf{u} = cabcbacdcabcbaccabcbacdcabcbacdcabcbaccabcbacdcabcbaccabcbacdcabcbacdcabcbac\ldots$$

It is easy to show that all palindromes of length greater than one and the palindromes a, b, and d have only palindromic complete return words. Hint: long palindromes in $\mathbf{u}$ contain in their center images of non-empty palindromes from $\mathbf{u}_{\mathrm{F}}$ that have palindromic complete return words by the richness of $\mathbf{u}_{\mathrm{F}}$. The only non-palindromic complete return word of c is *cabc*. In order to show that $D(\mathbf{u}) = 1$, it suffices to verify that no prefixes longer than *cabc* have c as their longest palindromic suffix. This follows directly from the form of σ. The lowest values of the constants K and H are: $K = 2$, $H = 5$.

In the paper **Infinite Words with Finite Defect** [9] (the complete text starts on the page 57), we have proved a new characterization of infinite words with finite defect based on a relation between the palindromic and factor complexity.

Theorem 4 ([9])**.** *Let* $\mathbf{u}$ *be a uniformly recurrent word. Then* $\mathbf{u}$ *is almost rich if and only if there exists an integer* N *such that*

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$

holds for all $n \geq N$*.*

Notice that if we set $N = 0$ in the previous theorem, then we obtain the known characterization of rich words from Theorem 2 (which holds even under a weaker assumption that $\mathcal{L}(\mathbf{u})$ is closed under reversal).

We will present here only the main ingredient of the proof of Theorem 4. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Using the proof of Proposition 1, those $n \in \mathbb{N}$ for which the equality

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$$

holds can be characterized in the graph language.

An n-**simple path** e is a factor of $\mathbf{u}$ of length at least $n+1$ such that the only special (right or left) factors of length n occurring in e are its prefix and suffix of length n. If w is the prefix of e of length n and v is the suffix of e of length n, we say that the n-simple path e starts in w and ends in v. We will denote by $G_n(\mathbf{u})$ an undirected graph whose set of vertices is formed by unordered pairs $\{w, \overline{w}\}$ such that $w \in \mathcal{L}_n(\mathbf{u})$ is right or left special. We connect two vertices $\{w, \overline{w}\}$ and $\{v, \overline{v}\}$ by an unordered pair $\{e, \overline{e}\}$ if e or $\overline{e}$ is an n-simple path starting in w or $\overline{w}$ and ending in v or $\overline{v}$. Note that the graph $G_n(\mathbf{u})$ may have multiple edges and loops.

Lemma 1 ([9])**.** *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Let* $n \in \mathbb{N}$*. Then* $\mathcal{P}(n) + \mathcal{P}(n+1) = \Delta\mathcal{C}(n) + 2$ *if and only if both of the following conditions are met:*

1. *The graph obtained from* $G_n(\mathbf{u})$ *by removing loops is a tree.*

2. *Any* n*-simple path forming a loop in the graph* $G_n(\mathbf{u})$ *is a palindrome.*

1.5.2 The Brlek–Reutenauer Conjecture

Despite the fact that numerous researchers study palindromes, only recently Brlek and Reutenauer [16] have noticed that the value of defect is closely tied with the expression in (1.4.1) – let us denote for an infinite word $\mathbf{u}$ by $T_{\mathbf{u}}(n) = \Delta\mathcal{C}(n) + 2 - \mathcal{P}(n) - \mathcal{P}(n+1)$. They have shown that for a periodic infinite word $\mathbf{u}$ with the language closed under reversal, it holds $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$. Their conjecture says that the same equality holds for all infinite words with the language closed under reversal.

Conjecture 1 (The Brlek–Reutenauer conjecture)**.** *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)\,. \tag{1.5.1}$$

It is known from the paper [17] that Conjecture 1 holds for rich words (both sides equal zero for them). As we have already mentioned, Brlek and Reutenauer provided a result for periodic words.

Theorem 5 ([16]). *Let $\mathbf{u}$ be a periodic infinite word. Then* (1.5.1) *holds.*

Partial Proof of the Brlek–Reutenauer Conjecture

In the paper **On the Brlek–Reutenauer Conjecture** [10] (see the complete version on the page 71), we have proved their conjecture for uniformly recurrent words.

Theorem 6 ([10]). *If $\mathbf{u}$ is a uniformly recurrent infinite word with the language closed under reversal, then* (1.5.1) *holds.*

In the proof of Theorem 6 we used our result from [9] (recalled here as Theorem 4) to show that either both sides of (1.5.1) are finite, or both of them are infinite. Further on, the main idea was to construct for any almost rich uniformly recurrent word $\mathbf{u}$ a periodic word $\mathbf{v}$ satisfying $D(\mathbf{u}) = D(\mathbf{v})$ and $T_{\mathbf{u}}(n) = T_{\mathbf{v}}(n)$ for all $n \in \mathbb{N}$. The proof was then finished because by Theorem 5 the conjecture holds for periodic words. However, in the construction we had to prove one more quite interesting statement on increasing squares in uniformly recurrent words with finite defect.

Lemma 2 ([10]). *Let $\mathbf{u}$ be an almost rich uniformly recurrent infinite word. Then the set*

$$\{w \in \mathcal{A}^* \mid ww \in \mathcal{L}(\mathbf{u})\}$$

is infinite.

The proof relied essentially on the uniform recurrence of the infinite word in question.

Proof of the Brlek–Reutenauer Conjecture

In the paper **Proof of the Brlek–Reutenauer Conjecture** [11] (consult the paper on the page 79), we managed to find completely different arguments than in the previous paper [10] that enabled us to prove Conjecture 1 in full generality without exploiting the result for periodic words.

Theorem 7 ([11]). *Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)\,. \tag{1.5.2}$$

Let us provide the main ideas of the proof. We divided the proof into two steps:

1. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Assume $D(\mathbf{u}) < +\infty$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty$. Then $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$.

   Since the defect is finite, there exists H_1 such that $D(q) = D(\mathbf{u})$ for every prefix q of $\mathbf{u}$ of length greater than or equal to $H_1 - 1$. By the finiteness of $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$, there exists H_2 such that $T_{\mathbf{u}}(n) = 0$ for all $n \geq H_2$. Set $H = \max\{H_1, H_2\}$ and find a prefix p of $\mathbf{u}$ containing all factors of length H. Clearly, $D(p) = D(\mathbf{u})$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n)$.

   In the sequel, we will use a theorem by Brlek and Reutenauer for finite words [16].

Theorem 8 ([16]). *For every finite word w we have*

$$2D(w) = \sum_{n=0}^{|w|} T_w(n),$$

where $T_w(n) = \Delta\mathcal{C}_w(n) + 2 - \mathcal{P}_w(n+1) - \mathcal{P}_w(n)$ and the index w means that we consider only factors of w.

We deduce the following equalities:

$$2D(\mathbf{u}) = 2D(p) = \sum_{n=0}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_p(n) + \sum_{n=H}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n),$$

where everything follows from the previous arguments; only the equality $\sum_{n=H}^{|p|} T_p(n) = 0$ remains to be explained. Let us rewrite it in the following form:

$$\begin{aligned} \textstyle\sum_{n=H}^{|p|} T_p(n) &= \textstyle\sum_{n=H}^{|p|} \Big(\mathcal{C}_p(n+1) - \mathcal{C}_p(n) + 2 - \mathcal{P}_p(n+1) - \mathcal{P}_p(n)\Big) \\ &= -\mathcal{C}_p(H) + 2(|p| - H + 1) - 2\textstyle\sum_{n=H}^{|p|} \mathcal{P}_p(n) + \mathcal{P}_p(H). \end{aligned}$$

We have to explain why the last expression equals zero: The factors of length H are either palindromes of length H – their number is equal to $\mathcal{P}_p(H)$ – or they are non-palindromic factors whose longest palindromic suffix is of course of length less than H. Such non-palindromic factors are certainly not contained in the prefix q of length $H-1$, therefore their longest palindromic suffix is unioccurrent. Consequently, to any palindrome of length less than H that does not occur in q, there are exactly two non-palindromic factors w and $\overline{w}$ having it as its longest palindromic suffix. The number of such factors is therefore given by twice the number of palindromes of length less than H that are not contained in q, which is equal to $2(|p| - |q| - \sum_{n=H}^{|p|} \mathcal{P}_p(n))$. This concludes the proof.

2. Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then $D(\mathbf{u})$ is finite if and only if $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ is finite.

   This second part was proved using the graph theory, more precisely, using mainly Lemma 1.

Chapter 2

Applications of Combinatorics on Words in Random Number Generation

We have seen that Combinatorics on Words may be considered as a more than 100 years old discipline founded by Axel Thue. When studying square-free and cube-free words, Thue explained that he had no concrete application in mind, but he studied the questions because he found them interesting. Unlike Thue we will show in this chapter that Combinatorics on Words may be applied in such a practical domain as pseudorandom number generation.

Pseudorandom number generators aim to produce random numbers using a deterministic process. No wonder they suffer from many defects. The most usual ones in the past – linear congruential generators – are known to produce periodic sequences having a defect called the lattice structure. Guimond et al. [27] proved that when two linear congruential generators are combined using infinite words coding certain classes of quasicrystals or, equivalently, of cut-and-project sets, the resulting sequence is aperiodic and has no lattice structure. For some other related results concerning aperiodic pseudorandom generators we refer to [25, 26].

In the papers **Pseudorandom Number Generators Based on Infinite Words** [6] and in its preliminary version **Infinite Words with Well Distributed Occurrences** [5], we have substantially generalized the results of Guimond et al. [25, 26, 27]. For Czech readers, we have in an accessible way presented the topic in [7]. The papers [5] and [6] (for their full version see pages 86, resp. 99) make part of this habilitation.

We have found a combinatorial condition – **well distributed occurrences**, or WELLDOC for short (the acronym WDO instead of WELLDOC was used in the preliminary paper [5]) – that also guarantees absence of the lattice structure. The WELLDOC property for an infinite word $\mathbf{u}$ over an alphabet $\mathcal{A}$ means that for any integer m and any factor w of $\mathbf{u}$, the set of Parikh vectors modulo m of prefixes of $\mathbf{u}$ preceding the occurrences of w coincides with $\mathbb{Z}_m^{|\mathcal{A}|}$. In other words, among Parikh vectors modulo m of such prefixes one has all possible vectors, where the **Parikh vector** of a finite word v over an alphabet $\mathcal{A}$ has its i-th component equal to the number of occurrences of the i-th letter of $\mathcal{A}$ in v. Besides giving generators without lattice structure, the WELLDOC property is an interesting combinatorial property of infinite words itself. We have proved that the WELLDOC property holds for the family of Sturmian words, and more generally for Arnoux–Rauzy words.

An infinite word with the WELLDOC property is then used to combine two linear congruential generators (or other periodic generators) and form an infinite aperiodic sequence with good statistical behavior. Using the TestU01 [23] and PractRand [19] statistical tests, we have moreover shown that not only the lattice structure is absent, but also other important properties of pseudorandom number generators are improved when linear congruential generators are combined using infinite words having the WELLDOC property. Let us emphasize that the theoretical results on aperiodicity and absence of the lattice structure hold when combining any two periodic generators with the same output, not only linear congruential generators. In practice, we recommend of course to use better generators than linear congruential generators as the underlying ones.

2.1 Lattice Structure of Pseudorandom Number Generators

In the sequel, for the seek of simplicity, we understand under the notion of pseudorandom number generator (PRNG) any sequence of nonnegative integers. The most illustrative example of generators having the weakness called lattice structure are linear congruential generators (already from dimension two, as shown by Marsaglia [32]). Let $Z = (Z_n)_{n\in\mathbb{N}}$ be a PRNG whose output is a finite set $M \subset \mathbb{N}$. We say that Z has the **lattice structure** if there exists a positive integer t such that the set

$$\{(Z_i, Z_{i+1}, \ldots, Z_{i+t-1}) \bigm| i \in \mathbb{N}\}$$

is covered by a family of equidistant parallel hyperplanes in the Euclidean space $\mathbb{R}^t$ and at the same time, these hyperplanes do not cover all points of the lattice

$$M^t = \{(A_1, A_2, \ldots, A_t) \bigm| A_i \in M \text{ for all } i \in \{1, \ldots, t\}\}.$$

Let us recall that a **linear congruential generator** (LCG for short) $(Z_n)_{n\in\mathbb{N}}$ is given by parameters $a, m, c \in \mathbb{N}$ and defined by the recurrence relation $Z_{n+1} = aZ_n + c \mod m$. An example of an LCG with striking lattice structure is the generator RANDU, widely used in the sixties of the 20th century. For $t = 3$ the consecutive triples of the generator RANDU, i.e., $\{(Z_i, Z_{i+1}, Z_{i+2}) \bigm| i \in \mathbb{N}\}$, are covered by as few as 15 equidistant parallel hyperplanes that do not cover by far the whole lattice $\{1, 2, \ldots, m-1\}^3$, see Figure 2.1.

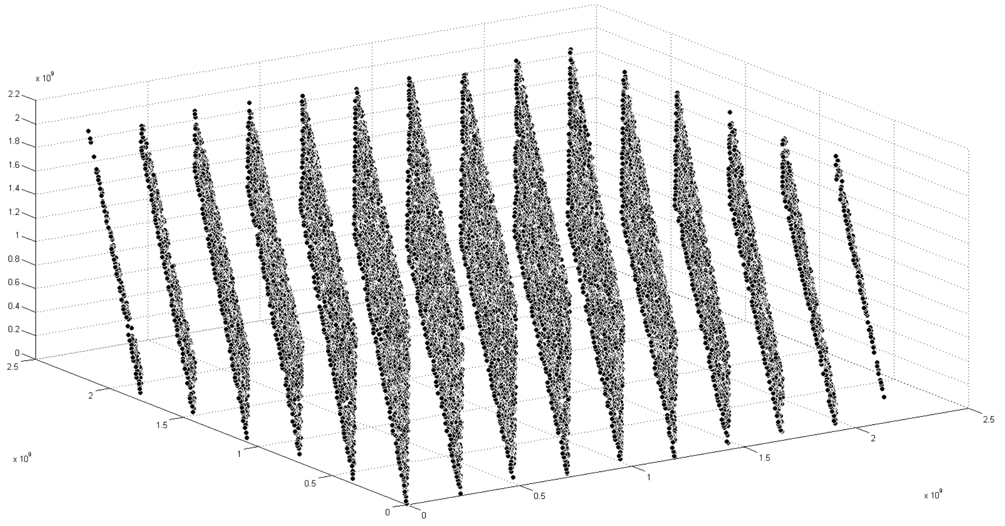


Fig. 2.1: Triples of the RANDU generator, LCG with parameters $a = (2^{16} + 3), m = 2^{31}, c = 0$, are covered by as few as 15 equidistant parallel hyperplanes.

Guimond et al. [27] provided a sufficient condition for absence of the lattice structure – even if they formulated it in a less general form (Lemma 2.3).

Theorem 9 ([27])**.** *Let Z be a PRNG whose output is a finite set $M \subset \mathbb{N}$ containing at least two elements. Assume that for every $A, B \in M$ and for every $\ell \in \mathbb{N}$ there exists an ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ such that both $(A_1, A_2, \ldots, A_\ell, A)$ and $(A_1, A_2, \ldots, A_\ell, B)$ are $(\ell+1)$-tuples of the generator Z. Then Z has no lattice structure.*

In words of Combinatorics on Words, the statement of the previous theorem may be reformulated in the following form: Let Z be a PRNG whose output is a finite set $M \subset \mathbb{N}$ containing at least two elements. If Z contains for any two letters $A, B \in M$ and for any length $\ell \in \mathbb{N}$ a right special factor of length ℓ with right extensions A and B, then Z has no lattice structure.

2.2 Combination of Pseudorandom Number Generators

If we want to eliminate the lattice structure, it helps to combine generators in a smart way. One of such methods has been introduced by Guimond et al. [25]. Let $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ be two not necessarily distinct PRNGs with the same output $M \subset \mathbb{N}$ and the same period $m \in \mathbb{N}$. Let $\mathbf{u}$ be an infinite binary word over the alphabet $\{a, b\}$. Then the **generator**

$$Z = (Z_n)_{n\in\mathbb{N}} \tag{2.2.1}$$

based on the word u is obtained by the following algorithm:

1. Read step by step the letters of $\mathbf{u}$.
2. If you read a for the ith time, copy the ith symbol of X to the end of the constructed sequence Z.
3. If you read b for the ith time, copy the ith symbol of Y to the end of the constructed sequence Z.

It is easy to generalize this construction for infinite words over larger alphabets and to combine thus more generators.

Example 4. Let $X = (X_1X_2X_3X_4)^\omega$, $Y = (Y_1Y_2Y_3Y_4)^\omega$ and $\mathbf{u} = (abb)^\omega$. Then

$$Z = X_1Y_1Y_2X_2Y_3Y_4X_3Y_1Y_2X_4Y_3Y_4X_1Y_1Y_2\ldots$$

2.3 Words with Well Distributed Occurrences

Guimond et al. [27] proved that PRNGs based on a certain class of infinite words coding cut-and-project sequences are aperiodic and have no lattice structure. We have generalized their result and have found larger classes of words that guarantee aperiodicity and absence of the lattice structure for generators based on such words [5, 6].

Definition 3 (WELLDOC property for binary words)**.** We say that an aperiodic infinite word $\mathbf{u}$ over $\{a, b\}$ has **well distributed occurrences** (or has the **WELLDOC property**) if $\mathbf{u} = u_0u_1u_2\ldots$ satisfies for every positive integer m and for every factor w of $\mathbf{u}$ the following condition: If we denote $i_0, i_1, i_2, \ldots$ the occurrences of w in the word $\mathbf{u}$, then

$$\left\{\left(|u_0u_1\ldots u_{i_j-1}|_a, |u_0u_1\ldots u_{i_j-1}|_b\right) \bmod m \mid j \in \mathbb{N}\right\} = \mathbb{Z}_m^2\,,$$

where mod m is applied component-wise and $|w|_a$ denotes the number of occurrences of a in the word w and $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$.

The WELLDOC property is defined for aperiodic words since it is evident that it never holds for eventually periodic words. For words with well distributed occurrences we have proved the following theorem.

Theorem 10 ([5])**.** *Let Z be a PRNG from* (2.2.1) *based on a binary infinite word* $\mathbf{u}$ *with the WELLDOC property. Then Z is aperiodic and has no lattice structure.*

If we are given two concrete generators with period m and we want to guarantee absence of the lattice structure only for them, it is sufficient to check whether the WELLDOC property holds for the modulus equal to m.

Example 5. Let us show at first that there exist infinite words which do not have the WELLDOC property. Consider the Thue–Morse word defined in Section 1.1.1

$$\mathbf{u}_{\mathrm{TM}} = abbabaabbaababbabaababbaabbabaab\ldots,$$

being the fixed point of the morphism $\varphi_{\mathrm{TM}}(a) = ab$, $\varphi_{\mathrm{TM}}(b) = ba$.

Indeed, if we take $m = 2$ and $w = aa$, then it follows from the form of the morphism φ_{TM} that w has only odd occurrences i_j. For example

$$\begin{array}{lll} i_0 = 5, & u_0 \ldots u_4 = abbab, & (|u_0 \ldots u_4|_a, |u_0 \ldots u_4|_b) = (2,3), \\ i_1 = 9, & u_0 \ldots u_8 = abbabaabb, & (|u_0 \ldots u_8|_a, |u_0 \ldots u_8|_b) = (4,5), \\ i_2 = 17, & u_0 \ldots u_{16} = abbabaabbaababbab, & (|u_0 \ldots u_{16}|_a, |u_0 \ldots u_{16}|_b) = (8,9). \end{array}$$

Therefore $|u_0u_1 \ldots u_{i_j-1}|_a + |u_0u_1 \ldots u_{i_j-1}|_b = i_j$ is an odd number. Consequently, for instance

$$(|u_0u_1 \ldots u_{i_j-1}|_a, |u_0u_1 \ldots u_{i_j-1}|_b) \bmod 2 \neq (0,0).$$

We introduced a combinatorial condition – well distributed occurrences – guaranteeing absence of the lattice structure. Then it was important to find words having such a property. We have proved [5, 6] that among binary words, Sturmian words have the WELLDOC property.

Besides the generalization of the results from [27] for binary infinite words, we have found a combinatorial condition for words over multiliteral alphabets; the generators based on such words combine then more input generators and have again no lattice structure. It turned out that it suffices to generalize the WELLDOC property in the most natural way to larger alphabets.

Definition 4 (WELLDOC property for multiliteral words)**.** We say that an aperiodic infinite word $\mathbf{u}$ over the alphabet $\{a_1, a_2, \ldots, a_d\}$ has **well distributed occurrences** (or has the **WELLDOC property**) if $\mathbf{u}$ satisfies for every positive integer m and for every factor w of the word $\mathbf{u}$ the following condition: If we denote $i_0, i_1, i_2, \ldots$ the occurrences of w in the word $\mathbf{u}$, then

$$\left\{\left(|u_0u_1 \ldots u_{i_j-1}|_{a_1}, |u_0u_1 \ldots u_{i_j-1}|_{a_2}, \ldots, |u_0u_1 \ldots u_{i_j-1}|_{a_d}\right) \bmod m \mid j \in \mathbb{N}\right\} = \mathbb{Z}_m^d\,,$$

where mod m is applied component-wise.

Remark 1. The WELLDOC property is sufficient, however it is not necessary for absence of the lattice structure. It is not difficult to show that the modified Fibonacci word that is obtained from the Fibonacci word over the alphabet $\{a, b\}$ when replacing step by step a with ac and b with bc, i.e.,

$$\mathbf{u} = acbcacacbcacbcac\ldots,$$

does not have the WELLDOC property, but it does not produce generators with the lattice structure neither. Nevertheless, we have seen in statistical tests that the WELLDOC property is important not only for absence of the lattice structure, but it probably guarantees other good properties of the combined generators, too.

Example 6. Let us provide at first a trivial example of a word with well distributed occurrences. We say that an infinite word over the alphabet $\mathcal{A}$ is **universal** if it contains all finite words over $\mathcal{A}$ as its factors. A universal word over $\{0, 1, \dots, d-1\}$ may be obtained for instance by concatenation of representations of consecutive natural numbers in base d. It is readily seen that universal words have the WELLDOC property.

Similarly as over the binary alphabet, we had to show over multiliteral alphabets that there exists a large class of infinite words with the WELLDOC property. We have shown that Arnoux–Rauzy words (AR words) have well distributed occurrences. The best known example of such words over the ternary alphabet is the Tribonacci word $\mathbf{u}_{\mathrm{T}}$ being the fixed point of the substitution $\varphi_{\mathrm{T}}(a) = ab,\ \varphi_{\mathrm{T}}(b) = ac,\ \varphi_{\mathrm{T}}(c) = a$, i.e.,

$$\mathbf{u}_{\mathrm{T}} = abacabaabacab\dots$$

Remark 2. There exists a simple method of how to produce from words with well distributed occurrences again words with well distributed occurrences over smaller alphabets. If the alphabet equals $\mathcal{A} = \{a_1, a_2, \dots, a_d\}$, where $d \geq 3$, and the word $\mathbf{u}$ over $\mathcal{A}$ has the WELLDOC property, then if we replace in $\mathbf{u}$ the letter a_d with another arbitrary, but fixed letter a_i, then the arising word will have again well distributed occurrences. In such a way we obtain words different from AR words.

Remark 3. A further transformation of words that preserves the WELLDOC property and this time it preserves the alphabet as well uses a morphism whose matrix is unimodular, i.e., has the determinant equal to ±1. Under the **matrix of a morphism** φ over the alphabet $\{a_1, a_2, \dots, a_d\}$ is understood the matrix $\mathbf{\Phi}$ whose ijth element is defined as $\mathbf{\Phi}_{ij} = |\varphi(a_i)|_{a_j}$.

Example 7. Consider the morphism $\varphi\colon\ a \to aab,\ b \to ab$. Then its matrix is of the form $\mathbf{\Phi} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. If we apply this morphism to the Fibonacci word $\mathbf{u}_{\mathrm{F}} = abaababaabaab\dots$, then we get the word $\varphi(\mathbf{u}_{\mathrm{F}}) = aababaabaababaababaabaababaabaabab\dots$ with well distributed occurrences.

2.4 Statistical Tests of Pseudorandom Number Generators

In the previous part we have seen that PRNGs based on infinite words with the WELLDOC property are aperiodic and have no lattice structure. In the sequel, we will show that such generators have as well much better results in empirical statistical tests than the original generators that are combined. This programming part is due to Jiří Hladký, one of the authors of [6].

It is necessary to generate efficiently long prefixes of infinite words with the WELLDOC property for both practical use and for statistical tests of PRNGs based on such infinite words. Fortunately, a lot of Sturmian and Arnoux–Rauzy words, i.e., infinite words with well distributed occurrences as we have learned in the previous section, are fixed points of morphisms. Such fixed points may be generated in a fast way [35]: the prefix of length n is obtained in time $O(n)$ and with memory consumption $O(\log n)$. Hladký suggested an improvement: in order to speed up the generation of prefixes, it is better to remember instead of letter images $\varphi(a)$ a suitable higher iteration $\varphi^k(a)$ for all $a \in \mathcal{A}$. Thanks to that simple improvement, the speed of generation became higher than the speed of generation of the output of the combined LCGs. For instance 10^{10} 32-bit values of the output of an LCG with modulus 2^{64} were obtained in 14.3 seconds, while using the same hardware, 10^{10} letters of a fixed point were generated in 0.5 seconds. To sum up, use of PRNGs based on infinite words that are fixed points of a morphism means only a negligible time penalization in comparison to use of the original generators.

Let us show here results of statistical tests of PRNGs based on:

- the Fibonacci word (as an example of Sturmian words): the fixed point of the morphism $a \mapsto ab$ and $b \mapsto a$;
- the Tribonacci word (as an example of ternary AR words): the fixed point of the morphism $a \mapsto ab, b \mapsto ac, c \mapsto a$.

PRNGs based on fixed points of morphisms were implemented for more Sturmian and AR words. Since the results are similar, we will present here only the two above cases. A program generating PRNGs based on fixed points of morphisms is available online together with a description [28].

When combining LCGs, we do not use all output bits. We combine LCGs having the period between $2^{47} - 115$ and 2^{64}, however we consider only 32 upper bits as their output. The reason is that 32-bit sequences are needed as the input of the batteries of statistical tests we have chosen.[1] We apply two batteries of statistical tests – TestU01 BigCrush and PractRand. They work differently. The first one contains 160 statistical tests – many of them are tailored to specific types of generators. It is a battery of tests with a good reputation, nevertheless its disadvantage is that it works with a fixed number of bits and it always throws away the two lowest bits. The second battery is composed of three types of tests: the first one focuses on short range correlations, the second one reveals long range violations, and the last one is a variation on the classical gap test. The PractRand tests are able to treat very long input sequences, up to a few exabytes. To control the runtime we have limited the length of input sequences to 16TB.

The first column of Table 2.1 shows the list of tested LCGs. The BigCrush column shows how many tests of the TestU01 BigCrush battery failed. The PractRand column gives the $\log_2$ of sample datasize in bytes for which the results of the PractRand tests started to be "very suspicious" (p-values smaller than 10^{-5}). One LCG did not show any failures in the PractRand tests which is denoted as > 44 – the meaning is that the PractRand test has passed successfully $16\text{TB} \doteq 2^{44}\text{B}$ of input data and the test was stopped there. The last column provides time in seconds to generate the first 10^{10} 32-bit sequences of output on Intel i7-3520M CPU running at 2.90GHz.

Generator	Symbol	BigCrush	PractRand	Time 10^{10}
$\mathrm{LCG}(2^{47} - 115, 71971110957370, 0)$	L47-115	14	40	281 s
$\mathrm{LCG}(2^{63} - 25, 2307085864, 0)$	L63-25	2	>44	277 s
$\mathrm{LCG}(2^{59}, 13^{13}, 0)$	L59	19	27	14.1 s
$\mathrm{LCG}(2^{63}, 5^{19}, 1)$	L63	19	33	14.4 s
$\mathrm{LCG}(2^{64}, 2862933555777941757, 1)$	L64_28	18	35	14.0 s
$\mathrm{LCG}(2^{64}, 3202034522624059733, 1)$	L64_32	14	34	14.1 s
$\mathrm{LCG}(2^{64}, 3935559000370003845, 1)$	L64_39	13	33	14.0 s

Tab. 2.1: The list of used $\mathrm{LCG}(m, a, c)$ and their results in tests BigCrush and PractRand.

From Table 2.1 we can observe that LCGs with $m \in \{2^{47} - 115, 2^{63} - 25\}$ give the best statistical results. At the same time, they are much slower than the other generators. The

[1] We do not use LCGs with modulo 2^{32} because for such generators it is known that their k-th bit has the period of length at most 2^k.

reason is that we have to really compute the operation modulo, while for $m = 2^\ell$ we only shift the binary point.

In Tables 2.2 and 2.3 we summarize results for PRNGs based on the Fibonacci and the Tribonacci word and combining various LCGs from Table 2.1. (We combine in such a way that reading of the letter a in the infinite word corresponds to taking the output of the generators in the column a and analogously for the other letters b, c and columns b, c.) It includes also the situations where instances of the same LCG are used. In such a case, the LCGs were seeded with the value 1. The PRNGs were warmed up by generating 10^9 values before statistical tests started. Since the relative frequency of letters in the aperiodic words differ a lot (for example for the Fibonacci word the ratio of zeroes to ones is given by $\tau = \frac{1+\sqrt{5}}{2}$), the warming procedure will guarantee that the state of instances of LCGs will differ even when the same LCGs are used.

The BigCrush column is using the following notation: the first number indicates how many tests from the BigCrush battery have clearly failed and the optional second number in parenthesis denotes how many tests have suspiciously low p-value in the range from 10^{-6} to 10^{-4}. The rest of the notation has been explained above, it stays the same as in Table 2.1.

Word	Group	a	b	BigCrush	PractRand	Time 10^{10}
Fib	A	L64_28	L64_28	0	41	30.2 s
	A	L64_32	L64_28	0(1)	41	29.3 s
	A	L64_39	L64_28	0 (2)	41	31 s
	A	L64_28	L64_32	0	41	30.2 s
	A	L64_32	L64_32	0	41	30.1 s
	B	L47-115	L47-115	1(1)	>44	302 s
	B	L63-25	L63-25	0(1)	>44	299 s
	B	L59	L59	0(1)	34	28.7 s
	C	L63-25	L59	0	38	198 s
	C	L59	L63-25	0(1)	35	134 s
	C	L63-25	L64_39	0	>44	199 s
	C	L64_39	L63-25	0	41	135 s
	C	L59	L64_39	0	35	30.4 s
	C	L64_39	L59	0	37	31.3 s

Tab. 2.2: Summary of results of statistical tests for generators based on the Fibonaci word and various combinations of LCGs from Table 2.1.

The following observations are based on the results in statistical tests:

1. The quality of LCGs has increased substantially when they were combined using words with well distributed occurrences. This is visible in the BigCrush test. While for the

Word	Group	a	b	c	BigCrush	PractRand	Time 10^{10}
Trib	A	L64_28	L64_28	L64_28	0(2)	42	27.2
	A	L64_39	L64_28	L64_28	0	43	27.1
	A	L64_39	L64_32	L64_28	0(1)	42	28.0
	A	L64_28	L64_39	L64_28	0(1)	42	28.1
	A	L64_32	L64_39	L64_28	0	42	27.1
	B	L47-115	L47-115	L47-115	1	>44	299.0
	B	L63-25	L63-25	L63-25	0(1)	>44	298.0
	B	L59	L59	L59	0	35	27.2
	B	L63	L63	L63	0(1)	41	27.2
	C	L63-25	L59	L64_39	0(1)	39	172.0
	C	L63-25	L64_39	L59	0(1)	41	173.0
	C	L59	L63-25	L64_39	0	35	106.0
	C	L59	L64_39	L63-25	0	34	70.5
	C	L64_39	L63-25	L59	0	41	107.0
	C	L64_39	L59	L63-25	0(1)	40	74.3

Tab. 2.3: Summary of results of statistical tests for generators based on the Tribonacci word and various combinations of LCGs from Table 2.1.

original LCGs 13 to 19 tests failed (see Table 2.1), after the combination almost all of them have passed. The worst result was one failed test for the combination based on the Tribonacci word, resp. on the Fibonacci word for LCG L47-115. The likely reason is however the shortest period of this generator among all tested ones. The results of the PractRand battery confirms an improvement, too. For instance, in the case of the LCGs with modulus 2^{64} (see the group A in Tables 2.2 and 2.3), the test begins to find irregularities in the distribution of the last bit only for the output of 2TB in the case of the Fibonacci word and 4TB to 8TB in the case of the Tribonacci word. The reader may compare this with 8GB to 32GB, where the original LCGs showed up weaknesses in this test.

2. The quality of the arising generators is closely related to the quality of the input generators, see for instance some generators of the group B in Tables 2.2 and 2.3 that had good results already themselves.

3. Another interesting observation is the fact that using LCGs with the same parameters and different initial values does not spoil anything. It is only important to check that the initial values are distinct enough. See the groups A and B in Tables 2.2 and 2.3.

4. On one hand, if we combine LCGs of different quality, then the weakest LCG determines the quality of the arising generator. See the group C in Tables 2.2 and 2.3. If we cannot avoid combining LCGs of different quality, we should at least use the best one so that it

correspond to the most frequent letter in the word with well distributed occurrences. On the other hand, if we combine generators of similar quality, then their order does not play any role. See the group A in Tables 2.2 and 2.3.

5. The Tribonacci word shows better results than the Fibonacci word. This observation holds for all tested AR words in comparison to Sturmian words.

2.5 Open Problems

There are a lot of open problems left. Concerning the combinatorial part, it would be good to find other large classes of infinite words with well distributed occurrences and to detect among fixed points of morphisms those ones with the WELLDOC property. It seems as well meaningful to study the WELLDOC property for some particular modulae, where we consider in the definition of the WELLDOC property one specific m instead of all $m \in \mathbb{N}$. Concerning the statistical tests, the field of open questions is even larger – besides aperiodicity and absence of the lattice structure no other success of PRNGs based on infinite words with well distributed occurrences in empirical tests has been explained theoretically. We continue of course in statistical tests of PRNGs combining other than LCGs, too, and our aim is to provide a useful, practical, well documented open source library of "our" PRNGs in the future.

Bibliography

[1] J.-P. Allouche, J. Shallit, *The ubiquitous Prouhet-Thue-Morse sequence*, Sequences and their applications, Proceedings of SETA'98, C. Ding, T. Helleseth and H. Niederreiter (Eds.)(1999), Springer Verlag, 1–16.

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

[3] P. Baláži, Z. Masáková, E. Pelantová, *Factor versus palindromic complexity of uniformly recurrent infinite words*, Theoret. Comput. Sci. **380** (2007), 266–275.

[4] L'. Balková, *Nahlédnutí pod pokličku kombinatoriky na nekonečných slovech*, PMFA **56** (2011), 9–18.

[5] L'. Balková, M. Bucci, A. De Luca, S. Puzynina, *Infinite words with well distributed occurrences*, In: J. Karhumäki, A. Lepistö, L. Zamboni (Eds.), Combinatorics on Words, LNCS **8079**, Springer (2013), 46–57

[6] L'. Balková, M. Bucci, A. De Luca, J. Hladký, S. Puzynina, *Pseudorandom number generators based on infinite words*, submitted to Math. Comput. (2014)

[7] L'. Balková, J. Hladký, *Generátory pseudonáhodných čísel založené na nekonečných slovech*, PMFA **59** (2014), 211–222.

[8] L'. Balková, E. Pelantová, Š. Starosta, *Sturmian jungle (or garden?) on multiliteral alphabets*, RAIRO Theor. Inf. Appl. **44** (2010), 443–470.

[9] L'. Balková, E. Pelantová, Š. Starosta, *Infinite words with finite defect*, Adv. Appl. Math. **47** (2011), 562–574.

[10] L'. Balková, E. Pelantová, Š. Starosta, *On the Brlek–Reutenauer conjecture*, Theoret. Comput. Sci. **412** (2011), 5649–5655.

[11] L'. Balková, E. Pelantová, Š. Starosta, *Proof of the Brlek–Reutenauer conjecture*, Theoret. Comput. Sci. **475** (2013), 120–125.

[12] L'. Balková, E. Pelantová, Š. Starosta, *Palindromes in infinite ternary words*, RAIRO Theor. Inf. Appl. **43** (2009), 687–702.

[13] L'. Balková, E. Pelantová, W. Steiner, *Sequences with constant number of return words*, Monatsh. Math. **155** (2008), 251–263.

[14] L'. Balková, E. Pelantová, Š. Starosta, *Corrigendum: "On Brlek-Reutenauer conjecture"*, Theoret. Comput. Sci. **465** (2012), 73–74.

[15] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, *On the palindromic complexity of infinite words*, Internat. J. Found. Comput. Sci. **2** (2004), 293–306.

[16] S. Brlek, Ch. Reutenauer, *Complexity and palindromic defect of infinite words*, Theoret. Comput. Sci. **412** (2011), 493–497.

[17] M. Bucci, A. De Luca, A. Glen, L. Q. Zamboni, *A connection between palindromic and factor complexity using return words*, Adv. Appl. Math **42** (2009), 60–74.

[18] J. Cassaigne, *Complexity and special factors*, Bull. Belg. Math. Soc. Simon Stevin **4** (1997), 67–88.

[19] Ch. Doty-Humphrey, *Practically Random: C++ library of statistical tests for RNGs*, https://sourceforge.net/projects/pracrand

[20] X. Droubay, J. Justin, G. Pirillo, *Episturmian words and some constructions of de Luca and Rauzy*, Theoret. Comput. Sci. **255** (2001), 539–553.

[21] X. Droubay, G. Pirillo, *Palindromes and Sturmian words*, Theoret. Comput. Sci. **223** (1999), 73–85.

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

[23] P. L'Ecuyer, R. Simard, *TestU01: A C library for empirical testing of random number generators*, ACM Trans. Math. Software **33** (2007)

[24] A. Glen, J. Justin, S. Widmer, L. Q. Zamboni, *Palindromic richness*, European J. Combin. **30** (2009), 510–531.

[25] L.-S. Guimond, Jan Patera, Jiří Patera, *Combining random number generators using cut-and-project sequences*, Czech. J. Phys. **51** (2001), 305–311.

[26] L.-S. Guimond, J. Patera, *Proving the deterministic period breaking of linear congruential generators using two tile quasicrystals*, Math. Comput. **71** (2002), 319–332.

[27] L.-S. Guimond, Jan Patera, Jiří Patera, *Statistical properties and implementation of aperiodic pseudorandom number generators*, Appl. Numer. Math. **46** (2003), 295–318.

[28] J. Hladký, *Random number generators based on the aperiodic infinite words*, https://github.com/jirka-h/aprng

[29] J. Joyce, *Ulysses*, Sylvia Beach's Shakespeare and Company in Paris, 1922.

[30] M. Lothaire, *Combinatorics on words*, Addison-Wesley, 1983

[31] M. Lothaire, *Algebraic combinatorics on words*, Cambridge University Press, 2002.

[32] G. Marsaglia, *Random numbers fall mainly in the planes*, Proc. Natl. Acad. Sci. **61** (1968), 25–28.

[33] M. Morse, *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.

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

[35] Jan Patera, *Generating the Fibonacci chain in $O(\log n)$ space and $O(n)$ time*, Phys. Part. Nuclei **33** (2002), 118–122.

[36] A. Thue, *Über unendliche Zeichenreihen*, Norske vid. Selsk. Skr. Mat. Nat. Kl. **7** (1906), 1–22.

[37] L. Vuillon, *A characterization of Sturmian words by return words*, European J. Combin. **22** (2001), 263–275.

Sturmian Jungle (or Garden?) on Multiliteral Alphabets

Authors:	**Ľubomíra Balková, Edita Pelantová, Štěpán Starosta**
Journal:	RAIRO – Theoretical Informatics and Applications
Year:	2010
Volume:	44
Pages:	443–470

RAIRO-Theor. Inf. Appl. 44 (2010) 443–470
DOI: 10.1051/ita/2011002

Available online at:
www.rairo-ita.org

STURMIAN JUNGLE (OR GARDEN?) ON MULTILITERAL ALPHABETS

Ľubomíra Balková[1], Edita Pelantová[1]
and Štěpán Starosta[1]

Abstract. The properties characterizing Sturmian words are considered for words on multiliteral alphabets. We summarize various generalizations of Sturmian words to multiliteral alphabets and enlarge the list of known relationships among these generalizations. We provide a new equivalent definition of rich words and make use of it in the study of generalizations of Sturmian words based on palindromes. We also collect many examples of infinite words to illustrate differences in the generalized definitions of Sturmian words.

Mathematics Subject Classification. 68R15.

1. Introduction

Sturmian words, *i.e.*, aperiodic words with the lowest factor complexity, appeared first in the paper of Hedlund and Morse in 1940. Since then Sturmian words have been in the center of interest of many mathematicians and the number of discoveries of new properties and connections keeps growing. The charm of Sturmian words consists in their natural appearance while studying diverse problems. Many equivalent definitions have been found that way. Sturmian words are binary and every property characterizing Sturmian words asks for a fruitful extension to an analogy on a larger alphabet. Well-known examples of such efforts are Arnoux-Rauzy words, words coding interval exchange transformations, or billiard words. All these words belong to well established classes and their descriptions

Keywords and phrases. Sturmian words, generalizations of Sturmian words, palindromes, rich words.

[1] Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic; lubomira.balkova@fjfi.cvut.cz; edita.pelantova@fjfi.cvut.cz; staroste@fjfi.cvut.cz

Article published by EDP Sciences

and properties can be found in many works [5,6,11,27,35,40,46]. An overview of some generalizations of Sturmian words is provided in [12,50].

The aim of this paper is to attract attention to other generalizations of Sturmian words. Our motivation stems from recent results on palindromes in infinite words that have ended in the definition of words rich in palindromes, the definition of defect, the description of a relation between factor and palindromic complexity, etc. [3,7,15]. Impulses for such an intensive research of palindromes come concededly from the article [22] which characterizes Sturmian words by palindromes, the article [23] which investigates the number of palindromes in prefixes of infinite words and last, but not least, the discovery of the role of palindromes in description of the spectrum of Schrödinger operators with aperiodic potentials [31]. While generalizing Sturmian words we have taken into consideration the characterization of Sturmian words by return words from [49] and a recent definition of Abelian complexity [42,43], which is closely connected with balance properties.

We consider the following properties (k denotes the cardinality of alphabet $\mathcal{A}$):

(1) Property $\mathcal{C}$:
the factor complexity of $\mathbf{u}$ satisfies $\mathcal{C}(n) = (k-1)n + 1$ for all $n \in \mathbb{N}$.

(2) Property $\mathcal{LR}$:
$\mathbf{u}$ contains one left special and one right special factor of every length.

(3) Property $\mathcal{BO}$:
all bispecial factors of $\mathbf{u}$ are ordinary.

(4) Property $\mathcal{R}$:
any factor of $\mathbf{u}$ has exactly k return words.

(5) Property $\mathcal{P}$:
the palindromic complexity of $\mathbf{u}$ satisfies $\mathcal{P}(n) + \mathcal{P}(n+1) = k + 1$ for all $n \in \mathbb{N}$.

(6) Property $\mathcal{PE}$:
every palindrome has a unique palindromic extension in $\mathbf{u}$.

(7) Balance properties:

(a) Property $\mathcal{B}_\forall$:
$\mathbf{u}$ is aperiodic and for all $a \in \mathcal{A}$ and for all factors $w, v \in \mathcal{L}(\mathbf{u})$ with $|w| = |v|$ it holds

$$||w|_a - |v|_a| \le k - 1.$$

(b) Property $\mathcal{B}_\exists$:
$\mathbf{u}$ is aperiodic and there exists $a \in \mathcal{A}$ such that for all factors $w, v \in \mathcal{L}(\mathbf{u})$ with $|w| = |v|$ it holds

$$||w|_a - |v|_a| \le k - 1.$$

(c) Property $\mathcal{AC}$:
$\mathbf{u}$ is aperiodic and the abelian complexity of $\mathbf{u}$ satisfies $\mathcal{AC}(n) = k$ for all $n \in \mathbb{N},\ n \ge 1$.

All properties are equivalent on a binary alphabet and they characterize Sturmian words. No two of them are equivalent on the set of infinite words over a multiliteral alphabet. The non-equivalence is shown by counterexamples. However some properties imply others, or it can be shown that a couple of properties are equivalent on a certain class of infinite words. For instance, on the class of uniformly recurrent ternary words properties $\mathcal{R}$ and $\mathcal{BO}$ are equivalent.

There exist more equivalent definitions of Sturmian words, for instance the definition based on balance properties of subfactors of factors [25], on the index of an infinite word [37], or Richomme's characteristics of Sturmian words [41]. We do not pay attention to these definitions in our survey.

The paper is organized as follows. In Section 2 we recall the notions playing an important role in the definitions of properties (1) through (7). We recall the notion of substitution which is irrelevant for the generalizations of Sturmian words but is used to construct most of examples of infinite words. Section 3 is focused on the study of palindromes in infinite words: we summarize older and new results concerning palindromes, we define palindromic branches. A new result in this section is Theorem 3.10 providing a new characterization of rich words by means of bilateral orders. Section 4 shortly summarizes essential results on Sturmian words. Section 5 is devoted to an overview of known relations among different generalizations of Sturmian words, mostly from articles [7,9,16,30,42,43]. New results are in Theorems 5.9 and 5.13, and Corollaries 5.11 and 5.12. The last section is a brief summary of selected relations and examples illustrating the studied properties.

2. Notation and definitions

By $\mathcal{A}$ we denote a finite set of symbols, usually called *letters*; the set $\mathcal{A}$ is therefore called an *alphabet*. A finite string $w = w_0w_1 \dots w_{n-1}$ of letters of $\mathcal{A}$ is said to be a *finite word*, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ε as the neutral element form a free monoid $\mathcal{A}^*$. The map

$$w = w_0w_1 \dots w_{n-1} \quad \mapsto \quad \tilde{w} = w_{n-1}w_{n-2} \dots w_0$$

is a bijection on $\mathcal{A}^*$, the word $\tilde{w}$ is called the *reversal* or the *mirror image* of w. A word w which coincides with its mirror image is a *palindrome*.

Under an *infinite word* $\mathbf{u}$ over the alphabet $\mathcal{A}$ we understand an infinite string $\mathbf{u} = u_0u_1u_2 \dots$ of letters from $\mathcal{A}$ such that every letter of $\mathcal{A}$ occurs in $\mathbf{u}$. We call an infinite word $\mathbf{u}$ *eventually periodic* if there exist finite words w, v such that $\mathbf{u} = wv^\omega$, where ω means ‘repeated infinitely many times’. If $w = \varepsilon$, then $\mathbf{u}$ is said to be *(purely) periodic*. If $\mathbf{u}$ is not eventually periodic, then we call $\mathbf{u}$ *aperiodic*.

A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \varepsilon$, then w is said to be a *prefix* of v, if $s = \varepsilon$, then w is a *suffix* of v. We say that a prefix or a suffix is *proper* if it is not equal to the word itself.

The *language* $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. The factors of $\mathbf{u}$ of length n form the set denoted by $\mathcal{L}_n(\mathbf{u})$. Using this notation, we may write $\mathcal{L}(\mathbf{u}) = \cup_{n\in\mathbb{N}}\mathcal{L}_n(\mathbf{u})$.

We say that the language $\mathcal{L}(\mathbf{u})$ is *closed under reversal* if $\mathcal{L}(\mathbf{u})$ contains with every factor w also its reversal $\tilde{w}$.

An infinite word $\mathbf{u}$ over $\mathcal{A}$ is called *c-balanced* if for every $a \in \mathcal{A}$ and for every pair of factors w, v of $\mathbf{u}$ of the same length $|w| = |v|$, we have $||w|_a - |v|_a| \leq c$, where $|w|_a$ means the number of letters a contained in w. Note that in the case of a binary alphabet, say $\mathcal{A} = \{0, 1\}$, this condition may be rewritten in a simpler way: an infinite word $\mathbf{u}$ is *c-balanced*, if for every pair of factors w, v of $\mathbf{u}$ with $|w| = |v|$, we have $||w|_0 - |v|_0| \leq c$. We call 1-balanced words simply *balanced*.

We say that two words $w, v \in \mathcal{A}^*$ are *abelian equivalent* if for each letter $a \in \mathcal{A}$, it holds $|w|_a = |v|_a$. It is easy to see that the abelian equivalence defines indeed an equivalence relation on $\mathcal{A}^*$. If $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$, then the *Parikh vector* associated with the word $w \in \mathcal{A}^*$ is defined as

$$\Psi(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k}).$$

We call *abelian complexity* (as defined in [42]) of an infinite word $\mathbf{u}$ the function $\mathcal{AC} : \mathbb{N} \to \mathbb{N}$ given by

$$\mathcal{AC}(n) = \#\{\Psi(w) \bigm| w \in \mathcal{L}_n(\mathbf{u})\}.$$

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2}\ldots$ Such an index i is called an *occurrence* of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ has at least two occurrences in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be *recurrent*. It can be easily shown that each factor of a recurrent word occurs infinitely many times. It is readily seen that if the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent. The infinite word $\mathbf{u}$ is said to be *uniformly recurrent* if for any factor w of $\mathbf{u}$ the distances between successive occurrences of w form a bounded sequence.

Let j, k, $j < k$, be two successive occurrences of a factor w in $\mathbf{u}$. Then $u_ju_{j+1}\ldots u_{k-1}$ is called a *return word* of w. Return words were first studied in [24,32]. The set of all return words of w is denoted by $R(w)$,

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

If v is a return word of w, then the word vw is called a *complete return word* of w. It is obvious that an infinite recurrent word is uniformly recurrent if and only if the set of return words of any of its factors is finite.

The *(factor) complexity* of an infinite word $\mathbf{u}$ is the map $\mathcal{C} : \mathbb{N} \mapsto \mathbb{N}$, defined by $\mathcal{C}(n) = \#\mathcal{L}_n(\mathbf{u})$. To determine the increment of complexity, one has to count the possible extensions of factors of length n. A *left extension* of $w \in \mathcal{L}(\mathbf{u})$ is any letter $a \in \mathcal{A}$ such that $aw \in \mathcal{L}(\mathbf{u})$. The set of all left extensions of a factor w will be denoted by $\mathrm{Lext}(w)$. We will mostly deal with recurrent infinite words $\mathbf{u}$. In this case, any factor of $\mathbf{u}$ has at least one left extension. A factor w is called *left*

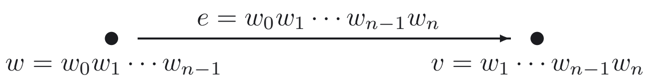

$e = w_0w_1\cdots w_{n-1}w_n$
$w = w_0w_1\cdots w_{n-1}$
$v = w_1\cdots w_{n-1}w_n$

Figure 1. Incidence relation between an edge and vertices in a Rauzy graph.

special (or LS for short) if w has at least two left extensions. Clearly, any prefix of a LS factor is LS as well. It makes therefore sense to define an *infinite LS branch* which is an infinite word whose all prefixes are LS factors of $\mathbf{u}$. Similarly, one can define a *right extension*, a *right special* (or RS) factor, $\mathrm{Rext}(w)$, and an *infinite RS branch* which is a left-sided infinite word whose all suffixes are RS factors of $\mathbf{u}$.

We say that a factor w of $\mathbf{u}$ is a *bispecial* (or BS) factor if it is both RS and LS. The role of BS factors for the computation of complexity can be nicely illustrated on Rauzy graphs (introduced in [6]).

Let $\mathbf{u}$ be an infinite word and $n \in \mathbb{N}$. The *Rauzy graph* Γ_n of $\mathbf{u}$ is a directed graph whose set of vertices is $\mathcal{L}_n(\mathbf{u})$ and set of edges is $\mathcal{L}_{n+1}(\mathbf{u})$. An edge $e \in \mathcal{L}_{n+1}(\mathbf{u})$ starts in the vertex w and ends in the vertex v if w is a prefix and v is a suffix of e, see Figure 1. If the word $\mathbf{u}$ is recurrent, the graph Γ_n is strongly connected for every $n \in \mathbb{N}$, *i.e.*, there exists a directed path from every vertex w to every vertex v of the graph.

If the language $\mathcal{L}(\mathbf{u})$ of the infinite word $\mathbf{u}$ is closed under reversal, then the operation that to every vertex w of the graph associates its mirror image, the vertex $\tilde{w}$, and to every edge e associates $\tilde{e}$ maps the Rauzy graph Γ_n onto itself.

The *outdegree* (*indegree*) of a vertex $w \in \mathcal{L}_n(\mathbf{u})$ is the number of edges which start (end) in w. Obviously the outdegree of w is equal to $\#\mathrm{Rext}(w)$ and the indegree of w is $\#\mathrm{Lext}(w)$. The sum of outdegrees over all vertices is equal to the number of edges in every directed graph. Similarly, it holds for indegrees. In particular, for the Rauzy graph Γ_n we have

$$\sum_{w\in\mathcal{L}_n(\mathbf{u})} \#\mathrm{Rext}(w) \;=\; \mathcal{C}(n+1) \;=\; \sum_{w\in\mathcal{L}_n(\mathbf{u})} \#\mathrm{Lext}(w).$$

The first difference of complexity $\Delta\mathcal{C}(n) = \mathcal{C}(n+1) - \mathcal{C}(n)$ is thus given by

$$\Delta\mathcal{C}(n) = \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Rext}(w) - 1\big) \;=\; \sum_{w\in\mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Lext}(w) - 1\big).$$

A non-zero contribution to $\Delta\mathcal{C}(n)$ in the left-hand sum is given only by those factors $w \in \mathcal{L}_n(\mathbf{u})$ for which $\#\mathrm{Rext}(w) \geq 2$, and for recurrent words, a non-zero contribution to $\Delta\mathcal{C}(n)$ in the right-hand sum is provided only by those factors $w \in \mathcal{L}_n(\mathbf{u})$ for which $\#\mathrm{Lext}(w) \geq 2$. The last relation can be thus rewritten for recurrent words $\mathbf{u}$ as

$$\Delta\mathcal{C}(n) = \sum_{w\in\mathcal{L}_n(\mathbf{u}),\ w\ \mathrm{RS}} \big(\#\mathrm{Rext}(w) - 1\big) \;=\; \sum_{w\in\mathcal{L}_n(\mathbf{u}),\ w\ \mathrm{LS}} \big(\#\mathrm{Lext}(w) - 1\big).$$

If we denote Bext(w) $= \{awb \in \mathcal{L}(\mathbf{u}) \mid a, b \in \mathcal{A}\}$, then the second difference of complexity $\Delta^2\mathcal{C}(n) = \Delta\mathcal{C}(n+1) - \Delta\mathcal{C}(n) = \mathcal{C}(n+2) - 2\mathcal{C}(n+1) + \mathcal{C}(n)$ is given by

$$\Delta^2\mathcal{C}(n) = \sum_{w \in \mathcal{L}_n(\mathbf{u})} \big(\#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1\big). \tag{2.1}$$

Denote by $\mathrm{b}(w)$ the quantity

$$\mathrm{b}(w) := \#\mathrm{Bext}(w) - \#\mathrm{Rext}(w) - \#\mathrm{Lext}(w) + 1.$$

The number $\mathrm{b}(w)$ is called the *bilateral order* of the factor w and was introduced in [18]. It is readily seen that if w is not a BS factor, then $\mathrm{b}(w) = 0$. Bispecial factors are distinguished according to their bilateral order in the following way

- if $\mathrm{b}(w) > 0$, then w is a *strong* BS factor;
- if $\mathrm{b}(w) < 0$, then w is a *weak* BS factor;
- if $\mathrm{b}(w) = 0$ then w is an *ordinary* BS factor.

A *substitution* on $\mathcal{A}$ is a morphism $\varphi : \mathcal{A}^* \to \mathcal{A}^*$ such that there exists a letter $a \in \mathcal{A}$ and a non-empty word $w \in \mathcal{A}^*$ satisfying $\varphi(a) = aw$ and $\varphi(b) \neq \varepsilon$ for all $b \in \mathcal{A}$. Since a morphism satisfies $\varphi(vw) = \varphi(v)\varphi(w)$ for all $v, w \in \mathcal{A}^*$, any substitution is uniquely determined by the images of letters. Instead of classical $\varphi(a) = w$, we sometimes write $a \to w$. A substitution can be naturally extended to an infinite word $\mathbf{u} = u_0u_1u_2\dots$ by the prescription $\varphi(\mathbf{u}) = \varphi(u_0)\varphi(u_1)\varphi(u_2)\dots$ An infinite word $\mathbf{u}$ is said to be a *fixed point* of the substitution φ if it fulfills $\mathbf{u} = \varphi(\mathbf{u})$. It is obvious that every substitution φ has at least one fixed point, namely $\lim_{n\to\infty} \varphi^n(a)$ (to be understood in the sense of product topology).

3. Words opulent in palindromes

In resemblance to the factor complexity $\mathcal{C}(n)$ of an infinite word $\mathbf{u}$, let us define the *palindromic complexity* of $\mathbf{u}$ as the map $\mathcal{P} : \mathbb{N} \to \mathbb{N}$ given by

$$\mathcal{P}(n) = \#\{w \in \mathcal{L}_n(\mathbf{u}) | \ w = \tilde{w}\}.$$

If $a \in \mathcal{A}$ and w is a palindrome and $awa \in \mathcal{L}(\mathbf{u})$, then awa is said to be a *palindromic extension* of w. The set of all palindromic extensions of w is denoted by $\mathrm{Pext}(w)$.

Similarly as in the case of left special and right special branches, one can define a *palindromic branch* of $\mathbf{u}$.

Definition 3.1. Let $\mathbf{u}$ be an infinite word. A two-sided infinite word $v = \dots v_3v_2v_1 v_1v_2v_3\dots$ is a palindromic branch with center ε of the word $\mathbf{u}$ if for every $n \in \mathbb{N}$ the word $v_nv_{n-1}\dots v_2v_1v_1v_2\dots v_{n-1}v_n$ is a factor of $\mathbf{u}$. Let a be a letter. A two-sided infinite word $v = \dots v_3v_2v_1av_1v_2v_3\dots$ is a palindromic branch with center a of the word $\mathbf{u}$ if for every $n \in \mathbb{N}$ the word $v_nv_{n-1}\dots v_2v_1av_1v_2\dots v_{n-1}v_n$ is a factor of $\mathbf{u}$.

It follows from the König's theorem that if $\mathbf{u}$ has infinitely many palindromes, then $\mathbf{u}$ has at least one palindromic branch. In any Sturmian word on $\{0,1\}$ there exist exactly three palindromic branches with centers ε, 0 and 1. See also Section 5.1.

Uniformly recurrent words containing infinitely many distinct palindromes satisfy that for any factor w, every sufficiently large palindrome in $\mathbf{u}$ contains w, thus such a palindrome contains $\tilde{w}$ as well. As a consequence, we have the following theorem.

Theorem 3.2. *If $\mathbf{u}$ is a uniformly recurrent word that contains infinitely many distinct palindromes, then its language $\mathcal{L}(\mathbf{u})$ is closed under reversal.*

The opposite implication is not true as illustrated by the following example.

Example 3.1 (uniform recurrence + language closed under reversal $\not\Rightarrow$ infinitely many palindromes)**.** The infinite word $\mathbf{u}$ on $\{a,b\}$ (constructed in [13]) whose prefixes u_n are given by the following recurrent formula

$$u_0 = ab, \quad u_{n+1} = u_n ab \widetilde{u_n},$$

is uniformly recurrent and its language is closed under reversal. However, $\mathbf{u}$ contains only a finite number of palindromes.

When we relax the condition of uniform recurrence, the statement of Theorem 3.2 is not true any more.

Example 3.2 (infinitely many palindromes $\not\Rightarrow$ language closed under reversal)**.** The infinite word $\mathbf{u}$ on $\{a,b,c\}$ whose prefixes u_n are given by the following recurrent formula

$$u_0 = \varepsilon, \quad u_{n+1} = u_n abc^{n+1} u_n$$

is clearly recurrent. Infinitely many palindromes are represented by the factors c^n for every n. As the factor ba does not occur, the set of factors is not closed under reversal. A similar example can be found in [16].

The word $\mathbf{u}$ may be recoded to a binary alphabet while preserving the mentioned properties. We may for instance recode $\mathbf{u}$ using the following mapping:

$$a \to 0110, \ b \to 1001, \ c \to 1.$$

An interesting relation between the palindromic and factor complexity has been revealed in [7].

Theorem 3.3. *Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then*

$$\mathcal{P}(n+1) + \mathcal{P}(n) \le \Delta\mathcal{C}(n) + 2 \quad \textit{for all } n \in \mathbb{N}. \tag{3.1}$$

In fact, the above relation is stated in [7] for uniformly recurrent words, however the proof requires only recurrent words. Theorem 3.3 implies that infinite words reaching the equality in (3.1) are in a certain sense opulent in palindromes. Another measure of opulence in palindromes has been provided in [23].

Theorem 3.4. *Every finite word w contains at most $|w|+1$ palindromes (including the empty word).*

Definition 3.5. An infinite word $\mathbf{u}$ satisfying that every factor w of $\mathbf{u}$ contains $|w|+1$ palindromes is called rich in palindromes.

The following equivalent definitions of richness have been proved in [30], [16], [17], respectively.

Theorem 3.6. *For any infinite word $\mathbf{u}$ the following conditions are equivalent:*

(1) $\mathbf{u}$ is rich;
(2) any complete return word of a palindromic factor of $\mathbf{u}$ is a palindrome;
(3) for any factor w of $\mathbf{u}$, every factor of $\mathbf{u}$ that contains w only as its prefix and $\tilde{w}$ only as its suffix is a palindrome;
(4) each factor of $\mathbf{u}$ is uniquely determined by its longest palindromic prefix and its longest palindromic suffix.

We will need for our further purposes an implication that holds only for languages closed under reversal.

Corollary 3.7 [16]**.** *Let $\mathbf{u}$ be a rich infinite word with the language closed under reversal. Then for any factor w of $\mathbf{u}$, the occurrences of w and $\tilde{w}$ alternate.*

A natural question is whether infinite words reaching the equality in (3.1) coincide with rich words. The following theorem proved in [16] provides an answer.

Theorem 3.8. *Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then $\mathbf{u}$ is rich if and only if $\mathcal{P}(n+1)+\mathcal{P}(n)=\Delta\mathcal{C}(n)+2$ for all $n \in \mathbb{N}$.*

Let us mention as an open problem the following question. "Does the equivalence of richness and the equality in (3.1) hold for a larger class than words with the language closed under reversal? For instance for all recurrent words?"

The following observations may serve as hints:

- It does not hold for non-recurrent infinite words in general. The infinite word ab^ω is given in [16] as an example of a rich non-recurrent infinite word (with the language of course not closed under reversal), which does not reach the equality in (3.1) for all $n \in \mathbb{N}$.
- Notice that both rich infinite words and infinite words reaching the equality in (3.1) contain infinitely many palindromes.
- If $\mathbf{u}$ is rich and recurrent, then $\mathcal{L}(\mathbf{u})$ is closed under reversal (proved in [30], Prop. 2.11).

The rest of this section is devoted to the relation between richness and bilateral orders of factors. The following proposition reveals some information on bilateral orders of palindromic bispecial factors in an infinite word with the language closed under reversal.

Proposition 3.9. *Let $\mathbf{u}$ be an infinite word whose language is closed under reversal. Then the bilateral order $\mathrm{b}(w)$ of a palindromic bispecial factor $w \in \mathcal{L}(\mathbf{u})$ has a different parity than the number of palindromic extensions of w.*

Proof. Let w be a palindromic BS factor of $\mathbf{u}$. On one hand, as the language is closed under reversal, we have $\#\mathrm{Lext}(w) = \#\mathrm{Rext}(w)$. Consequently, from the definition of bilateral order one can see that the parity of $\#\mathrm{Bext}(w)$ is different from the parity of $\mathrm{b}(w)$. On the other hand, the parity of the number of palindromic extensions of w equals the parity of $\#\mathrm{Bext}(w)$ since for any $a, b \in \mathcal{A}$, if $awb \in \mathcal{L}(\mathbf{u})$, then $bwa \in \mathcal{L}(\mathbf{u})$. □

In the sequel, we will state and prove a new equivalent definition of rich words by means of bilateral orders.

Theorem 3.10. *Let $\mathbf{u}$ be an infinite word with the language closed under reversal. Then $\mathbf{u}$ is rich if and only if any bispecial factor w of $\mathbf{u}$ satisfies:*

- *if w is non-palindromic, then*
$$\mathrm{b}(w) = 0;$$
- *if w is a palindrome, then*
$$\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1.$$

The following lemma will provide the most important tool for the proof of Theorem 3.10.

Lemma 3.11. *Let $\mathbf{u}$ be a rich infinite word whose language is closed under reversal. Then it holds for any bispecial factor w:*

- *if w is non-palindromic, then*
$$\mathrm{b}(w) \geq 0;$$
- *if w is a palindrome, then*
$$\mathrm{b}(w) \geq \#\mathrm{Pext}(w) - 1.$$

Proof. Let w be a non-palindromic BS factor. By the definition of $\mathrm{b}(w)$, we want to prove
$$\#\mathrm{Bext}(w) \geq \#\mathrm{Rext}(w) + \#\mathrm{Lext}(w) - 1.$$
We will construct a bipartite oriented graph G having its set of vertices V defined as
$$V = \{wa | a \in \mathrm{Rext}(w)\} \cup \{\tilde{w}a | a \in \mathrm{Rext}(\tilde{w})\}.$$
There is an oriented edge from wa to $\tilde{w}b$ if there exists a factor $vb \in \mathcal{L}(\mathbf{u})$ such that wa is its prefix, $\tilde{w}b$ is its suffix and factors w and $\tilde{w}$ occur each exactly once in vb. Furthermore, there is an oriented edge from $\tilde{w}x$ to wy if there exists a factor $vy \in \mathcal{L}(\mathbf{u})$ such that $\tilde{w}x$ is its prefix, wy is its suffix and factors w and $\tilde{w}$ occur each exactly once in v.

Due to Theorem 3.6, such a factor v is a palindrome. Therefore the existence of an edge from wa to $\tilde{w}b$ implies $a\tilde{w}b \in \mathcal{L}(\mathbf{u})$, and so $bwa \in \mathcal{L}(\mathbf{u})$, too. Analogously, if there is an edge from $\tilde{w}x$ to wy, we have $xwy \in \mathcal{L}(\mathbf{u})$.

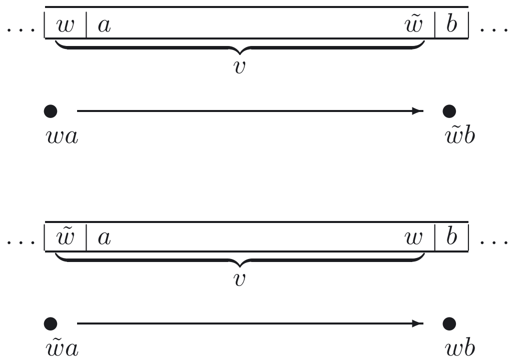


FIGURE 2. Incidence relation in the graph G.

By Corollary 3.7, the occurrences of w and $\tilde{w}$ alternate. Thus, to any factor of $\mathbf{u}$ corresponds a path in G. As $\mathbf{u}$ is recurrent, the graph G is strongly connected.

As a consequence, the number of pairs of its vertices which are connected by an edge is greater than or equal to the number of its vertices minus 1. We have

$$\#\text{Bext}(w) \geq \#\text{Rext}(w) + \#\text{Rext}(\tilde{w}) - 1.$$

Since $\text{Rext}(\tilde{w}) = \text{Lext}(w)$ the proof of the first part is finished.

Let w be a palindromic BS factor. Let us consider this time a graph G whose set of factors V is defined as

$$V = \{wa | a \in \text{Rext}(w)\}\,.$$

There is an edge from wa to wb if there exists a factor $vb \in \mathcal{L}(\mathbf{u})$ such that v is a complete return word to w that has wa as a prefix. As $\mathbf{u}$ is rich, v is a palindrome. Due to the recurrence of $\mathbf{u}$, for every $awb \in \mathcal{L}(\mathbf{u})$, $a \neq b$, there exists an edge in G going from wa to wb. As the language is closed under reversal, the edge going from wb to wa is in G, too. Therefore

$$\#\left\{awb \in \mathcal{L}(\mathbf{u}) | a \neq b\right\} = 2 \times \text{ the number of pairs of distinct vertices connected by an edge.}$$

Owing to the recurrence of $\mathbf{u}$, the graph G is strongly connected, thus the number of pairs of distinct vertices connected by an edge is greater than or equal to the number of vertices of G minus 1, which equals $\#\text{Rext}(w) - 1$. We find

$$\#\text{Bext}(w) = \#\left\{awb \in \mathcal{L}(\mathbf{u}) | a \neq b\right\} + \#\text{Pext}(w) \geq 2\left(\#\text{Rext}(w) - 1\right) + \#\text{Pext}(w).$$

As $\text{Rext}(w) = \text{Lext}(w)$, the statement is proved. □

Proof of Theorem 3.10. ($\Leftarrow$): Let us show by mathematical induction that

$$\Delta\mathcal{C}(n) + 2 = \mathcal{P}(n+1) + \mathcal{P}(n) \quad \text{for all } n \in \mathbb{N}.$$

Since $\mathcal{L}(\mathbf{u})$ is closed under reversal, this means by Theorem 3.8 that $\mathbf{u}$ is rich.

The assumption on bilateral orders and the fact that non-bispecial palindromic factors have a unique palindromic extension guarantee the following equality for all $n \in \mathbb{N}$:

$$\Delta^2\mathcal{C}(n) = \sum_{w \in \mathcal{L}_n(\mathbf{u})} \mathrm{b}(w) = \sum_{\substack{w \in \mathcal{L}_n(\mathbf{u}) \\ w = \tilde{w}}} (\#\mathrm{Pext}(w) - 1) = \mathcal{P}(n+2) - \mathcal{P}(n). \tag{3.2}$$

For $n = 0$, we can write $\Delta\mathcal{C}(0) + 2 = \mathcal{C}(1) - \mathcal{C}(0) + 2 = \#\mathcal{A} + 1$. On the other hand we have $\mathcal{P}(1) + \mathcal{P}(0) = \#\mathcal{A} + 1$.
Take $N \in \mathbb{N}$. Assume $\Delta\mathcal{C}(n) + 2 = \mathcal{P}(n+1) + \mathcal{P}(n)$ holds for all $n < N$. Using the induction assumption and (3.2), we obtain

$$\begin{aligned}
\Delta\mathcal{C}(N) + 2 &= (\Delta\mathcal{C}(N) - \Delta\mathcal{C}(N-1)) + (\Delta\mathcal{C}(N-1) + 2) \\
&= \Delta^2\mathcal{C}(N-1) + (\mathcal{P}(N-1) + \mathcal{P}(N)) \\
&= (\mathcal{P}(N+1) - \mathcal{P}(N-1)) + (\mathcal{P}(N-1) + \mathcal{P}(N)) \\
&= \mathcal{P}(N+1) + \mathcal{P}(N).
\end{aligned}$$

($\Rightarrow$): Take $n \in \mathbb{N}$ arbitrary. We will prove the statement of the theorem for all BS factors of length n.

As $\mathbf{u}$ is rich and the language $\mathcal{L}(\mathbf{u})$ is closed under reversal, we have by Theorem 3.8

$$\Delta\mathcal{C}(k) + 2 = \mathcal{P}(k+1) + \mathcal{P}(k) \quad \text{for all } k \in \mathbb{N}.$$

Applying this equality, we will deduce the form of $\Delta^2\mathcal{C}(n)$.

$$\begin{aligned}
\Delta^2\mathcal{C}(n) &= (\Delta\mathcal{C}(n+1) + 2) - (\Delta\mathcal{C}(n) + 2) \\
&= (\mathcal{P}(n+2) + \mathcal{P}(n+1)) - (\mathcal{P}(n+1) + \mathcal{P}(n)) \\
&= \mathcal{P}(n+2) - \mathcal{P}(n).
\end{aligned}$$

Consequently, we obtain

$$\sum_{w \in \mathcal{L}_n(\mathbf{u})} \mathrm{b}(w) = \Delta^2\mathcal{C}(n) = \mathcal{P}(n+2) - \mathcal{P}(n) = \sum_{\substack{w \in \mathcal{L}_n(\mathbf{u}) \\ w = \tilde{w}}} (\#\mathrm{Pext}(w) - 1)\,.$$

Palindromic factors that are not BS have obviously exactly one palindromic extension. Thus, we can rewrite the previous equality

$$\sum_{w \in \mathcal{L}_n(\mathbf{u})} \mathrm{b}(w) = \sum_{\substack{w \in \mathcal{L}_n(\mathbf{u}) \\ w = \tilde{w}, w \text{ BS}}} (\#\mathrm{Pext}(w) - 1)\,. \tag{3.3}$$

Let us split the sum of bilateral orders into two parts and use Lemma 3.11

$$\begin{aligned}\sum_{w\in\mathcal{L}_n(\mathbf{u})} \mathrm{b}(w) &= \sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w\neq\tilde{w},\ w\ \mathrm{BS}}} \mathrm{b}(w) + \sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w=\tilde{w},\ w\ \mathrm{BS}}} \mathrm{b}(w)\\ &\geq \sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w\neq\tilde{w},\ w\ \mathrm{BS}}} \mathrm{b}(w) + \sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w=\tilde{w},\ w\ \mathrm{BS}}} (\#\mathrm{Pext}(w)-1). \qquad (3.4)\end{aligned}$$

This in combination with (3.3) gives $\sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w\neq\tilde{w},\ w\ \mathrm{BS}}} \mathrm{b}(w) = 0$. By Lemma 3.11, bilateral orders of such factors are non-negative, which implies $\mathrm{b}(w) = 0$ for all non-palindromic BS factors. Since the equality is reached in (3.4), we obtain $\sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w=\tilde{w},\ w\ \mathrm{BS}}} \mathrm{b}(w) = \sum_{\substack{w\in\mathcal{L}_n(\mathbf{u})\\ w=\tilde{w},\ w\ \mathrm{BS}}} (\#\mathrm{Pext}(w)-1)$. Together with Lemma 3.11, this results in $\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1$ for all palindromic BS factors. □

4. Equivalent definitions of Sturmian words

Let us stress a close link between periodicity and complexity (revealed by Morse and Hedlund [38]). On one hand, the complexity of eventually periodic words is bounded. On the other hand, if there exists $n \in \mathbb{N}$ such that $\mathcal{C}(n) \leq n$, then the complexity is bounded and the infinite word $\mathbf{u}$ is eventually periodic. In consequence, the complexity of aperiodic words satisfies $\mathcal{C}(n) \geq n+1$ for all $n \in \mathbb{N}$. *Sturmian words* are defined as infinite words with the complexity $\mathcal{C}(n) = n+1$ for all $n \in \mathbb{N}$. This condition on complexity implies many properties. Let us list some of them. If $\mathbf{u}$ is a Sturmian word, then $\mathbf{u}$ has the following properties:

- $\mathbf{u}$ is a binary word;
- $\mathbf{u}$ is aperiodic;
- the language $\mathcal{L}(\mathbf{u})$ is closed under reversal;
- the language $\mathcal{L}(\mathbf{u})$ contains infinitely many palindromes;
- the word $\mathbf{u}$ is uniformly recurrent;
- the language $\mathcal{L}(\mathbf{u})$ contains no weak bispecial factors;
- $\mathbf{u}$ is rich.

There exist many equivalent definitions of Sturmian words. The following theorem summarizes several of their well-known combinatorial characterizations.

Theorem 4.1. *Let $\mathbf{u}$ be an infinite word over the alphabet $\mathcal{A}$. The properties listed below are equivalent:*

(i) $\mathbf{u}$ *is Sturmian, i.e., $\mathcal{C}(n) = n+1$ for all n;*
(ii) $\mathbf{u}$ *is binary and contains a unique left special factor of every length;*
(iii) $\mathbf{u}$ *is binary, aperiodic and every bispecial factor is ordinary;*
(iv) *any factor of* $\mathbf{u}$ *has exactly two return words;*

(v) $\mathbf{u}$ *contains one palindrome of every even length and two palindromes of every odd length;*
(vi) $\mathbf{u}$ *is binary and every palindrome has a unique palindromic extension;*
(vii) $\mathbf{u}$ *is aperiodic and balanced;*
(viii) $\mathbf{u}$ *is aperiodic and* $\mathcal{AC}(n) = 2$ *for all* $n \in \mathbb{N},\ n \geq 1$.

The characterization by return words is due to Vuillon [49] and the one by the abelian complexity is a consequence of the works by Coven and Hedlund [20]. The equivalent definition based on the balance property comes already from Morse and Hedlund [39]. The two equivalent properties concerning palindromes have been proved by Droubay and Pirillo [22]. Notice that the sixth property can be equivalently rewritten as

$$\mathcal{P}(n) + \mathcal{P}(n+1) = 3 \quad \text{for all } n \in \mathbb{N},$$

and also as

$$\mathcal{P}(n+2) = \mathcal{P}(n) \quad \text{for all } n \in \mathbb{N}.$$

Let us recall that $\mathcal{P}(0) = 1$ since the empty word is considered to be a palindrome.

5. Generalizations of Sturmian words

We have seen that Sturmian words can be defined in many equivalent ways. As a matter of course, various generalizations to multiliteral alphabets have been suggested and studied.

5.1. Two well-known generalizations

The most studied generalizations are Arnoux-Rauzy words and words coding k-interval exchange transformation.

Arnoux-Rauzy words (or *AR words* for simplicity) are infinite words with the language closed under reversal and containing exactly one LS factor w of every length, and such that every LS factor has the same number k of left extensions, *i.e.*, $\#\mathrm{Lext}(w) = k$. Their alphabet $\mathcal{A}$ has k letters since the empty word has exactly k left extensions. AR words are aperiodic and satisfy $\mathcal{C}(n) = (k-1)n+1$ for all $n \in \mathbb{N}$. They have been defined and studied in [23], the following properties have been proved ibidem. The language of AR words contains infinitely many palindromes, they are uniformly recurrent, rich, and have only ordinary BS factors. AR words form a subclass of extensively studied *episturmian words* (see for instance [29]), defined as infinite words that have the language closed under reversal and contain at most one LS factor of every length.

Another well-known generalization of Sturmian words is provided by *words coding k-interval exchange transformation.* Let us state their definition and then explain why such words generalize Sturmian words to k-letter alphabets. Take positive numbers $\alpha_1, \ldots, \alpha_k$ such that $\sum_{i=1}^{k} \alpha_i = 1$. They define a partition of the interval $I = [0, 1)$ into k subintervals

$$I_j = \left[\sum_{i=1}^{j-1}\alpha_i, \sum_{i=1}^{j}\alpha_i\right),\ j \in \{1, 2, \ldots, k\}.$$

The *interval exchange transformation* is a bijection $T : I \to I$ given by the prescription

$$T(x) = x + c_j \quad \text{for all } x \in I_j,\ j \in \{1, 2, \ldots, k\},$$

where c_j are suitably chosen constants. Since T is a bijection, the intervals $T(I_1), T(I_2), \ldots, T(I_k)$ form a partition of I. The orders of $T(I_j)$ in the partition define a permutation $\pi : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ and this permutation π determines uniquely the constants c_j. For instance, if the permutation π is symmetric, *i.e.*, $\pi = \left(\begin{smallmatrix} 1 & 2 & \ldots & k-1 & k \\ k & k-1 & \ldots & 2 & 1 \end{smallmatrix}\right)$, then the transformation T is of the following form

$$T(x) = x + \sum_{i>j}\alpha_i - \sum_{i<j}\alpha_i \quad \text{for} \quad x \in I_j.$$

The infinite word $\mathbf{u} = u_0u_1u_2\ldots$ over $\mathcal{A} = \{a_1, \ldots, a_k\}$ associated with T is defined as

$$u_n := a_j \quad \text{if} \quad T^n(x) \in I_j$$

and is called a *word coding k-interval exchange transformation* (*k-iet word* for short).

From the point of view of combinatorics on words, an important role is played by those transformations whose orbit for an arbitrary $x \in I$ is dense in I, *i.e.*, the closure of $\{T^n(x) \mid n \in \mathbb{N}\}$ is the whole interval I. A sufficient condition for this property represents the so-called i.d.o.c. (consult [35]) and the irreducibility of the permutation π. In the sequel, let us assume that T satisfies both of these properties. The k-iet word is then uniformly recurrent, its language does not depend on the position of the starting point x, but only on the transformation T, its complexity satisfies $\mathcal{C}(n) = (k-1)n + 1$ for all $n \in \mathbb{N}$ and no BS factor is weak.

The language of the k-iet word $\mathbf{u}$ is closed under reversal if and only if the permutation π is symmetric. In such a case, the language $\mathcal{L}(\mathbf{u})$ contains infinitely many palindromes and, as shown in [7], the equality in (3.1) is attained. Hence, according to Theorem 3.8, the k-iet words are rich. It is easy to describe the infinite palindromic branches for such k-iet words. The one with the empty word as its center is obtained as the coding of the orbit $\{T^n(x) | n \in \mathbb{Z}\}$ with the starting point $x = 1/2$ and the branch with the center $a_j \in \mathcal{A}$ as the coding of the orbit with the starting point $x = \sum_{i<j}\alpha_i + \alpha_j/2$.

The k-iet words provide a generalization of Sturmian words due to the well-known connection between Sturmian and mechanical words [36].

Theorem 5.1. *Let $\mathbf{u}$ be an infinite word. Then $\mathbf{u}$ is Sturmian if and only if $\mathbf{u}$ is a 2-iet word with an irrational partition of the unit interval.*

Recently, in [45], a different generalization of Sturmian sequences is considered. It in fact corresponds to a special subclass of k-iet words given by coding a trajectory in a regular $2n$-gon.

5.2. Combinatorial generalizations

Let us write down and baptize the generalizations of properties from Theorem 4.1. We will then refer to them and study their relations. Let $\mathbf{u}$ be an infinite word over the alphabet $\mathcal{A}$. Denote $k = \#\mathcal{A}$.

(1) Property $\mathcal{C}$:
the factor complexity of $\mathbf{u}$ satisfies $\mathcal{C}(n) = (k-1)n+1$ for all $n \in \mathbb{N}$.

(2) Property $\mathcal{LR}$:
$\mathbf{u}$ contains one left special and one right special factor of every length.

(3) Property $\mathcal{BO}$:
all bispecial factors of $\mathbf{u}$ are ordinary.

(4) Property $\mathcal{R}$:
any factor of $\mathbf{u}$ has exactly k return words.

(5) Property $\mathcal{P}$:
the palindromic complexity of $\mathbf{u}$ satisfies $\mathcal{P}(n) + \mathcal{P}(n+1) = k+1$ for all $n \in \mathbb{N}$.

(6) Property $\mathcal{PE}$:
every palindrome has a unique palindromic extension in $\mathbf{u}$.

(7) Balance properties:

(a) Property $\mathcal{B}_\forall$:
$\mathbf{u}$ is aperiodic and for all $a \in \mathcal{A}$ and for all factors $w, v \in \mathcal{L}(\mathbf{u})$ with $|w| = |v|$ it holds

$$||w|_a - |v|_a| \leq k-1.$$

(b) Property $\mathcal{B}_\exists$:
$\mathbf{u}$ is aperiodic and there exists $a \in \mathcal{A}$ such that for all factors $w, v \in \mathcal{L}(\mathbf{u})$ with $|w| = |v|$ it holds

$$||w|_a - |v|_a| \leq k-1.$$

(c) Property $\mathcal{AC}$:
$\mathbf{u}$ is aperiodic and the abelian complexity of $\mathbf{u}$ satisfies $\mathcal{AC}(n) = k$ for all $n \in \mathbb{N},\ n \geq 1$.

At first, let us mention which properties are satisfied by the two generalizations of Sturmian words from Section 5.1. AR words fulfill Properties: $\mathcal{C}, \mathcal{LR}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$ and k-iet words satisfy Properties: $\mathcal{C}, \mathcal{BO}, \mathcal{R}$. If moreover the permutation defining the k-iet word is symmetric, then these words have Properties $\mathcal{P}$ and $\mathcal{PE}$. Property $\mathcal{LR}$ does not hold for k-iet words.

It follows directly from the definition that some Properties imply others. For instance, by (2.1) $\mathcal{BO}$ implies $\mathcal{C}$. They are not equivalent as shown by the following example taken from [26].

Example 5.1 ($\mathcal{C} \not\Rightarrow \mathcal{BO}$)**.** The infinite ternary word $\lim_{n\to\infty} \varphi^n(a)$, where $\varphi(a) = ab$, $\varphi(b) = cab$, $\varphi(c) = ccab$ – a recoding of the Chacon substitution – has the complexity $2n+1$ for every $n \in \mathbb{N}$, but contains infinitely many strong and weak BS factors.

In the sequel, we will show that no two of these properties are equivalent on a multiliteral alphabet.

Concerning Properties $\mathcal{B}_\forall, \mathcal{B}_\exists$ and $\mathcal{AC}$, we will not treat them but in the last section since they are very restrictive, and consequently, satisfied only by a small class of infinite words.

5.3. Property $\mathcal{LR}$

Property $\mathcal{LR}$ does not characterize AR words since it is satisfied by a larger class of words. Infinite words with the language closed under reversal and satisfying Property $\mathcal{LR}$ coincide with extensively studied aperiodic episturmian words. Nevertheless, Property $\mathcal{LR}$ may be satisfied by words whose language is not closed under reversal, as illustrated in [23] by the following example. It shows also that Property $\mathcal{LR}$ does not guarantee Properties $\mathcal{C}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$.

Example 5.2 ($\mathcal{LR} \not\Rightarrow$ language closed under reversal, $\mathcal{C}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$)**.** If we construct an infinite word $\mathbf{u}$ so that we replace b with bc in the Fibonacci word $abaababaabaabab\ldots$, the fixed point of $\varphi : a \to ab,\ b \to a$, then bc is a factor of $\mathcal{L}(\mathbf{u})$, however cb not. It is easy to see that such a word has still a unique infinite RS and a unique LS branch (the infinite word $\mathbf{u}$ itself). Consequently, Property $\mathcal{LR}$ is preserved. However, both of these infinite special branches have only two extensions, hence Property $\mathcal{C}$ (and $\mathcal{BO}$ as well) fails. The factor c has only two return words $caab$ and cab, hence Property $\mathcal{R}$ does not hold. Moreover, as $\mathbf{u}$ is uniformly recurrent and its language is not closed under reversal, it contains by Theorem 3.2 only a finite number of palindromes. Therefore, Properties $\mathcal{P}$ and $\mathcal{PE}$ are not satisfied.

On the other hand, observing k-iet words, we learn that none of Properties $\mathcal{C}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$ imply $\mathcal{LR}$. The problem to describe the class of infinite words with Property $\mathcal{LR}$ whose language is not closed under reversal requires a further study.

5.4. Property $\mathcal{R}$

Let us recall that infinite words with Property $\mathcal{R}$ are necessarily uniformly recurrent. If their language is not closed under reversal, then it cannot contain infinitely many palindromes by Theorem 3.2. Such words exist, as illustrated by the following example, therefore, Property $\mathcal{R}$ does not imply $\mathcal{P}$.

Example 5.3 ($\mathcal{R} \not\Rightarrow \mathcal{P}$)**.** The fixed point $\mathbf{u}$ of φ, where $\varphi(a) = aab$, $\varphi(b) = ac$, $\varphi(c) = a$, contains bac, but cab is not its factor. The fact that every factor of $\mathbf{u}$ has three return words is explained in [9] for a whole class of infinite words coding β-integers.

We have seen that AR words and k-iet words have both Property $\mathcal{R}$ and $\mathcal{C}$, however, as shown in [26] by the following example, Property $\mathcal{C}$ does not imply Property $\mathcal{R}$ on multiliteral alphabets.

Example 5.4 ($\mathcal{C} \not\Rightarrow \mathcal{R}$)**.** The fixed point of $\varphi : a \to ab,\ b \to cab, c \to ccab$ – the above mentioned recoding of the Chacon substitution – has the complexity $2n+1$ for every $n \in \mathbb{N}$, but contains more than three return words of certain factors (for example the factor bc has 4 return words: bca, $bcca$, $bcaba$ and $bccaba$.

The following theorems come from the paper [9] that is devoted to the study of Property $\mathcal{R}$ for infinite words on multiliteral alphabets. Let us observe once more AR words and k-iet words, these classes satisfy not only Property $\mathcal{C}$, but also Property $\mathcal{BO}$. It is thus natural to ask whether Property $\mathcal{BO}$ guarantees $\mathcal{R}$. The corollary of the following theorem will provide an answer.

Theorem 5.2. *If* $\mathbf{u}$ *is an infinite word with no weak BS factors, then* $\mathbf{u}$ *has Property* $\mathcal{R}$ *if and only if* $\mathbf{u}$ *is uniformly recurrent and satisfies* $\mathcal{C}$*.*

Let us underline, an infinite word $\mathbf{u}$ has Property $\mathcal{BO}$ if and only if it has Property $\mathcal{C}$ and contains no weak BS factors. It results in the advertised corollary.

Corollary 5.3. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word. Then*

$$\mathcal{BO} \Rightarrow \mathcal{R}.$$

If we restrict our consideration to the ternary alphabet, the implication can be reversed.

Theorem 5.4. *Let* $\mathbf{u}$ *be a ternary uniformly recurrent infinite word. Then*

$$\mathcal{BO} \Leftrightarrow \mathcal{R}.$$

As soon as the alphabet has more than three letters, Property $\mathcal{R}$ does not imply Property $\mathcal{BO}$ any more.

Example 5.5 ($\mathcal{R} \not\Rightarrow \mathcal{BO}$)**.** The uniformly recurrent infinite word $\mathbf{u} = \lim_{n\to\infty} \varphi^n(a)$, where

$$\varphi(a) = acbca,\ \varphi(b) = acbcadbdaca,\ \varphi(c) = dbcbdacadbd,\ \varphi(d) = dbcbd,$$

satisfies $\mathcal{R}$, but not $\mathcal{C}$ (since $\mathcal{C}(n)$ is even for all $n \in \mathbb{N}$) and $\mathbf{u}$ contains, of course, weak BS factors. For details consult [9].

The question whether there exists a nice characterization of words with Property $\mathcal{R}$ on alphabets with more than three letters remains open.

5.5. Property $\mathcal{P}$ and $\mathcal{PE}$

The paper [8] is focused on the study of Properties $\mathcal{P}$ and $\mathcal{PE}$. As soon as an infinite word $\mathbf{u}$ has Property $\mathcal{PE}$, then $\mathbf{u}$ has exactly one infinite palindromic branch with center a for every letter $a \in \mathcal{A}$ and one infinite palindromic branch with center ε. Therefore, $\mathbf{u}$ contains exactly $\#\mathcal{A}$ palindromes for every odd length (central factors of palindromic branches with centers $a \in \mathcal{A}$) and one palindrome for every even length (central factor of the infinite palindromic branch with center ε). Consequently, Property $\mathcal{P}$ is also satisfied by $\mathbf{u}$.

Let us recall that Property $\mathcal{P}$ may be reformulated in the following way

$$\mathcal{P}(n+2) = \mathcal{P}(n) \quad \text{for all } n \in \mathbb{N}, \tag{5.1}$$

where $\mathcal{P}(0) = 1$. We will equally use both of the forms of Property $\mathcal{P}$.

Let $\mathbf{u}$ be an infinite word satisfying $\mathcal{PE}$. The language $\mathcal{L}(\mathbf{u})$ contains infinitely many palindromes, but it need not be closed under reversal, neither recurrent nor rich as illustrated by the following example.

Example 5.6 ($\mathcal{PE} \not\Rightarrow$ language closed under reversal, $\mathcal{PE} \not\Rightarrow$ richness)**.** The infinite word $\mathbf{u}$ on the alphabet $\{a, b, c\}$ defined in the following way:

$$\mathbf{u} = caccb\,\underbrace{ccc}_{3\times}\,a\,\underbrace{cccc}_{4\times}\,b\,\underbrace{ccccc}_{5\times}\,a\,\underbrace{cccccc}_{6\times}\,b\,\underbrace{ccccccc}_{7\times}\,a\ldots$$

has three infinite palindromic branches with centers a, b and c

$$\ldots cccaccc\ldots, \quad \ldots cccbccc\ldots, \quad \ldots ccccccc\ldots$$

and one infinite palindromic branch with central factors of even length of the form $\ldots cccccccc\ldots$ The factor $accb$ occurs only once in $\mathbf{u}$, thus $\mathbf{u}$ is not recurrent and hence $\mathcal{L}(\mathbf{u})$ is not closed under reversal. Moreover, $\mathbf{u}$ is not rich since the prefix $caccbccca$ of length 9 contains only 9 palindromes:
ε, a, b, c, cc, cac, cbc, ccc and $ccbcc$.

However, if the language $\mathcal{L}(\mathbf{u})$ is closed under reversal, then it is possible to say more about the relation of Properties $\mathcal{P}$ and $\mathcal{C}$ and the richness of $\mathbf{u}$. When both $\mathcal{P}$ and $\mathcal{C}$ are satisfied, the equality in (3.1) is reached. Application of Theorem 3.8 provides us with the following corollary.

Corollary 5.5. *Let* $\mathbf{u}$ *be an infinite word whose language is closed under reversal. Then*

$$\mathcal{P} + \mathcal{C} \Rightarrow \textit{richness of } \mathbf{u}.$$

The first example shows that Property $\mathcal{P}$ itself does not guarantee richness even if the language is closed under reversal. The second one illustrates that the implication in Corollary 5.5 cannot be reversed.

Example 5.7 ($\mathcal{PE} \not\Rightarrow$ richness, $\mathcal{PE} \not\Rightarrow \mathcal{C}$)**.** A known example of an infinite word with the language closed under reversal and with a higher factor complexity is the billiard sequence on three letters, for which $\mathcal{C}(n) = n^2 + n + 1$. As shown in [14], such words satisfy Property $\mathcal{PE}$, hence $\mathcal{P}$ as well. Consequently, billiard sequences do not reach the upper bound in (3.1) and by Theorem 3.8 cannot be rich.

Example 5.8 (richness $\not\Rightarrow \mathcal{P}$, richness $\not\Rightarrow \mathcal{C}$)**.** Let φ be defined on an m-letter alphabet as follows:

$$\varphi(0) = 0^t1, \quad \varphi(1) = 0^t2, \quad \ldots, \varphi(m-2) = 0^t(m-1), \quad \varphi(m-1) = 0^s,$$

where $s, t \in \mathbb{N}$ and $t \geq s \geq 2$. The fixed point $\mathbf{u}$ of φ satisfies the equality $\mathcal{P}(n+1) + \mathcal{P}(n) = \Delta\mathcal{C}(n) + 2$ for all n. As the language is closed under reversal, by Theorem 3.8 $\mathbf{u}$ is rich. Property $\mathcal{P}$ is not satisfied since the sum $\mathcal{P}(n+1)+\mathcal{P}(n)$ is not constant. Further properties of palindromes in $\mathbf{u}$ can be found in [4].

Let us examine in the sequel the connection between Properties $\mathcal{C}$ and $\mathcal{P}$, resp. $\mathcal{C}$ and $\mathcal{PE}$.

5.5.1. *Ternary alphabet*

Let us limit our considerations to the ternary alphabet. The following theorem and examples come from [8].

Theorem 5.6. *Let* $\mathbf{u}$ *be an infinite ternary word with the language closed under reversal. Then*

(1) $\mathcal{C} \Rightarrow \mathcal{P}$*;*
(2) $\mathcal{BO} \Rightarrow \mathcal{PE}$*.*

The implication in Theorem 5.6 cannot be reversed. We have already illustrated in Example 5.7 that even the stronger property $\mathcal{PE}$ does not ensure $\mathcal{C}$. Let us provide one more counterexample – a fixed point of a substitution.

Example 5.9 ($\mathcal{PE} \not\Rightarrow \mathcal{C}$)**.** Denote by $\mathbf{u}$ the infinite ternary word being the fixed point of the substitution Φ defined by

$$\Phi(a) = aba, \quad \Phi(b) = cac, \quad \Phi(c) = aca. \tag{5.2}$$

Then the language of $\mathbf{u}$ is closed under reversal. On one hand, $\mathbf{u}$ has Property $\mathcal{PE}$, consequently, $\mathbf{u}$ has Property $\mathcal{P}$, too. On the other hand, Property $\mathcal{C}$ fails and $\mathcal{L}(\mathbf{u})$ contains infinitely many weak BS factors.

Properties $\mathcal{P}$ and $\mathcal{PE}$ are equivalent for binary words. However already for ternary words, the implication $\mathcal{P} \Rightarrow \mathcal{PE}$ does not hold any more.

Example 5.10 ($\mathcal{P} \not\Rightarrow \mathcal{PE}$)**.** Let $\mathbf{v}$ be the ternary infinite word defined by $\mathbf{v} = \Psi(\mathbf{u})$, where $\Psi : \{A, B\}^* \to \{a, b, c\}^*$ is the morphism given by

$$\Psi(A) = bc \quad \text{and} \quad \Psi(B) = baa,$$

and $\mathbf{u}$ is the fixed point of the substitution φ defined by

$$\varphi(A) = ABBABBA, \quad \varphi(B) = ABA.$$

Then $\mathbf{v}$ satisfies $\mathcal{P}$, but does not satisfy $\mathcal{PE}$.

The relation between $\mathcal{R}$ and $\mathcal{P}$ follows from Theorems 5.6 and 5.4.

Corollary 5.7. *Let* $\mathbf{u}$ *be an infinite ternary word with the language closed under reversal. Then*

$$\mathcal{R} \Rightarrow \mathcal{PE}.$$

The implication cannot be reversed.

Example 5.11 ($\mathcal{PE} \not\Rightarrow \mathcal{R}$)**.** Consider the fixed point $\mathbf{u}$ of the substitution in (5.2). As mentioned above, $\mathbf{u}$ contains weak BS factors. Then by Theorem 5.4, $\mathbf{u}$ does not satisfy $\mathcal{R}$.

Putting together Theorems 5.6 and Corollary 5.5, we obtain one more corollary.

Corollary 5.8. *Let* $\mathbf{u}$ *be an infinite ternary word with the language closed under reversal. Then*

$$\mathcal{C} \Rightarrow \textit{richness of } \mathbf{u}.$$

In contrast with Corollary 5.5, we see that on a ternary alphabet already Property $\mathcal{C}$ itself ensures richness.

Neither in this case, the reversed implication holds. Consult Example 5.8 or the following example with a periodic word.

Example 5.12 (richness $\not\Rightarrow \mathcal{C}$)**.** The periodic infinite word $(abcba)^\omega$ is rich (since complete return words of palindromic factors are palindromes) and has a bounded complexity.

5.5.2. *Multiliteral alphabet*

In this section, two new theorems concerning Properties $\mathcal{P}$ and $\mathcal{PE}$ for multiliteral infinite words will be proved.

Theorem 5.9. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal.*

$$\textit{Assume } \mathcal{C}: \quad \mathcal{PE} \Leftrightarrow \mathcal{BO}.$$

Proof. ($\Leftarrow$): Let us prove the statement by contradiction. Assume that Property $\mathcal{BO}$ holds and Property $\mathcal{PE}$ does not. It is clear that the property $\mathcal{PE}$ can only be violated on a palindromic BS factor. By Property $\mathcal{BO}$, all palindromic factors have their bilateral order equal to zero. By Proposition 3.9, they have an odd number of palindromic extensions, particularly at least one.

Since the language is closed under reversal, Theorem 3.3 implies the inequality (3.1) for all $n \in \mathbb{N}$

$$\mathcal{P}(n) + \mathcal{P}(n+1) \leq 2 + \Delta\mathcal{C}(n).$$

Let w denote the shortest palindromic BS factor that does not have exactly one palindromic extension. Denote $N = |w|$. Then we have for all $n \leq N$,

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \#\mathcal{A} + 1.$$

Since Property $\mathcal{BO}$ implies Property $\mathcal{C}$, we have $2 + \Delta\mathcal{C}(n) = 2 + (\#\mathcal{A} - 1)$, hence the equality in (3.1) is attained for all $n \leq N$.

Since w has to have at least 3 palindromic extensions, one can see that $\mathcal{P}(N+2) \geq \mathcal{P}(N) + 2$. Thus, we obtain $\mathcal{P}(N+1) + \mathcal{P}(N+2) \geq \mathcal{P}(N+1) + \mathcal{P}(N) + 2 = \#\mathcal{A} + 3 = \Delta\mathcal{C}(N+1) + 4$, which is a contradiction with (3.1). We conclude that Property $\mathcal{PE}$ holds.

($\Rightarrow$): Assume Property $\mathcal{PE}$ holds. Then Property $\mathcal{P}$ holds as well. By Corollary 5.5 $\mathbf{u}$ is rich. Consequently, we can apply Theorem 3.10 and we obtain $\mathrm{b}(w) = 0$ for all non-palindromic BS factors and $\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1$ for all palindromic BS factors. By Property $\mathcal{PE}$ every palindromic BS factor has a unique palindromic extension, thus $\mathrm{b}(w) = 0$ for palindromic BS factors, too. □

Let us deduce several corollaries of Theorem 5.9. The most straightforward concerns richness and Property $\mathcal{BO}$. It follows combining Theorems 5.9 and 3.8.

Corollary 5.10. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. Then*

$$\mathcal{BO} \Rightarrow \textit{richness of } \mathbf{u}.$$

Putting together Theorems 3.2, 5.2 and 5.9, we obtain the following corollaries.

Corollary 5.11. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word.*

$$\textit{Assume } \mathcal{C}: \quad \mathcal{PE} \Rightarrow \mathcal{R}.$$

The reversed implication does not hold. Property $\mathcal{R}$ does not even guarantee the weaker property $\mathcal{P}$.

Example 5.13 ($\mathcal{R} + \mathcal{C} \not\Rightarrow \mathcal{P}$)**.** Consider again the infinite word from the previous section: the fixed point $\mathbf{u}$ of φ, where $\varphi(a) = aab$, $\varphi(b) = ac$, $\varphi(c) = a$. Properties $\mathcal{C}$ and $\mathcal{R}$ are satisfied (as explained in [9]), $\mathbf{u}$ is uniformly recurrent and the language $\mathcal{L}(\mathbf{u})$ is not closed under reversal. By Theorem 3.2, $\mathbf{u}$ contains only a finite number of palindromes.

Notice that the assumptions in Corollary 5.11 imply that the language $\mathcal{L}(\mathbf{u})$ is closed under reversal. It is natural to ask whether the implication $\mathcal{R} \Rightarrow \mathcal{PE}$ holds for infinite words with the language closed under reversal. The answer is however negative. Property $\mathcal{R}$ does not imply even the weaker property $\mathcal{P}$.

Example 5.14 ($\mathcal{R}$ + language closed under reversal $\not\Rightarrow \mathcal{P}$)**.** Consider again the uniformly recurrent infinite word from [9] given by $\mathbf{u} = \lim_{n\to\infty} \varphi^n(a)$, where

$$\varphi(a) = acbca,\ \varphi(b) = acbcadbdaca,\ \varphi(c) = dbcbdacadbd,\ \varphi(d) = dbcbd.$$

It satisfies $\mathcal{R}$, but $\mathcal{C}$ and $\mathcal{BO}$ are violated. It is not difficult to find infinitely many palindromes among weak BS factors. Thus, the language $\mathcal{L}(\mathbf{u})$ is closed under reversal. However $\mathcal{PE}$ is not satisfied because $cbc, dbd \in \mathcal{L}(\mathbf{u})$. Nor $\mathcal{P}$ holds since $\mathcal{P}(1) + \mathcal{P}(2) = 4 \neq 5$.

We notice in the previous examples that to demand either only the language closed under reversal or only Property $\mathcal{C}$ in order to reverse the implication in Corollary 5.11 is not sufficient. It is however not solved whether any infinite word with the language closed under reversal and having Properties $\mathcal{C}$ and $\mathcal{R}$ satisfies Property $\mathcal{PE}$ or at least $\mathcal{P}$ as well.

Corollary 5.12. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word.*

$$\textit{Assume } \mathcal{PE}\textit{:} \quad \textit{richness of } \mathbf{u} \Leftrightarrow \mathcal{R}.$$

Proof. Recall that by Theorem 3.2, the language is closed under reversal.
($\Rightarrow$): Suppose $\mathbf{u}$ is rich. Then Property $\mathcal{PE}$ guarantees that Property $\mathcal{P}$ holds as well. Since the language is closed under reversal, Property $\mathcal{P}$ together with Theorem 3.8 implies $\mathcal{C}$ is also satisfied. The statement follows then by Corollary 5.11.
($\Leftarrow$): Let us prove the second implication by contradiction. Assume $\mathcal{R}$ is satisfied and $\mathbf{u}$ is not rich. Theorem 3.6 claims that there exists a palindrome w which has a complete return word that is not a palindrome itself. As $\mathcal{PE}$ holds, the language has $\#\mathcal{A}+1$ biinfinite palindromic branches. As w is a palindrome, we can find it in the middle of one branch. Since $\mathbf{u}$ is uniformly recurrent, we can find w in a bounded distance from the center (on both sides) of the remaining $\#\mathcal{A}$ branches. Thus we have $\#\mathcal{A}$ distinct palindromic complete return words of w. As w was supposed to have a non-palindromic complete return word, we have a contradiction with $\mathcal{R}$. □

In Theorem 5.9 for infinite words having Property $\mathcal{C}$, we have proved that Property $\mathcal{PE}$ coincides with Property $\mathcal{BO}$. Under the same assumption on the complexity, we are again able to characterize Property $\mathcal{P}$ imposing this time a weaker condition on bilateral orders of BS factors.

Theorem 5.13. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal and satisfying Property* $\mathcal{C}$*. Then Property* $\mathcal{P}$ *holds if and only if any bispecial factor* w *of* $\mathbf{u}$ *satisfies:*

- *if* w *is non-palindromic, then*
$$\mathrm{b}(w) = 0;$$
- *if* w *is a palindrome, then*
$$\mathrm{b}(w) = \#\mathrm{Pext}(w) - 1.$$

Proof. ($\Leftarrow$): Theorem 3.10 implies that $\mathbf{u}$ is rich. Since the language is closed under reversal, we can use Theorem 3.8. By Property $\mathcal{C}$, we have $\mathcal{P}(n+1)+\mathcal{P}(n) = \Delta\mathcal{C}(n)+2 = \#\mathcal{A}+1$, thus Property $\mathcal{P}$ holds.

($\Rightarrow$): Corollary 5.5 states that $\mathbf{u}$ is rich. The statement about bilateral orders follows then by Theorem 3.10. □

This theorem may be immediately reformulated using Theorem 3.10.

Corollary 5.14. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal.*

$$\textit{Assume } \mathcal{C}\textit{:} \quad \mathcal{P} \Leftrightarrow \textit{richness of } \mathbf{u}.$$

Non-palindromic bispecial factors can really occur in infinite words with the language closed under reversal and satisfying Properties $\mathcal{C}$ and $\mathcal{PE}$, thus $\mathcal{P}$ as well. This means that there exist rich words with non-palindromic BS factors.

Example 5.15. A ternary word with such properties is $\mathbf{v} = \pi(\mathbf{u})$, where $\mathbf{u} = \varphi^2(\mathbf{u})$ and

$$\varphi : A \to CAC,\ B \to CACBD,\ C \to BDBCA,\ D \to BDB,$$

$$\pi : A \to ba,\ B \to b,\ C \to a,\ D \to abc.$$

The substitution φ satisfies for any letter $x \in \{A, B, C, D\}$, if we cut off the last two letters of $\varphi^{2n}(x)$, we get a palindrome. Together with the uniform recurrence of $\mathbf{u}$, Theorem 3.2 implies that the language $\mathcal{L}(\mathbf{u})$ is closed under reversal. Every LS factor of $\mathbf{u}$ is a prefix of $\varphi^{2n}(B)$ or $\varphi^{2n}(C)$ for some $n \in \mathbb{N}$, consequently, $\Delta\mathcal{C}(n) = 2$ for all $n \in \mathbb{N},\ n \geq 1$.

For every non-empty palindrome $w \in \mathcal{L}(\mathbf{u})$ (except for B and C), its morphic image $\pi(w)$ without first two letters is a palindrome. As $\mathbf{v}$ contains infinitely many distinct palindromes and is a morphic image of a uniformly recurrent word, thus uniformly recurrent, too, the language $\mathcal{L}(\mathbf{v})$ is closed under reversal. The word $\mathbf{v}$ has two infinite LS branches: every LS factor of $\mathbf{v}$ is either a prefix of $\pi(\varphi^{2n}(B))$ or of $\pi(\varphi^{2n}(C))$. Therefore, $\mathbf{v}$ satisfies Property $\mathcal{C}$. Moreover, $\mathbf{v}$ contains only ordinary BS factors. Applying Theorem 5.9, Property $\mathcal{PE}$ holds as well. Remark that the factor ba is a non-palindromic BS factor of $\mathbf{v}$.

5.6. Balance properties

It is a direct consequence of the definition that

$$\mathcal{AC} \ \Rightarrow\ \mathcal{B}_\forall \ \Rightarrow\ \mathcal{B}_\exists. \tag{5.3}$$

The first implication follows from the fact that if there are two factors v, w of the same length that contain a distinct number of letters a, say l and r, then there exist factors containing any number of letters a between l and r (they may be found in any factor having v as its prefix and w as its suffix, or *vice versa*).

Let us point out that our favorite generalizations of Sturmian words, namely AR words and k-iet words, violate the property $\mathcal{B}_\forall$. The paper [19] provides a construction of an AR word $\mathbf{u}$ that is not c-balanced for any c. The same property have also all 3-iet words given by the transformation T associated with the symmetric permutation and verifying the property i.d.o.c., which can be shown using methods from [1].

It is natural to ask whether infinite words on multiliteral alphabets with Property $\mathcal{AC}$ exist. A recent answer has been provided in [21]: there are no infinite words satisfying $\mathcal{AC}$ on alphabets containing more than 3 letters. On the other hand, there exist ternary infinite words with Property $\mathcal{AC}$ as shown by the example taken from [42].

Example 5.16. Let $\mathbf{v}$ be any aperiodic infinite word on $\{A, B\}$ and put $\mathbf{u} = \pi(\mathbf{v})$, where π is the morphism defined by $\pi(A) = abc$, $\pi(B) = acb$. Then $\mathcal{AC}(n) = 3$ for all $n \in \mathbb{N}$, $n \geq 1$.

A more general theorem has been proved ibidem.

Theorem 5.15. *If an aperiodic uniformly recurrent infinite word* $\mathbf{u}$ *on a ternary alphabet is 1-balanced, then* $\mathbf{u}$ *has Property* $\mathcal{AC}$*.*

Let us underline in the following examples that none of the implications in (5.3) can be reversed. The first example comes from [43] and the second one is taken from [47].

Example 5.17 ($\mathcal{B}_\forall \not\Rightarrow \mathcal{AC}$)**.** The ternary Tribonacci word – the fixed point of the substitution $\varphi : a \to ab,\ b \to ac,\ c \to a$ – is 2-balanced, however its abelian complexity reaches five values: $3, 4, 5, 6, 7$. Notice that the Tribonacci word belongs to AR words, which satisfy Properties $\mathcal{C}, \mathcal{LR}, \mathcal{BO}, \mathcal{R}, \mathcal{P}, \mathcal{PE}$.

Example 5.18 ($\mathcal{B}_\exists \not\Rightarrow \mathcal{B}_\forall$)**.** The fixed point $\mathbf{u}$ of the substitution $\varphi : a \to aab,\ b \to c,\ c \to ab$ has the following properties (shown in [47]):

- for any factors $v, w \in \mathcal{L}(\mathbf{u})$ with $|v| = |w|$, it holds

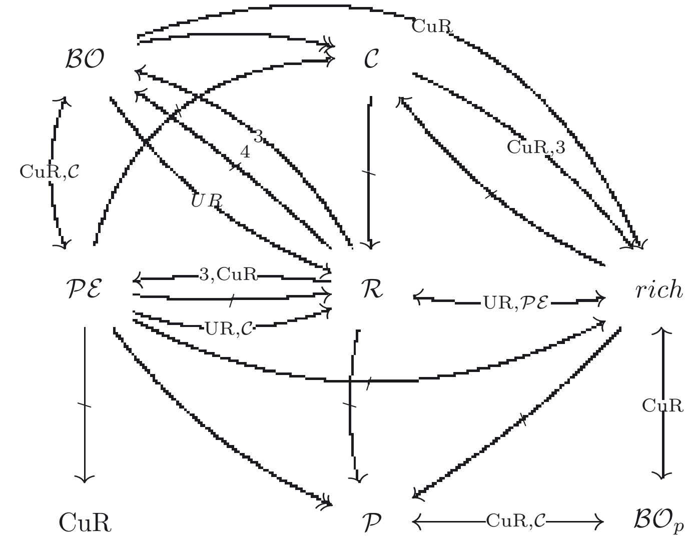
$$||v|_x - |w|_x| \leq 2 \quad \text{if } x \in \{b, c\};$$

- there exist $v, w \in \mathcal{L}(\mathbf{u})$ with $|v| = |w|$ such that

$$||v|_a - |w|_a| = 3.$$

Thus, $\mathbf{u}$ has Property $\mathcal{B}_\exists$. The word $\mathbf{u}$ is a coding of distances between neighboring β-integers, where β is the largest root of the polynomial $x^3 - 2x^2 - x + 1$. The word $\mathbf{u}$ is moreover known (see [28]) to verify Property $\mathcal{BO}$, but not $\mathcal{LR}$. Theorem 5.2 implies that $\mathbf{u}$ has Property $\mathcal{R}$ as well. Its language is not closed under reversal, consequently, neither $\mathcal{PE}$ nor $\mathcal{P}$ holds.

Generally, it is difficult to decide whether an infinite word has Property $\mathcal{B}_\exists$ or $\mathcal{B}_\forall$. A slightly simpler problem is to study infinite words that are c-balanced for some c. The criterion for existence of such a constant c for fixed points of a primitive substitution has been provided in [2], observing the spectra of adjacence matrices of substitutions. In general, it is however impossible to determine the minimal value of c from the spectrum. To our knowledge, besides the ternary words considered in Examples 5.17 and 5.18, the only non-sturmian fixed points of primitive substitutions, for which the minimal value of c is known, have been examined in [10] and [48].



$\longrightarrow$	implication
$\twoheadrightarrow$	irreversible implication
$\longleftrightarrow$	equivalence
$\not\longrightarrow$	invalid implication
number	$\#\mathcal{A}$
UR	uniform recurrence
CuR	language closed under reversal
$\mathcal{BO}_p$	$\mathrm{b}(w) = \begin{cases} \#\mathrm{Pext}(w) - 1 & \text{if } w \text{ is bispecial palindromic} \\ 0 & \text{otherwise} \end{cases}$

FIGURE 3. Diagram of known relations (assumptions are marked as labels of arrows).

6. Overview of relations and examples

In this section we provide a brief overview of relations and examples presented in the paper. Most of the relations are depicted in Figure 3. Examples are listed in Table 1. The word is either a fixed point of the given substitution, the image by the morphism π of a fixed point of the substitution φ, the limit of the sequence (u_n) or otherwise specified.

Table 1. Example overview.

word	properties	reference
$u_0 = ab$, $u_{n+1} = u_n ab\widetilde{u_n}$	uniformly recurrent, closed under reversal, finite number of palindromes	ex. 3.1 on p. 449, [13]
$u_0 = \varepsilon$, $u_{n+1} = u_n abc^{n+1} u_n$	recurrent, ∞-many palindromes, not closed under reversal	ex. 3.2 on p. 449, [16]
$a \to ab$, $b \to cab$, $c \to ccab$	$\mathcal{C}$, not $\mathcal{BO}$, not $\mathcal{R}$	ex. 5.1 on p. 457, ex. 5.4 on p. 459, [26]
φ: $A \to AB$, $B \to A$; π: $A \to a$, $B \to bc$	$\mathcal{LR}$, not closed under reversal, finite number of palindromes, not $\mathcal{C}$, not $\mathcal{R}$	ex. 5.2 on p. 458
$a \to aab$, $b \to ac$, $c \to a$	$\mathcal{R}$, not closed under reversal	ex. 5.3 on p. 458, [9]
$a \to acbca$, $b \to acbcadbdaca$, $c \to dbcbdacadbd$, $d \to dbcbd$	$\mathcal{R}$, closed under reversal, not $\mathcal{C}$, not $\mathcal{P}$	ex. 5.5 on p. 459, ex. 5.14 on p. 463, [9]
$\mathbf{u} = ca\underbrace{cc}_{2\times}b\underbrace{ccc}_{3\times}a\underbrace{cccc}_{4\times}b\underbrace{ccccc}_{5\times}a\dots$	∞-many palindromes, not closed under reversal, not rich	ex. 5.6 on p. 460
billiard sequence on three letters	closed under reversal, $\mathcal{PE}$, not $\mathcal{C}$, not rich	ex. 5.7 on p. 460, [14]
$a \to aab$, $b \to aac$, $c \to aa$	rich, not $\mathcal{C}$, not $\mathcal{P}$	ex. 5.8 on p. 461, [4]
$a \to aba$, $b \to cac$, $c \to aca$	closed under reversal, $\mathcal{PE}$, not $\mathcal{C}$, not $\mathcal{R}$	ex. 5.9 on p. 461, ex. 5.11 on p. 462, [8]
φ: $A \to ABBABBA$, $B \to ABA$; π: $A \to bc$, $B \to baa$	closed under reversal, $\mathcal{C}$,$\mathcal{P}$, not $\mathcal{PE}$	ex. 5.10 on p. 461, [8]
$(abcba)^\omega$	rich, not $\mathcal{C}$	ex. 5.12 on p. 462
$a \to aab$, $b \to ac$, $c \to a$	$\mathcal{C}$, $\mathcal{R}$, not closed under reversal	ex. 5.13 on p. 463, [9]
φ: $A \to CAC$, $B \to CACBD$, $C \to BDBCA$, $D \to BDB$; π: $A \to ba$, $B \to b$, $C \to a$, $D \to abc$	$\mathcal{PE}$, $\mathcal{C}$, closed under reversal, rich, contains non-palindromic BS factors	ex. 5.15 on p. 465
$\mathbf{u} = \pi(\mathbf{v})$, π: $A \to abc$, $B \to acb$, $\mathbf{v}$ is an aperiodic word over $\{A, B\}$	$\mathcal{AC}$	ex. 5.16 on p. 466, [42]
$a \to ab$, $b \to ac$, $c \to a$	$\mathcal{LR}, \mathcal{BO}, \mathcal{R}, \mathcal{PE}$, $\mathcal{B}_\forall$, not $\mathcal{AC}$	ex. 5.17 on p. 466, [43]
$a \to aab$, $b \to c$, $c \to ab$	$\mathcal{B}_\exists$, not $\mathcal{B}_\forall$, not closed under reversal, $\mathcal{BO}$, not $\mathcal{LR}$, $\mathcal{R}$	ex. 5.18 on p. 466, [47]

Acknowledgements. We would like to thank Pierre Arnoux for his valuable remarks and careful reviewing. We acknowledge financial support by the Czech Science Foundation grant 201/09/0584 and by the grants MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

[1] B. Adamczewski, Codages de rotations et phénomènes d'autosimilarité. *J. Théor. Nombres Bordeaux* **14** (2002) 351–386.

[2] B. Adamczewski, Balances for fixed points of primitive substitutions. *Theoret. Comput. Sci.* **307** (2003) 47–75.

[3] J.P. Allouche, M. Baake, J. Cassaigne and D. Damanik, Palindrome complexity. *Theoret. Comput. Sci.* **292** (2003) 9–31.

[4] P. Ambrož, Ch. Frougny, Z. Masáková and E. Pelantová, Palindromic complexity of infinite words associated with simple Parry numbers. *Ann. Inst. Fourier* **56** (2006) 2131–2160.

[5] P. Arnoux, C. Mauduit, I. Shiokawa and J.-I. Tamura, Complexity of sequences defined by billiards in the cube. *Bull. Soc. Math. France* **122** (1994) 1–12.

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

[7] P. Baláži, Z. Masáková and E. Pelantová, Factor versus palindromic complexity of uniformly recurrent infinite words. *Theoret. Comput. Sci.* **380** (2007) 266–275.

[8] L'. Balková, E. Pelantová and Å . Starosta, Palindromes in infinite ternary words. *RAIRO-Theor. Inf. Appl.* **43** (2009) 687–702.

[9] L'. Balková, E. Pelantová and W. Steiner, Sequences with constant number of return words. *Monatsh. Math.* **155** (2008) 251–263.

[10] L'. Balková, E. Pelantová and O. Turek, Combinatorial and arithmetical properties of infinite words associated with quadratic non-simple Parry numbers. *RAIRO-Theor. Inf. Appl.* **41** (2007) 307–328.

[11] Y. Baryshnikov, Complexity of trajectories in rectangular billiards. *Commun. Math. Phys.* **174** (1995) 43–56.

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

[13] J. Berstel, L. Boasson, O. Carton and I. Fagnot, Infinite words without palindromes. `arXiv:0903.2382` (2009), in Proc. CoRR 2009.

[14] J.P. Borel, Complexity and palindromic complexity of billiards words, in *Proceedings of WORDS 2005*, edited by S. Brlek, C. Reutenauer (2005) 175–183.

[15] S. Brlek, S. Hamel, M. Nivat and C. Reutenauer, On the palindromic complexity of infinite words. *Int. J. Found. Comput. Sci.* **2** (2004) 293–306.

[16] M. Bucci, A. De Luca, A. Glen and L.Q. Zamboni, A connection between palindromic and factor complexity using return words. *Adv. Appl. Math.* **42** (2009) 60–74.

[17] M. Bucci, A. De Luca, A. Glen and L.Q. Zamboni, A new characteristic property of rich words. *Theoret. Comput. Sci.* **410** (2009) 2860–2863.

[18] J. Cassaigne, Complexity and special factors. *Bull. Belg. Math. Soc. Simon Stevin 4* **1** (1997) 67–88.

[19] J. Cassaigne, S. Ferenczi and L.Q. Zamboni, Imbalances in Arnoux-Rauzy sequences. *Ann. Inst. Fourier* **50** (2000) 1265–1276.

[20] E.M. Coven and G.A. Hedlund, Sequences with minimal block growth. *Math. Syst. Theor.* **7** (1973) 138–153.

[21] J. Currie and N. Rampersad, Recurrent words with constant Abelian complexity. *Adv. Appl. Math.* (2010) DOI: 10.1016/j.aam.2010.05.001

[22] X. Droubay and G. Pirillo, Palindromes and Sturmian words. *Theoret. Comput. Sci.* **223** (1999) 73–85.

[23] X. Droubay, J. Justin and G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy. *Theoret. Comput. Sci.* **255** (2001) 539–553.

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

[25] I. Fagnot and L. Vuillon, Generalized balances in Sturmian words. *Discrete Appl. Math.* **121** (2002) 83–101.

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

[27] S. Ferenczi and L. Zamboni, Languages of k-interval exchange transformations. *Bull. Lond. Math. Soc.* **40** (2008) 705–714.

[28] C. Frougny, Z. Masáková and E. Pelantová, Complexity of infinite words associated with beta-expansions. *RAIRO-Theor. Inf. Appl.* **38** (2004) 162–184.

[29] A. Glen and J. Justin, Episturmian words: a survey. *RAIRO-Theor. Inf. Appl.* **43** (2009) 403–442.

[30] A. Glen, J. Justin, S. Widmer and L.Q. Zamboni, Palindromic richness. *Eur. J. Comb.* **30** (2009) 510–531.

[31] A. Hof, O. Knill and B. Simon, Singular continuous spectrum for palindromic Schröodinger operators. *Commun. Math. Phys.* **174** (1995) 149–159.

[32] C. Holton and L.Q. Zamboni, Geometric realizations of substitutions. *Bull. Soc. Math. France* **126** (1998) 149–179.

[33] J. Justin and G. Pirillo, Episturmian words and episturmian morphisms. *Theoret. Comput. Sci.* **276** (2002) 281–313.

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

[35] M.S. Keane, Interval exchange transformations. *Math. Z.* **141** (1975) 25–31.

[36] M. Lothaire, *Algebraic combinatorics on words*. Encyclopedia of Mathematics and its Applications, **90**, Cambridge University Press (2002).

[37] Z. Masáková, E. Pelantová, Relation between powers of factors and the recurrence function characterizing Sturmian words. *Theoret. Comput. Sci.* **410** (2009) 3589–3596.

[38] M. Morse and G.A. Hedlund, Symbolic dynamics. *Amer. J. Math.* **60** (1938) 815–866.

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

[40] G. Rauzy, Échanges d'intervalles et transformations induites. *Acta Arith.* **34** (1979) 315–328.

[41] G. Richomme, Another characterization of Sturmian words (one more). *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* **67** (1999) 173–175.

[42] G. Richomme, K. Saari and L.Q. Zamboni, Abelian complexity of minimal subshifts. *J. London Math. Soc.* (2010) DOI: 10.1112/jlms/jdq063

[43] G. Richomme, K. Saari and L.Q. Zamboni, Balance and abelian complexity of the Tribonacci word. *Adv. Appl. Math.* **45** (2010) 212–231.

[44] G. Rote, Sequences with subword complexity $2n$. *J. Number Theory* **46** (1993) 196–213.

[45] J. Smillie and C. Ulcigrai, Beyond Sturmian sequences: coding linear trajectories in the regular octagon. *Proc. London Math. Soc.* (2010) DOI: 10.1112/plms/pdq018

[46] S. Tabachnikov, *Billiards*. Panoramas et synthèse, SMF, Numéro 1 (1995).

[47] O. Turek, Balances and Abelian complexity of a certain class of infinite ternary words. *RAIRO-Theor. Inf. Appl.* **44** (2010) 317–341.

[48] O. Turek, Balance properties of the fixed point of the substitution associated to quadratic simple Pisot numbers. *RAIRO-Theor. Inf. Appl.* **41** (2007) 123–135.

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

[50] L. Vuillon, Balanced words. *Bull. Belg. Math. Soc. Simon Stevin* **10** (2003) 787–805.

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

Communicated by J. Berstel.
Received March 14, 2010. Accepted September 28, 2010.

Infinite Words with Finite Defect

Authors:	**Ľubomíra Balková, Edita Pelantová, Štěpán Starosta**
Journal:	Advances in Applied Mathematics
Year:	2011
Volume:	47
Pages:	562–574

Advances in Applied Mathematics 47 (2011) 562–574

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Infinite words with finite defect

L'ubomíra Balková *, Edita Pelantová, Štěpán Starosta

Department of Mathematics, FNSPE Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic

ARTICLE INFO

Article history:
Received 27 September 2010
Accepted 9 November 2010
Available online 12 January 2011

MSC:
68R15

Keywords:
Defect
Rich words
Palindrome
Palindromic complexity
Factor complexity
Morphisms of class P

ABSTRACT

In this paper, we provide a new characterization of uniformly recurrent words with finite defect based on a relation between the palindromic and factor complexity. Furthermore, we introduce a class of morphisms P_{ret} closed under composition and we show that a uniformly recurrent word with finite defect is an image of a rich (also called full) word under a morphism of class P_{ret}. This class is closely related to the well-known class P defined by Hof, Knill, and Simon; every morphism from P_{ret} is conjugate to a morphism of class P.

1. Introduction

The upper bound $|w|+1$ on the number of palindromes occurring in a finite word w given by X. Droubay, J. Justin, and G. Pirillo in [9] initiated many interesting investigations on palindromes in infinite words as well. An infinite word for which the upper bound is attained for any of its factors is called rich or full. There exist several characterizations of rich words based on the notion of complete return words [11], on the longest palindromic suffix and prefix of a factor [9,7], on the palindromic and factor complexity [6] and most recently on the bilateral orders of factors [3]. Brlek et al. suggested in [5] to study the defect of a finite word w defined as the difference between the upper bound $|w|+1$ and the actual number of palindromes contained in w. The defect of an infinite word is then defined as the maximal defect of a factor of the infinite word. In this convention, rich words are precisely the words with zero defect. In this paper we focus on uniformly recurrent words with finite

[*] Corresponding author.
E-mail addresses: lubomira.balkova@fjfi.cvut.cz (L. Balková), edita.pelantova@fjfi.cvut.cz (E. Pelantová), staroste@fjfi.cvut.cz (Š. Starosta).

0196-8858/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.aam.2010.11.006

defect. Let us point out that periodic words with finite defect have been already described in [5] and in [11]. In Section 2 we introduce notation and summarize known results on rich words and words with finite defect. In Section 3 the notion of oddities and the characterization of uniformly recurrent words with finite defect based on oddities from [11] is recalled and, as an immediate consequence, two more useful characterizations are deduced. The main result is a new characterization of uniformly recurrent words with finite defect based on a relation between the palindromic and factor complexity, see Theorem 4.1 in Section 4. Furthermore, we introduce a class of morphisms P_{ret} closed under composition of morphisms and we show that a uniformly recurrent word with finite defect is an image of a rich word under a morphism of class P_{ret}, see Theorem 5.5 in Section 5. This class is closely related to the well-known class P defined by Hof, Knill, and Simon in [12]; every morphism from P_{ret} is conjugate to a morphism of class P.

2. Preliminaries

By $\mathcal{A}$ we denote a finite set of symbols, usually called *letters*; the set $\mathcal{A}$ is therefore called an *alphabet*. A finite string $w = w_0w_1 \dots w_{n-1}$ of letters of $\mathcal{A}$ is said to be a *finite word*, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ϵ as the neutral element form a free monoid $\mathcal{A}^*$. The map

$$w = w_0w_1 \dots w_{n-1} \quad \mapsto \quad \overline{w} = w_{n-1}w_{n-2} \dots w_0$$

is a bijection on $\mathcal{A}^*$, the word $\overline{w}$ is called the *reversal* or the *mirror image* of w. A word w which coincides with its mirror image is a *palindrome*.

Under an *infinite word* we understand an infinite string $\mathbf{u} = u_0u_1u_2 \dots$ of letters from $\mathcal{A}$. A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \epsilon$, then w is said to be a *prefix* of v, if $s = \epsilon$, then w is a *suffix* of v.

The *language* $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. Factors of $\mathbf{u}$ of length n form the set denoted by $\mathcal{L}_n(\mathbf{u})$. Clearly, $\mathcal{L}(\mathbf{u}) = \bigcup_{n \in \mathbb{N}} \mathcal{L}_n(\mathbf{u})$. We say that the language $\mathcal{L}(\mathbf{u})$ is *closed under reversal* if $\mathcal{L}(\mathbf{u})$ contains with every factor w also its reversal $\overline{w}$.

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2} \dots$. Such an index is called an *occurrence* of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ has infinitely many occurrences in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be *recurrent*. It is easy to see that if the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent (a proof can be found in [11]). For a recurrent infinite word $\mathbf{u}$, we may define the notion of a *complete return word* of any $w \in \mathcal{L}(\mathbf{u})$. It is a factor $v \in \mathcal{L}(\mathbf{u})$ such that w is a prefix and a suffix of v and w occurs in v exactly twice. Under a *return word* of a factor w is usually meant a word $q \in \mathcal{L}(\mathbf{u})$ such that qw is a complete return word of w. If any factor $w \in \mathcal{L}(\mathbf{u})$ has only finitely many return words, then the infinite word $\mathbf{u}$ is called *uniformly recurrent*. If $\mathbf{u}$ is a uniformly recurrent word, we can assign to any $n \in \mathbb{N}$ the minimal number $R_{\mathbf{u}}(n) \in \mathbb{N}$ such that we have for any $v \in \mathcal{L}(\mathbf{u})$ with $|v| \geqslant R_{\mathbf{u}}(n)$

$$\big\{w \;\big|\; |w| = n,\ w \text{ is a factor of } v\big\} = \mathcal{L}_n(\mathbf{u}),$$

or equivalently, any piece of $\mathbf{u}$ which is longer than or equal to $R_{\mathbf{u}}(n)$ contains already all factors of $\mathbf{u}$ of length n. The map $n \to R_{\mathbf{u}}(n)$ is usually called the *recurrence function* of $\mathbf{u}$. In particular, any fixed point of a primitive morphism is uniformly recurrent, where a morphism φ over an alphabet $\mathcal{A}$ is *primitive* if there exists an integer k such that for every $a \in \mathcal{A}$ the k-th iteration $\varphi^k(a)$ contains all letters of $\mathcal{A}$.

The *factor complexity* of an infinite word $\mathbf{u}$ is a map $\mathcal{C} : \mathbb{N} \mapsto \mathbb{N}$ defined by the prescription $\mathcal{C}(n) := \#\mathcal{L}_n(\mathbf{u})$. To determine the first difference of the factor complexity, one has to count the possible extensions of factors of length n. A *right extension* of $w \in \mathcal{L}(\mathbf{u})$ is any letter $a \in \mathcal{A}$ such that $wa \in \mathcal{L}(\mathbf{u})$. Of course, any factor of $\mathbf{u}$ has at least one right extension. A factor w is called *right special* if w has at least two right extensions. Similarly, one can define a *left extension* and a *left special* factor. We will deal only with recurrent infinite words $\mathbf{u}$. In this case, any factor of $\mathbf{u}$ has at least one left extension. We say that w is a *bispecial* factor if it is right and left special.

In our article we focus on words in some sense opulent in palindromes, therefore we will introduce several notions connected with palindromic factors.

The *defect* $D(w)$ of a finite word w is the difference between the utmost number of palindromes $|w|+1$ and the actual number of palindromes contained in w. Finite words with zero defects – called *rich* words – can be viewed as the most saturated by palindromes. This definition may be extended to infinite words as follows.

Definition 2.1. An infinite word $\mathbf{u} = u_0u_1u_2\ldots$ is called *rich*, if for any index $n \in \mathbb{N}$ the prefix $u_0u_1u_2\ldots u_{n-1}$ of length n contains exactly $n+1$ different palindromes.

We keep here the terminology introduced by Glen et al. in [11] in 2007, which seems to us to be prevalent nowadays. However, Brlek et al. in [5] baptized such words full already in 2004.

Let us remark that not only all prefixes of rich words are rich, but also all factors are rich. A result from [9] will provide us with a handful tool which helps to evaluate the defect of a factor.

Proposition 2.2. *(See [9].) A finite or infinite word* $\mathbf{u}$ *is rich if and only if the longest palindromic suffix of* w *occurs exactly once in* w *for any prefix* w *of* $\mathbf{u}$.

The longest palindromic suffix of a factor w will occur often in our considerations, therefore we will denote it by $lps(w)$. In accordance with the terminology introduced in [9], the factor with a unique occurrence in another factor is called *unioccurrent*. From the proof of the previous proposition directly follows the next corollary.

Corollary 2.3. *The defect* $D(w)$ *of a finite word* w *is equal to the number of prefixes* w' *of* w*, for which the longest palindromic suffix of* w' *is not unioccurrent in* w'.

This corollary implies that $D(v) \geqslant D(w)$ whenever w is a factor of v. It enables to give a reasonable definition of the defect of an infinite word (see [5]).

Definition 2.4. The defect of an infinite word $\mathbf{u}$ is the number (finite or infinite)

$$D(\mathbf{u}) = \sup\{D(w) \mid w \text{ is a prefix of } \mathbf{u}\}.$$

Let us point out several facts concerning defects that are easy to prove:

(1) If we consider all factors of a finite or an infinite word $\mathbf{u}$, we obtain the same defect, i.e.,

$$D(\mathbf{u}) = \sup\{D(w) \mid w \in \mathcal{L}(\mathbf{u})\}.$$

(2) Any infinite word with finite defect contains infinitely many palindromes.

(3) Infinite words with zero defect correspond exactly to rich words.

Periodic words with finite defect have been studied in [5] and in [11]. It holds that the defect of an infinite periodic word with the minimal period w is finite if and only if $w = pq$, where both p and q are palindromes. In [11] words with finite defect have been baptized *almost rich* and the richness of a word was described using complete return words.

Proposition 2.5. *(See [11].) An infinite word* $\mathbf{u}$ *is rich if and only if all complete return words of any palindrome are palindromes.*

The authors of [9] who were the first ones to tackle this problem showed that Sturmian and episturmian words are rich. In [5], an insight into the richness of periodic words can be found.

The number of palindromes of a fixed length occurring in an infinite word is measured by the so-called *palindromic complexity* $\mathcal{P}$, a map which assigns to any non-negative integer n the number

$$\mathcal{P}(n) := \#\{w \in \mathcal{L}_n(u) \mid w \text{ is a palindrome}\}.$$

The palindromic complexity is bounded by the first difference of factor complexity. The following proposition is proven in [2] for uniformly recurrent words, however the uniform recurrence is not needed in the proofs, thus it holds for any infinite words with language closed under reversal.

Proposition 2.6. *(See [2].) Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. Then*

$$\mathcal{P}(n) + \mathcal{P}(n+1) \leqslant \mathcal{C}(n+1) - \mathcal{C}(n) + 2, \tag{1}$$

for all $n \in \mathbb{N}$.

It is shown in [6] that this bound can be used for a characterization of rich words as well. The following proposition states this fact.

Proposition 2.7. *(See [6].) An infinite word* $\mathbf{u}$ *with language closed under reversal is rich if and only if the equality in* (1) *holds for all* $n \in \mathbb{N}$.

The most recent characterization of rich words given in [3] exploits the notion of the bilateral order $b(w)$ of a factor and the palindromic extension of a factor. The bilateral order was introduced in [8] as $\mathrm{b}(w) = \#\{awb \mid awb \in \mathcal{L}(\mathbf{u}), a, b \in \mathcal{A}\} - \#\{aw \mid aw \in \mathcal{L}(\mathbf{u}), a \in \mathcal{A}\} - \#\{wb \mid wb \in \mathcal{L}(\mathbf{u}), b \in \mathcal{A}\} + 1$. The set of palindromic extensions of a palindrome $w \in \mathcal{L}(\mathbf{u})$ is defined by $\mathrm{Pext}(w) = \{awa \mid awa \in \mathcal{L}(\mathbf{u}), a \in \mathcal{A}\}$.

Proposition 2.8. *(See [3].) An infinite word* $\mathbf{u}$ *with language closed under reversal is rich if and only if any bispecial factor* w *satisfies*:

- *if* w *is non-palindromic, then* $\mathrm{b}(w) = 0$,
- *if* w *is a palindrome, then* $\mathrm{b}(w) = \#\,\mathrm{Pext}(w) - 1$.

3. Characterizations of words with finite defect

Uniformly recurrent words with finite defect are characterized using the notion of oddities in Proposition 4.8 from [11]. It is based on the following lower bound.

Proposition 3.1. *(See [11, Proposition 4.6].) For any infinite word* $\mathbf{u}$ *it holds*

$$D(\mathbf{u}) \geqslant \#\{\{v, \overline{v}\} \mid v \neq \overline{v} \textit{ and } v \textit{ or } \overline{v} \textit{ is a complete return word in } \mathbf{u} \textit{ of a palindrome } w\}.$$

The set $\{v, \overline{v}\}$ is called an *oddity*. It is clear that for uniformly recurrent words with a finite number of distinct palindromes, the defect is infinite, however the number of oddities is finite. Moreover, even for uniformly recurrent words with infinitely many palindromes, it can hold

$$D(\mathbf{u}) > \#\{\{v, \overline{v}\} \mid v \neq \overline{v} \text{ and } v \text{ or } \overline{v} \text{ is a complete return word in } \mathbf{u} \text{ of a palindrome } w\}.$$

We take an example for this situation from [11]. Let $\mathbf{u} = (abcabcacbacb)^{\omega}$, where ω denotes an infinite repetition, then $D(\mathbf{u}) = 4$, but the number of oddities is equal to 3. However, the defect of an aperiodic word can also exceed the number of oddities. For instance, if we replace in Example 3.4 the substitution σ with $0 \to cabcabcbacbac$, $1 \to d$, then it is easy to show that $D(\mathbf{u}) = 4$, but the number of oddities is 3.

We can now recall the characterization of words with finite defect based on oddities.

Proposition 3.2. *(See [11, Proposition 4.8].) A uniformly recurrent word* $\mathbf{u}$ *has infinitely many oddities if and only if* $\mathbf{u}$ *contains infinitely many palindromes and* $D(\mathbf{u}) = \infty$.

As an immediate consequence of Proposition 3.2, we obtain the following characterizations of infinite words with finite defect.

Theorem 3.3. *Let* $\mathbf{u}$ *be a uniformly recurrent word containing infinitely many palindromes. Then the following statements are equivalent:*

1. $D(\mathbf{u}) < \infty$,
2. $\mathbf{u}$ *has a finite number of oddities,*
3. *there exists an integer* K *such that all complete return words of any palindrome from* $\mathcal{L}(\mathbf{u})$ *of length at least* K *are palindromes,*
4. *there exists an integer* H *such that for any prefix* f *of* $\mathbf{u}$ *with* $|f| \geqslant H$ *the longest palindromic suffix of* f *is unioccurrent in* f.

Proof. 1. and 2. are equivalent by Proposition 3.2. It follows directly from the definition of oddities that 2. and 3. are equivalent. Corollary 2.3 implies that 1. and 4. are equivalent. □

It is easy to see that the last statement of Theorem 3.3 can be equivalently rewritten as: There exists an integer H such that for any factor f of $\mathbf{u}$ with $|f| \geqslant H$ the longest palindromic suffix of f is unioccurrent in f.

Let us stress that if we put in the previous theorem $D(\mathbf{u}) = K = H = 0$, the points 1., 3., and 4. become known results on rich words, see Propositions 2.5 and 2.2.

Example 3.4. Let us provide an example of a uniformly recurrent word $\mathbf{u}$ with finite defect and let us find for $\mathbf{u}$ the lowest values of constants K and H from Theorem 3.3. Take the Fibonacci word $\mathbf{v}$, i.e., the fixed point of $\varphi : 0 \to 01$, $1 \to 0$. Define $\mathbf{u}$ as its morphic image $\sigma(\mathbf{v})$, where $\sigma : 0 \to cabcbac$, $1 \to d$.

It is easy to show that all palindromes of length greater than 1 and the palindromes a, b, and d have only palindromic complete return words. Hint: long palindromes in $\mathbf{u}$ contains in their center images of non-empty palindromes from $\mathbf{v}$ that have palindromic complete return words by the richness of $\mathbf{v}$. The only non-palindromic complete return of c is $cabc$, thus there is exactly one oddity $\{cabc, cbac\}$. In order to show that $D(\mathbf{u}) = 1$, it suffices to verify that no prefixes longer than $cabc$ have c as their longest palindromic suffix. This follows directly from the form of σ. The lowest values of the constants K and H are: $K = 2$, $H = 5$.

4. Palindromic complexity of words with finite defect

The aim of this section is to prove the following new characterization of infinite words with finite defect based on a relation between the palindromic and factor complexity.

Theorem 4.1. *Let* $\mathbf{u}$ *be a uniformly recurrent word. Then* $D(\mathbf{u}) < \infty$ *if and only if there exists an integer* N *such that*

$$\mathcal{P}(n) + \mathcal{P}(n+1) = \mathcal{C}(n+1) - \mathcal{C}(n) + 2$$

holds for all $n \geqslant N$.

Notice that if we set $N = 0$ in the previous theorem, then we obtain the known characterization of rich words from Proposition 2.7 (which holds even under a weaker assumption that $\mathcal{L}(\mathbf{u})$ is closed under reversal).

In the sequel, we will prove two propositions that together with the equivalent characterizations of words with finite defect from Theorem 3.3 imply Theorem 4.1. As we have already mentioned, all words with language closed under reversal satisfy the inequality in Proposition 2.6. A direct consequence of its proof given in [2] is a necessary and sufficient condition for the equality in (1). To formulate this condition in Lemma 4.2, we introduce two auxiliary notions.

Let $\mathbf{u}$ be an infinite word with language closed under reversal and let n be a given positive integer.

An *n-simple path* e is a factor of $\mathbf{u}$ of length at least $n+1$ such that the only special (right or left) factors of length n occurring in e are its prefix and suffix of length n. If w is the prefix of e of length n and v is the suffix of e of length n, we say that the n-simple path e starts in w and ends in v.

We will denote by G_n an undirected graph whose set of vertices is formed by unordered pairs $(w, \overline{w})$ such that $w \in \mathcal{L}_n(u)$ is right or left special. We connect two vertices $(w, \overline{w})$ and $(v, \bar{v})$ by an unordered pair $(e, \bar{e})$ if e or $\bar{e}$ is an n-simple path starting in w or $\overline{w}$ and ending in v or $\bar{v}$.

Note that the graph G_n may have multiple edges and loops.

Lemma 4.2. *Let $\mathbf{u}$ be an infinite word with language closed under reversal. The equality in* (1) *holds for an integer $n \in \mathbb{N}$ if and only if both of the following conditions are met*:

1. *The graph G_n after removing loops is a tree.*
2. *Any n-simple path forming a loop in the graph G_n is a palindrome.*

Proposition 4.3. *Let $\mathbf{u}$ be an infinite word with language closed under reversal. Suppose that there exists an integer N such that for all $n \geqslant N$ the equality $\mathcal{P}(n) + \mathcal{P}(n+1) = \mathcal{C}(n+1) - \mathcal{C}(n) + 2$ holds. Then the complete return words of any palindromic factor of length $n \geqslant N$ are palindromes.*

Proof. Assume the contrary: Let $p = p_1 p_2 \dots p_k$ be a palindrome with $k \geqslant N$ and let v be its complete return word which is not a palindrome. Clearly $|v| > 2|p|$. Then there exist a factor f (possibly empty) and two different letters x and y such that $v = pfxv'y\overline{f}p$.

Let us consider the graph G_n, where n is the length of the factor $w := pf$, i.e., $n \geqslant N$. Since the language of $\mathbf{u}$ is closed under reversal, the factor w is right special – the letters x and y belong to its right extensions.

If the complete return word v contains no other right or left special factors, then the non-palindromic v is an n-simple path which starts in $w = pf$ and ends in $\overline{w} = \overline{f}p$ – a contradiction with the condition 2. in Lemma 4.2.

Let v contain other left or right special factors of length n. We find the prefix of v which is an n-simple path. This simple path starts in w, its ending point is a special factor, we denote it by A. Since v is a complete return word of p, we have $A \neq w, \overline{w}$. So in the graph G_n, the vertices $(w, \overline{w})$ and $(A, \overline{A})$ are connected with an edge. Similarly, we find the suffix of v which is an n-simple path and we denote its starting point by B, its ending point is $\overline{w}$. Again, $B \neq w, \overline{w}$ and the vertices $(w, \overline{w})$ and $(B, \overline{B})$ are connected with an edge. So in G_n we have a path with two edges which connects $(A, \overline{A})$ and $(B, \overline{B})$ and the vertex $(w, \overline{w})$ is its intermediate vertex.

The special factors A and B are factors of $p_2 \dots p_k fxv'y\overline{f}p_k \dots p_2$, it means that in the graph G_n there exists a walk, and therefore a path[1] as well, between the vertices $(A, \overline{A})$ and $(B, \overline{B})$ which does not use the vertex $(w, \overline{w})$.

Finally, if $(A, \overline{A})$ and $(B, \overline{B})$ coincide, then we have in G_n a multiple edge between $(A, \overline{A})$ and $(w, \overline{w})$. If $(A, \overline{A}) \neq (B, \overline{B})$, then in G_n we have two different paths connecting $(A, \overline{A})$ and $(B, \overline{B})$. Together, G_n is not a tree after removing loops – a contradiction with the condition 1. in Lemma 4.2. □

[1] Along a walk vertices may occur with repetition, in a path any vertex appears at most once.

Lemma 4.4. *Let* $\mathbf{u}$ *be an infinite word whose language is closed under reversal. Let* $\mathbf{u}$ *have the following property: there exists an integer* H *such that for any factor* $f \in \mathcal{L}(\mathbf{u})$ *with* $|f| \geqslant H$ *the longest palindromic suffix of* f *is unioccurrent in* f. *Let* w *be a non-palindromic factor of* $\mathbf{u}$ *with* $|w| \geqslant H$ *and* v *be a palindromic factor of* $\mathbf{u}$ *with* $|v| \geqslant H$. *Then*

- *occurrences of* w *and* $\overline{w}$ *in* $\mathbf{u}$ *alternate, i.e., any complete return word of* w *contains the factor* $\overline{w}$,
- *any factor* e *of* $\mathbf{u}$ *with a prefix* w *and a suffix* $\overline{w}$, *which has no other occurrences of* w *and* $\overline{w}$, *is a palindrome,*
- *any complete return word of* v *is a palindrome.*

Proof. Consider a non-palindromic factor w such that $|w| \geqslant H$. Let f be a complete return word of w. Since $|w| \geqslant H$, its complete return word satisfies $|f| \geqslant H$. According to the assumption, $lps(f)$, the longest palindromic suffix of f, is unioccurrent in f. Its length satisfies necessarily $|lps(f)| > |w|$ – otherwise a contradiction with the unioccurrence of $lps(f)$. Clearly, the palindrome $lps(f)$ has a suffix w and thus a prefix $\overline{w}$, i.e., the complete return word f of w contains $\overline{w}$ as well. Moreover, we have proven that any factor e, which has a prefix $\overline{w}$ and a suffix w and which has no other occurrences of w and $\overline{w}$, is the longest palindromic suffix of a complete return word of w, therefore $e = lps(f)$, i.e., the factor e is a palindrome.

Consider a palindromic factor v, its complete return word f and the longest palindromic suffix of f. Since v is a palindromic suffix of f, necessarily $|lps(f)| \geqslant |v|$. As $|v| \geqslant H$, $lps(f)$ is unioccurrent in f. Hence, $|lps(f)| > |v|$. If $lps(f)$ is shorter than the whole f, then the complete return word f contains at least three occurrences of w – a contradiction. Thus, $lps(f) = f$, i.e., f is a palindrome. □

Proposition 4.5. *Let* $\mathbf{u}$ *be an infinite word whose language is closed under reversal. Let* $\mathbf{u}$ *have the following property: there exists an integer* H *such that for any factor* $f \in \mathcal{L}(\mathbf{u})$ *with* $|f| \geqslant H$ *the longest palindromic suffix of* f *is unioccurrent in* f. *Then*

$$2 + \mathcal{C}(n+1) - \mathcal{C}(n) = \mathcal{P}(n+1) + \mathcal{P}(n) \quad \textit{for any } n \geqslant H.$$

Proof. We have to show that both conditions of Lemma 4.2 are satisfied for any $n \geqslant H$.

The condition 1.: Let $(w, \overline{w})$ and $(v, \overline{v})$ be two distinct vertices in the graph G_n, where $n \geqslant H$. We say that an unordered couple $(f, \overline{f})$ is a *realization* of a path between these two vertices if

- either the factor f or the factor $\overline{f}$ has the property: w or $\overline{w}$ is its prefix and v or $\overline{v}$ is its suffix,
- there exist indices $i, \ell \in \mathbb{N}, i < \ell$ such that either the factor f or the factor $\overline{f}$ coincides with the factor $u_i u_{i+1} \dots u_\ell$ and factors w, $\overline{w}$, v, and $\overline{v}$ do not occur in $u_{i+1} \dots u_{\ell-1}$.

The number i is called an *index* of the realization $(f, \overline{f})$.

Since $\mathbf{u}$ is recurrent, there exists at least one realization for any pair of vertices $(w, \overline{w})$ and $(v, \overline{v})$ and any realization has infinitely many indices. Consider a realization $(f, \overline{f})$ and its index i. WLOG $f = u_i u_{i+1} \dots u_\ell$ and w is a prefix of f and v a suffix of f. Since $\mathbf{u}$ is recurrent, we can find the smallest index $m > \ell$ such that $u' = u_i u_{i+1} \dots u_\ell \dots u_m$ has a suffix $\overline{w}$. According to Lemma 4.4, u' is a palindrome. Therefore its suffix of length $|f|$ is exactly $\overline{f}$. This means that the index $m - |f| + 1$ is an index of the same realization of a path between $(w, \overline{w})$ and $(v, \overline{v})$. As the factor $u_{i+1} \dots u_{m-1}$ does not contain neither the factor w nor $\overline{w}$, no index j strictly between i and $m - |f| + 1$ is an index of any realization of a path between $(w, \overline{w})$ and $(v, \overline{v})$.

We have shown that between any pair of two consecutive indices of one specific realization $(f, \overline{f})$ of a path between $(w, \overline{w})$ and $(v, \overline{v})$ there does not exist any index of any other realization $(g, \overline{g})$ of a path between $(w, \overline{w})$ and $(v, \overline{v})$. This means that there exists a unique realization of a path between $(w, \overline{w})$ and $(v, \overline{v})$, which implies that in the graph G_n there exists a unique path between vertices $(w, \overline{w})$ and $(v, \overline{v})$. Since this is true for all pairs of vertices of G_n, the graph G_n after removing loops is a tree.

The condition 2.: Let $w \in \mathcal{L}(\mathbf{u})$ be a special factor (palindromic or non-palindromic) with $|w| = n \geqslant H$. An n-simple path f starting in w and ending in $\overline{w}$ contains according to its definition no other special vertex inside the path, in particular w and $\overline{w}$ do not occur inside the path. According to Lemma 4.4, the path f is a palindrome. □

Proof of Theorem 4.1. It is a direct consequence of Propositions 4.3 and 4.5 and of Theorem 3.3, where the last statement is replaced with an equivalent one: There exists an integer H such that for any factor f of $\mathbf{u}$ with $|f| \geqslant H$ the longest palindromic suffix of f is unioccurrent in f. □

5. Morphisms of class P_{ret}

In this section, we will define a new class of morphisms and we will reveal their relation with well-known morphisms of class P (defined in [12]). We will show an important role these morphisms play in the description of words with finite defect.

Definition 5.1. We say that a morphism $\varphi : \mathcal{B}^* \mapsto \mathcal{A}^*$ is of class P_{ret} if there exists a palindrome $p \in \mathcal{A}^*$ such that

- $\varphi(b)p$ is a palindrome for any $b \in \mathcal{B}$,
- $\varphi(b)p$ contains exactly 2 occurrences of p, one as a prefix and one as a suffix, for any $b \in \mathcal{B}$,
- $\varphi(b) \neq \varphi(c)$ for all $b, c \in \mathcal{B}$, $b \neq c$.

Lemma 5.2. *The following properties of the morphisms of class P_{ret} are easy to prove.*

(1) *$\varphi(w) = \varphi(v)$, where $w, v \in \mathcal{B}^*$, implies $w = v$, i.e., φ is injective,*
(2) *$\overline{\varphi(x)p} = \varphi(\bar{x})p$ for any $x \in \mathcal{B}^*$,*
(3) *$\varphi(s)p$ is a palindrome if and only if $s \in \mathcal{B}^*$ is a palindrome.*

Proof. We will give a hint for the proof of the injectivity. The other statements are immediate consequences of Definition 5.1. If $\varphi(w) = \varphi(v)$, then $\varphi(w)p = \varphi(v)p$. This implies $w = v$ by induction on $\max\{|w|, |v|\}$: the assertion is true for $\max\{|w|, |v|\} = 1$ (i.e., $|w| = |v| = 1$, since the morphism φ is not erasing from the second point of Definition 5.1) from the third point of Definition 5.1; the induction is then proven using the second point of Definition 5.1. □

Another class of morphisms closely related to defects is *standard* (*special*) *morphisms of class P* defined in [11]. We will reveal their connection with P_{ret} in Section 6.

Proposition 5.3. *The class P_{ret} is closed under the composition of morphisms, i.e., for any $\varphi, \sigma \in P_{\text{ret}}$ we have $\varphi\sigma \in P_{\text{ret}}$ (if the composition is well defined).*

Proof. Let p_φ and p_σ be the corresponding palindromes from the definition of P_{ret} of the morphisms φ and σ, respectively. Then $p_{\varphi\sigma} := \varphi(p_\sigma)p_\varphi$ is a palindrome by point (3) of Lemma 5.2 for φ. It suffices to verify that $p_{\varphi\sigma}$ plays the role of the palindrome p for the morphism $\varphi\sigma$.

- Take b a letter. We have $\overline{(\varphi\sigma)(b)p_{\varphi\sigma}} = \overline{\varphi(\sigma(b)p_\sigma)p_\varphi}$. We obtain the following equalities using firstly point (2) of Lemma 5.2 for φ and then for σ:

$$\overline{\varphi\big(\sigma(b)p_\sigma\big)p_\varphi} = \varphi\big(\overline{\sigma(b)p_\sigma}\big)p_\varphi = \varphi\big(\sigma(\bar{b})p_\sigma\big)p_\varphi = \varphi\big(\sigma(b)p_\sigma\big)p_\varphi = (\varphi\sigma)(b)p_{\varphi\sigma},$$

i.e., $(\varphi\sigma)(b)p_{\varphi\sigma}$ is a palindrome for all b.

- Since $\varphi \in P_{\text{ret}}$, there is a one-to-one correspondence between the occurrences of $p_{\varphi\sigma} = \varphi(p_\sigma)p_\varphi$ in $(\varphi\sigma)(b)p_{\varphi\sigma} = \varphi(\sigma(b)p_\sigma)p_\varphi$ and the occurrences of p_σ in $\sigma(b)p_\sigma$. As $\sigma \in P_{\text{ret}}$, the word $\sigma(b)p_\sigma$ contains p_σ only as a prefix and as a suffix. Therefore $\varphi(\sigma(b)p_\sigma)p_\varphi$ has only two occurrences of $\varphi(p_\sigma)p_\varphi$ – as a prefix and as a suffix.
- The injectivity of φ and σ clearly guarantees that $(\varphi\sigma)(b) \neq (\varphi\sigma)(c)$ for all $b \neq c$. □

In [12] another class of morphisms is defined. We say that a morphism φ is of class P if there exist a palindrome p and for every letter a a palindrome q_a such that $\varphi(a) = pq_a$. The interest of the class P has been awoken by the following question stated ibidem (however formulated in terms of dynamical systems): "Given a fixed point of a primitive morphism φ containing infinitely many palindromes, can we find a primitive morphism σ of class P such that the factors of a fixed point of σ are the same?" Let us recall that for any primitive morphism, the languages of all its fixed points are the same. The previous question has been answered affirmatively in [13] for morphisms defined on binary alphabets and in [1] for periodic fixed points.

In order to reveal the relation between the classes P and P_{ret}, we have to define the conjugation of a morphism. A morphism σ is said to be *conjugate* to a morphism φ defined on an alphabet $\mathcal{A}$ if there exists a word $w \in \mathcal{A}^*$ such that

- either for every letter $a \in \mathcal{A}$, the image $\varphi(a)$ has w as its prefix and the image $\sigma(a)$ is obtained from $\varphi(a)$ by erasing w from the beginning and adding w to the end; we write $\sigma(a) = w^{-1}\varphi(a)w$,
- or for every letter $a \in \mathcal{A}$, the image $\varphi(a)$ has w as its suffix and the image $\sigma(a)$ is obtained from $\varphi(a)$ by erasing w from the end and adding w to the beginning; we write $\sigma(a) = w\varphi(a)w^{-1}$.

Proposition 5.4. *If φ is a morphism of class P_{ret}, then φ is conjugate to a morphism of class P.*

Proof. Let $\varphi \in P_{\text{ret}}$ and let p have the same meaning as in the definition of P_{ret}. We will write $p = qx\bar{q}$, where $q \in \mathcal{A}^*$ and x is either the empty word or a letter. Denote by σ a morphism defined for all letters a as $\sigma(a) = q^{-1}\varphi(a)q$. Thus, φ is conjugate to σ.

The word $q^{-1}\varphi(a)q$ can be written as xy_a since qx is a prefix of $\varphi(a)q$. Since $\varphi(a)qx\bar{q}$ is a palindrome, $q^{-1}\varphi(a)qx\bar{q}\ \bar{q}^{-1} = xy_ax$ is a palindrome too. Therefore y_a is a palindrome and σ is of class P. □

The implication cannot be reversed. Consider the alphabet $\{a, b\}$ and let $\varphi(a) = aa$ and $\varphi(b) = ab$. It is clear that $\varphi \in P$ (for $p = a$), but φ is not conjugate to any morphism of class P_{ret} (aaa is not a complete return word of a).

The following theorem shows the importance of morphisms of class P_{ret} for uniformly recurrent words with finite defect.

Theorem 5.5. *Let $\mathbf{u} \in \mathcal{A}^{\mathbb{N}}$ be a uniformly recurrent word with finite defect. Then there exist a rich word $\mathbf{v} \in \mathcal{B}^{\mathbb{N}}$ and a morphism $\varphi : \mathcal{B}^* \mapsto \mathcal{A}^*$ of class P_{ret} such that*

$$\mathbf{u} = \varphi(\mathbf{v}).$$

The word $\mathbf{v}$ is uniformly recurrent.

Proof. Consider a prefix z of $\mathbf{u}$ of length $|z| > \max\{2R_{\mathbf{u}}(K), H\}$, where K is the constant from Theorem 3.3 and H is an integer such that any factor of $\mathbf{u}$ of length $\geqslant H$ has its longest palindromic suffix unioccurrent. (Let us recall that the existence of H is also guaranteed by Theorem 3.3.) Since the language of $\mathbf{u}$ is closed under reversal (this follows from the fact that $\mathbf{u}$ is uniformly recurrent and contains infinitely many palindromes), $\bar{z}$ is a factor of $\mathbf{u}$ as well and its $lps(\bar{z})$ has a unique occurrence in $\bar{z}$. As $|\bar{z}| > 2R_{\mathbf{u}}(K)$ any factor shorter than or equal to K occurs in $\bar{z}$ at least twice. Therefore, $|lps(\bar{z})| > K$. Hence, $lps(\bar{z})$ is a palindromic prefix of $\mathbf{u}$ of length greater than K.

Denote $p := lps(\bar{z})$. Since $\mathbf{u}$ is uniformly recurrent, the set of return words of p is finite, say $q_0, q_1, \ldots, q_{m-1}$ is the list of all different return words. Let us define a morphism φ on the alphabet $\mathcal{B} = \{0, 1, \ldots, m-1\}$ by $\varphi(b) = q_b$ for all $b \in \mathcal{B}$. It is obvious that the morphism belongs to the class P_{ret}. Then we can write $\mathbf{u} = q_{i_0}q_{i_1}q_{i_2}\ldots$ for some sequence $(i_n)_{n\in\mathbb{N}} \in \mathcal{B}^{\mathbb{N}}$. Let us put $\mathbf{v} = (i_n)_{n\in\mathbb{N}}$.

We will show that any complete return word of any palindrome in the word $\mathbf{v}$ is a palindrome as well. According to Lemma 2.5 this implies the richness of $\mathbf{v}$.

Let s be a palindrome in $\mathbf{v}$ and x its complete return word. Then $\varphi(x)p$ has precisely two occurrences of the factor $\varphi(s)p$. As s is a palindrome, $\varphi(s)p$ is a palindrome as well of length $|\varphi(s)p| \geqslant |p| > K$. Therefore $\varphi(x)p$ is a complete return word of a long enough palindrome and according to our assumption $\varphi(x)p$ is a palindrome as well. This together with point (3) in Lemma 5.2 implies

$$\varphi(x)p = \overline{\varphi(x)p} = \varphi(\bar{x})p.$$

The point (2) then gives $x = \bar{x}$ as we claimed.

The uniform recurrence of $\mathbf{v}$ is obvious. □

The reverse implication does not hold, i.e., the set of uniformly recurrent words with finite defect is not closed under morphisms of class P_{ret}. Let us provide a construction of such a word.

Let $v_0 = \epsilon$. For $i > 0$ set

$$v_i = (v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}2v_{i-1}2)^{(+)}, \tag{2}$$

where $w^{(+)}$ denotes the shortest palindrome having w as a prefix.

Note that v_{i-1} is a prefix of v_i for all i. Thus we can set $\mathbf{v} = \lim_{i\to\infty} v_i$ and $\mathbf{v}$ is uniformly recurrent by construction.

Denote by φ a morphism from P_{ret} defined by

$$\varphi : \begin{cases} 0 \mapsto 0100, \\ 1 \mapsto 01011, \\ 2 \mapsto 010111. \end{cases} \tag{3}$$

As we will show in the sequel, the word $\mathbf{v}$ is rich and the defect $D(\varphi(\mathbf{v})) = \infty$.

Lemma 5.6. *For all i the palindrome v_i from* (2) *is rich.*

Proof. We will show for all i that v_i is rich and

$$v_i = v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}2v_{i-1}2v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}.$$

Furthermore, we will show that for all letters x, the word v_ixv_i contains exactly 2 occurrences of v_i and 1 occurrence of $0v_{i-1}xv_{i-1}0$.

We will proceed by induction on i. For $i = 1$ and 2 it is left up to the reader to verify the proposition.

Suppose the fact holds for i, $i \geqslant 2$. We will show the claim for $i + 1$. Denote by w the factor

$$w := v_i0v_i1v_i1v_i0v_i2v_i2.$$

Note that since v_ixv_i contains exactly 2 occurrences of v_i for all letters x, the factor w contains exactly 6 occurrences of v_i. In other words, if we find 1 occurrence of v_i, we know all the other occurrences.

Table 1
Enumeration of palindromic factors of w.

#	Palindromic factors of w	Count
1	Palindromic factors of v_i	$\lvert v_i\rvert + 1$
2	$0v_{i-1}0, \ldots, v_{i-1}0v_{i-1}0v_{i-1}$	$\lvert v_{i-1}\rvert + 1$
3	$1v_{i-1}0v_{i-1}1, \ldots, v_{i-1}1v_{i-1}0v_{i-1}1v_{i-1}$	$\lvert v_{i-1}\rvert + 1$
4	$2v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}2, \ldots, v_{i-1}2v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}2v_{i-1}$	$\lvert v_{i-1}\rvert + 1$
5	$0v_{i-1}0v_{i-1}0, \ldots, v_i0v_i$	$\lvert v_i\rvert - \lvert v_{i-1}\rvert$
6	$0v_{i-1}1v_{i-1}0, \ldots, v_i1v_i$	$\lvert v_i\rvert - \lvert v_{i-1}\rvert$
7	$0v_{i-1}2v_{i-1}0, \ldots, v_i2v_i$	$\lvert v_i\rvert - \lvert v_{i-1}\rvert$
8	$1v_i1, \ldots, v_i0v_i1v_i1v_i0v_i$	$2\lvert v_i\rvert + 2$
9	$2v_i2$	1
Total		$6\lvert v_i\rvert + 7$

In Table 1 we can see the total number of palindromic factors of w. Let us give a brief explanation for rows which may not be clear at first sight. Let us recall that by the induction assumption

$$v_i = v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}2v_{i-1}2v_{i-1}0v_{i-1}1v_{i-1}1v_{i-1}0v_{i-1}.$$

Since there are exactly 11 occurrences of v_{i-1} in v_i, one can easily see that factors in rows 2, 3, and 4 have not been counted in row 1. Rows 5, 6, and 7 exploit the fact that for all letters x v_ixv_i contains 1 occurrence of $0v_{i-1}xv_{i-1}0$. One can see that the total number of palindromic factors is $6|v_i| + 7 = |w| + 1$, therefore w is rich from the definition.

As the right palindromic closure preserves the richness, we can see that v_{i+1} is rich. Moreover, since there are exactly 2 occurrences of v_i in v_ixv_i for all letters x, one can see that the closure will produce the following palindrome

$$v_{i+1} = v_i0v_i1v_i1v_i0v_i2v_i2v_i0v_i1v_i1v_i0v_i.$$

Take a letter x. We can now rewrite $v_{i+1}xv_{i+1}$ in terms of v_i and see the factor $0v_ixv_i0$ occurs once and v_{i+1} occurs twice again arguing by the known count of factors v_i. □

Proposition 5.7. *The infinite word* $\mathbf{v}$ *defined in* (2) *is rich and* $D(\varphi(\mathbf{v})) = \infty$*, where* φ *is defined in* (3)*.*

Proof. Directly from the definition of $\mathbf{v}$, one can see using the previous lemma that all its prefixes v_i are rich and therefore $\mathbf{v}$ is rich.

Denote by p the palindrome from the definition of P_{ret} for the substitution φ. One can see that $p = 010$. Take $1v_i1$, a factor of $\mathbf{v}$. We have $\varphi(1v_i1) = 01011\varphi(v_i)p11$, a factor of $\varphi(\mathbf{v})$. Using point (3) of Lemma 5.2, we can see that $o_i := 1\varphi(v_i)p1$ is a palindrome. Now take $2v_i2$. One can see that $\varphi(2v_i2) = 010111\varphi(v_i)p111$. Note again the palindromic factor o_i.

We will now look for complete return words of o_i in $\varphi(r_i)$, where

$$r_i = 1v_i1v_i0v_i2v_i2.$$

The word r_i is clearly a factor of v_{i+1}, therefore a factor of $\mathbf{v}$. The first occurrence of o_i is produced by the factor $1v_i1$ in r_i. Since φ is injective, we need to look only at occurrences of v_i in r_i. The next two occurrences are in the factors $1v_i0$ and $0v_i2$. One can see that $\varphi(1v_i0) = 01011\varphi(v_i)p0$ and $\varphi(0v_i1) = 0100\varphi(v_i)p11$, i.e., the factor o_i does not occur in $\varphi(r_i)$ until the factor $\varphi(2v_i2)$ occurs. The complete return word of o_i is then $O_i := 1\varphi(v_i1v_i0v_i2v_i)p1$. By point (3) of Lemma 5.2, as $v_i1v_i0v_i2v_i$ is not a palindrome, neither is the complete return word O_i. Therefore for each i we have an oddity $\{O_i, \overline{O_i}\}$. According to Lemma 3.2, it implies the defect of $\varphi(\mathbf{v})$ is infinite. □

The last proposition shows that the set of uniformly recurrent words with finite defect is not closed under morphisms of class P_{ret}.

It is clear that the defect of an image by a morphism of class P_{ret} of a word with finite defect depends on the morphism. As the previous example shows, it depends also on the original word. To underline this fact we can take the morphism φ from (3) and $\mathbf{u}$ the Tribonacci word, i.e., the fixed point of the Tribonacci morphism $0 \mapsto 01$, $1 \mapsto 02$ and $2 \mapsto 0$ – a well-known rich word [9]. It is easy to see that $D(\varphi(\mathbf{u})) = 0$.

6. Comments

At the end of the article [4], the authors state several open questions, among them the following one: "Let $\mathbf{u}$ be a fixed point of a primitive morphism. If the defect is finite and non-zero, is the word $\mathbf{u}$ necessarily periodic?"

We are not able to answer this question. The following observation is just a small comment to it.

Observation 6.1. *Let $\mathbf{u}$ be a fixed point of a primitive morphism and let its defect $D(\mathbf{u})$ be finite. Then there exists a rich word $\mathbf{v}$ and a morphism $\varphi \in P_{\text{ret}}$ such that $\mathbf{u} = \varphi(\mathbf{v})$ and $\mathbf{v}$ itself is a fixed point of a primitive morphism as well.*

Proof. The rich word $\mathbf{v}$, which we have constructed in the proof of Theorem 5.5, is a derived word, as introduced by Durand in [10]. Lemma 19 of [10] says that any derived word of a fixed point of a primitive morphism is a fixed point of a primitive morphism as well. □

Theorem 5.5 has the form of implication, which cannot be reversed, since Proposition 5.7 demonstrates that a morphism from P_{ret} does not always preserve the set of words with finite defect. It is thus natural to ask the following questions:

1. Can the class P_{ret} be replaced with a smaller one in such a way that Theorem 5.5 can be stated in the form of equivalence?
2. Characterize those morphisms from P_{ret} that preserve the set of rich words.
3. Find an algorithm to compute $D(\varphi(\mathbf{u}))$ for a rich word $\mathbf{u}$ and a morphism from $\varphi \in P_{\text{ret}}$.
4. Characterize morphisms φ on $\mathcal{B}^*$ with the property that $\varphi(\mathbf{u})$ has finite defect for any infinite word $\mathbf{u} \in \mathcal{B}^{\mathbb{N}}$ with finite defect.

We end with some remarks on Question 1. The authors of [11] define another class of morphisms that play an important role in the study of finite defect. They call a morphism φ on $\mathcal{A}^*$ a *standard morphism of class P* (or a *standard P-morphism*) if there exists a palindrome r (possibly empty) such that, for all $x \in \mathcal{A}$, $\varphi(x) = rq_x$, where the q_x are palindromes. If r is non-empty, then some (or all) of the palindromes q_x may be empty or may even take the form $q_x = \pi_x^{-1}$ with π_x a proper palindromic suffix of r. They say that a standard P-morphism is *special* if:

1. all $\varphi(x) = rq_x$ end with different letters, and
2. whenever $\varphi(x)r = rq_x r$, with $x \in \mathcal{A}$, occurs in some $\varphi(y_1 y_2 \ldots y_n)r$, then this occurrence is $\varphi(y_m)r$ for some m with $1 \leqslant m \leqslant n$.

They prove the following theorem.

Theorem 6.2. *(See [11, Theorem 6.28].) If φ is a standard special P-morphism on $\mathcal{A}^*$ and $\mathbf{u} \in \mathcal{A}^*$, then $D(\mathbf{u})$ is finite if and only if $D(\varphi(\mathbf{u}))$ is finite.*

However, as shown in the following proposition, standard special P-morphisms are not the only ones that preserve the set of uniformly recurrent words with finite defect, thus the class of standard special morphisms is too small as an answer to Question 1. Let us add that standard special morphisms of class P do not form a subset of morphisms of class P_{ret}. For instance, $\varphi: a \to aabbaabba,\ b \to ab$ is a standard special P-morphism with $r = a$, but does not belong to P_{ret}.

Proposition 6.3. *Let* $\mathbf{u}$ *be a binary uniformly recurrent word such that* $D(\mathbf{u})$ *is finite. Let* φ *be a morphism of class* P_{ret}*. Then* $D(\varphi(\mathbf{u}))$ *is finite.*

Lemma 6.4. *Let* φ *be a morphism of class* P_{ret} *on* $\{0,1\}^*$*. Then* φ *is conjugate to a standard special* P*-morphism.*

Proof. Let p be the palindrome corresponding to φ in the definition of P_{ret}. Denote by p_1 the longest common suffix of $\varphi(0)$ and $\varphi(1)$. Denote by p_2 a word such that pp_2 is the longest common prefix of $\varphi(0)p$ and $\varphi(1)p$. Using properties of P_{ret} we have $p_1 = \overline{p_2}$. Define $\sigma(0) = p_1\varphi(0)p_1^{-1}$ and $\sigma(1) = p_1\varphi(1)p_1^{-1}$. Then φ is conjugate to σ and σ is a standard special P-morphism with the corresponding palindrome $r = p_1 p p_1$. □

Proof of Proposition 6.3. By Lemma 6.4 the morphism φ is conjugate to a standard special P-morphism σ. Clearly, the languages of $\varphi(\mathbf{u})$ and $\sigma(\mathbf{u})$ are the same, hence $D(\varphi(\mathbf{u})) = D(\sigma(\mathbf{u}))$. Theorem 6.2 implies that $D(\sigma(\mathbf{u})) < \infty$. □

Acknowledgments

We acknowledge financial support by the Czech Science Foundation grant 201/09/0584, by the grants MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic, and by the grant SGS10/085OHK4/1T/14 of the Grant Agency of the Czech Technical University in Prague.

References

[1] J.P. Allouche, M. Baake, J. Cassaigne, D. Damanik, Palindrome complexity, Theoret. Comput. Sci. 292 (2003) 9–31.

[2] P. Baláži, Z. Masáková, E. Pelantová, Factor versus palindromic complexity of uniformly recurrent infinite words, Theoret. Comput. Sci. 380 (2007) 266–275.

[3] L'. Balková, E. Pelantová, Š. Starosta, Sturmian jungle (or garden?) on multiliteral alphabets, RAIRO – Theor. Inform. Appl. (2010), in press, arXiv:1003.1224, the original publication is available at www.edpsciences.org/ita.

[4] A. Blondin-Massé, S. Brlek, A. Garon, S. Labbé, Combinatorial properties of f-palindromes in the Thue–Morse sequences, PUMA 19 (2008) 39–52.

[5] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, On the palindromic complexity of infinite words, in: J. Berstel, J. Karhumäki, D. Perrin (Eds.), Combinatorics on Words with Applications, Internat. J. Found. Comput. Sci. 15 (2) (2004) 293–306.

[6] M. Bucci, A. De Luca, A. Glen, L.Q. Zamboni, A connection between palindromic and factor complexity using return words, Adv. in Appl. Math. 42 (2009) 60–74.

[7] M. Bucci, A. De Luca, A. Glen, L.Q. Zamboni, A new characteristic property of rich words, Theoret. Comput. Sci. 410 (2009) 2860–2863.

[8] J. Cassaigne, Complexity and special factors, Bull. Belg. Math. Soc. Simon Stevin 4 (1) (1997) 67–88.

[9] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theoret. Comput. Sci. 255 (2001) 539–553.

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

[11] A. Glen, J. Justin, S. Widmer, L.Q. Zamboni, Palindromic richness, European J. Combin. 30 (2009) 510–531.

[12] A. Hof, O. Knill, B. Simon, Singular continuous spectrum for palindromic Schrödinger operators, Comm. Math. Phys. 174 (1995) 149–159.

[13] B. Tan, Mirror substitutions and palindromic sequences, Theoret. Comput. Sci. 389 (2007) 118–124.

On the Brlek–Reutenauer Conjecture

Authors: **Ľubomíra Balková, Edita Pelantová, Štěpán Starosta**
Journal: Theoretical Computer Science
Year: 2011
Volume: 412
Pages: 5649–5655

Theoretical Computer Science 412 (2011) 5649–5655

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

On Brlek–Reutenauer conjecture

L'. Balková [*], E. Pelantová, Š. Starosta

Department of Mathematics FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic

ARTICLE INFO

Article history:
Received 26 January 2011
Received in revised form 6 June 2011
Accepted 22 June 2011
Communicated by D. Perrin

Keywords:
Defect
Rich word
Full word
Infinite word
Factor complexity
Palindromic complexity

ABSTRACT

Brlek and Reutenauer conjectured that any infinite word $\mathbf{u}$ with language closed under reversal satisfies the equality $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ in which $D(\mathbf{u})$ denotes the defect of $\mathbf{u}$ and $T_{\mathbf{u}}(n)$ denotes $\mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) + 2 - \mathcal{P}_{\mathbf{u}}(n+1) - \mathcal{P}_{\mathbf{u}}(n)$, where $\mathcal{C}_{\mathbf{u}}$ and $\mathcal{P}_{\mathbf{u}}$ are the factor and palindromic complexity of $\mathbf{u}$, respectively. Brlek and Reutenauer verified their conjecture for periodic infinite words. Using their result, we prove the conjecture for uniformly recurrent words. Moreover, we summarize results and some open problems related to defects, which may be useful for the proof of the Brlek–Reutenauer conjecture in full generality.

1. Introduction

There have recently been quite a lot of papers devoted to palindromes in infinite words. Droubay, Justin, and Pirillo determined in [6] the upper bound on the number of distinct palindromes occurring in a finite word — a finite word w contains at most $|w| + 1$ different palindromes, where $|w|$ denotes the length of w. The difference between the utmost number $|w| + 1$ and the actual number of palindromes in w is called the defect of w and it is usually denoted by $D(w)$. An infinite word $\mathbf{u}$ whose factors all have zero defect was baptized rich or full. In [1], Baláži, et al. proved for infinite words with language closed under reversal an inequality relating the palindromic and factor complexity of an infinite word $\mathbf{u}$ denoted $\mathcal{P}_{\mathbf{u}}$ and $\mathcal{C}_{\mathbf{u}}$, respectively. For such infinite words, it holds

$$\mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) + 2 - \mathcal{P}_{\mathbf{u}}(n) - \mathcal{P}_{\mathbf{u}}(n+1) \geq 0 \quad \text{for all } n \in \mathbb{N}. \tag{1}$$

In [5], Bucci, et al. showed that rich words with language closed under reversal can be characterized by the equality in (1). Brlek, et al. in [3] defined the defect $D(\mathbf{u})$ of an infinite word $\mathbf{u}$ as the maximum defects of all its factors and they studied its value for periodic words.

Recently, in [2], the authors of this paper have proved that for a uniformly recurrent word $\mathbf{u}$, its defect $D(\mathbf{u})$ is finite if and only if the equality in (1) is attained for all but a finite number of indices n.

Despite the fact that numerous researchers study palindromes, only recently Brlek and Reutenauer have noticed that the value of defect is closely tied with the expression on the left-hand side of (1) — let us denote it by $T_{\mathbf{u}}(n)$. They have shown that for periodic infinite words with language closed under reversal, it holds $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$. Their conjecture says that the same equation holds for all infinite words with language closed under reversal.

In this paper, using the result of Brlek and Reutenauer for periodic words, we will prove that the Brlek–Reutenauer conjecture is true for uniformly recurrent words and in the last chapter we will discuss some aspects concerning the conjecture for infinite words that are not uniformly recurrent.

[*] Corresponding author. Tel.: +420 224358544.
E-mail addresses: lubomira.balkova@fjfi.cvut.cz, l.balkova@centrum.cz (L'. Balková).

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.06.031

2. Preliminaries

By $\mathcal{A}$ we denote a finite set of symbols called *letters*; the set $\mathcal{A}$ is therefore called an *alphabet*. A finite string $w = w_0w_1 \ldots w_{n-1}$ of letters from $\mathcal{A}$ is said to be a *finite word*, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ϵ as the neutral element form a free monoid $\mathcal{A}^*$. The map

$$w = w_0w_1 \ldots w_{n-1} \mapsto \overline{w} = w_{n-1}w_{n-2} \ldots w_0$$

is a bijection on $\mathcal{A}^*$, the word $\overline{w}$ is called the *reversal* or the *mirror image* of w. A word w which coincides with its mirror image is a *palindrome*.

Under an *infinite word* we understand an infinite string $\mathbf{u} = u_0u_1u_2 \ldots$ of letters from $\mathcal{A}$. A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \epsilon$, then w is said to be a *prefix* of v, if $s = \epsilon$, then w is a *suffix* of v.

The *language* $\mathcal{L}(\mathbf{u})$ of an infinite word $\mathbf{u}$ is the set of all its factors. Factors of $\mathbf{u}$ of length n form the set denoted by $\mathcal{L}_n(\mathbf{u})$. We say that the language $\mathcal{L}(\mathbf{u})$ is *closed under reversal* if $\mathcal{L}(\mathbf{u})$ contains with every factor w, also its reversal $\overline{w}$.

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2} \ldots$. Such an index is called an *occurrence* of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ has infinitely many occurrences in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be *recurrent*. It is easy to see that if the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent (a proof can be found in [7]). For a recurrent infinite word $\mathbf{u}$, we may define the notion of a *complete return word* of any $w \in \mathcal{L}(\mathbf{u})$. It is a factor $v \in \mathcal{L}(\mathbf{u})$ such that w is a prefix and a suffix of v and w occurs in v exactly twice. Under a *return word* of a factor w is usually understood a word $q \in \mathcal{L}(\mathbf{u})$ such that qw is a complete return word of w. If any factor $w \in \mathcal{L}(\mathbf{u})$ has only finitely many return words, then the infinite word $\mathbf{u}$ is called *uniformly recurrent*. If $\mathbf{u}$ is a uniformly recurrent word, we can find for any $n \in \mathbb{N}$ a number R such that any factor of $\mathbf{u}$ which is longer than R already contains all factors of $\mathbf{u}$ of length n.

The *factor complexity* of an infinite word $\mathbf{u}$ is the mapping $\mathcal{C}_{\mathbf{u}} : \mathbb{N} \mapsto \mathbb{N}$ defined by the prescription $\mathcal{C}_{\mathbf{u}}(n) := \#\mathcal{L}_n(\mathbf{u})$. To determine the first difference of the factor complexity, one has to count the possible extensions of factors of length n. A *right extension* of $w \in \mathcal{L}(\mathbf{u})$ is any letter $a \in \mathcal{A}$ such that $wa \in \mathcal{L}(\mathbf{u})$. Of course, any factor of $\mathbf{u}$ has at least one right extension. A factor w is called *right special* if w has at least two right extensions. Similarly, one can define a *left extension* and a *left special* factor. We will deal mainly with recurrent infinite words $\mathbf{u}$. In such a case, any factor of $\mathbf{u}$ has at least one left extension.

The *defect* $D(w)$ of a finite word w is the difference between the utmost number of distinct palindromes $|w| + 1$ and the actual number of distinct palindromes contained in w. Finite words with zero defects – called *rich* or *full* words – can be viewed as the most saturated by palindromes. This definition may be extended to infinite words as follows.

Definition 2.1. *An infinite word $\mathbf{u} = u_0u_1u_2 \ldots$ is called rich or full, if for any index $n \in \mathbb{N}$, the prefix $u_0u_1u_2 \ldots u_{n-1}$ of length n contains exactly $n + 1$ different palindromes.*

Let us remark that not only all prefixes of rich words are rich, but also all factors are rich. A result from [6] provides us with a handful tool which helps to evaluate the defect of a factor.

Proposition 2.2 (*[6]*). *A finite or infinite word $\mathbf{u}$ is rich if and only if the longest palindromic suffix of w occurs exactly once in w for any prefix w of $\mathbf{u}$.*

In accordance with the terminology introduced in [6], the factor with a unique occurrence in another factor is called *unioccurrent*. From the proof of the previous proposition directly follows the next corollary.

Corollary 2.3. *The defect $D(w)$ of a finite word w is equal to the number of prefixes w' of w, for which the longest palindromic suffix of w' is not unioccurrent in w'. In other words, if b is a letter and w a finite word, then $D(wb) = D(w) + \delta$, where $\delta = 0$ if the longest palindromic suffix of wb occurs exactly once in wb and $\delta = 1$ otherwise.*

This corollary implies that $D(v) \geq D(w)$ whenever w is a factor of v. It enables to give a reasonable definition of the defect of an infinite word (see [3]).

Definition 2.4. *The defect of an infinite word $\mathbf{u}$ is the number (finite or infinite)*

$$D(\mathbf{u}) = \sup\{D(w) \bigm| w \text{ is a prefix of } \mathbf{u}\}\,.$$

Let us point out several facts concerning defects that are easy to prove:

(1) If we consider all factors of a finite or an infinite word $\mathbf{u}$, we obtain the same defect, i.e.,

$$D(\mathbf{u}) = \sup\{D(w) \bigm| w \in \mathcal{L}(\mathbf{u})\}\,.$$

(2) Any infinite word with finite defect contains infinitely many palindromes.
(3) Infinite words with zero defect correspond exactly to rich words.

Periodic words with finite defect have been studied in [3] and in [7]. It holds that the defect of an infinite periodic word with the minimal period w is finite if and only if $w = pq$, where both p and q are palindromes. Words with finite defect have been studied in [2] and [7].

The number of palindromes of a fixed length occurring in an infinite word is measured by the so called *palindromic complexity* $\mathcal{P}_{\mathbf{u}}$, the mapping which assigns to any non-negative integer n the number

$$\mathcal{P}_{\mathbf{u}}(n) := \#\{w \in \mathcal{L}_n(u) \mid w \text{ is a palindrome}\}\,.$$

Denote by

$$T_{\mathbf{u}}(n) = \mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) + 2 - \mathcal{P}_{\mathbf{u}}(n+1) - \mathcal{P}_{\mathbf{u}}(n).$$

The following proposition is proved in [1] for uniformly recurrent words, however the uniform recurrence is not needed in the proof, thus it holds for any infinite word with language closed under reversal.

Proposition 2.5 ([1])**.** *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. Then*

$$T_{\mathbf{u}}(n) \geq 0, \tag{2}$$

for all $n \in \mathbb{N}$.

It is shown in [5] that this bound can be used for a characterization of rich words as well. The following proposition states this fact.

Proposition 2.6 ([5])**.** *An infinite word* $\mathbf{u}$ *with language closed under reversal is rich if and only if the equality in* (2) *holds for all* $n \in \mathbb{N}$.

Let $\mathbf{u}$ be an infinite word with language closed under reversal. Using the proof of Proposition 2.5, those $n \in \mathbb{N}$ for which $T_{\mathbf{u}}(n) = 0$ can be characterized in the graph language.

An *n-simple path* e is a factor of $\mathbf{u}$ of length at least $n + 1$ such that the only special (right or left) factors of length n occurring in e are its prefix and suffix of length n. If w is the prefix of e of length n and v is the suffix of e of length n, we say that the n-simple path e starts in w and ends in v. We will denote by $G_n(\mathbf{u})$ an undirected graph whose set of vertices is formed by unordered pairs $(w, \overline{w})$ such that $w \in \mathcal{L}_n(\mathbf{u})$ is right or left special. We connect two vertices $(w, \overline{w})$ and $(v, \overline{v})$ by an unordered pair $(e, \overline{e})$ if e or $\overline{e}$ is an n-simple path starting in w or $\overline{w}$ and ending in v or $\overline{v}$. Note that the graph $G_n(\mathbf{u})$ may have multiple edges and loops.

Remark 2.7. *Let us point out that if* $\mathcal{L}_n(\mathbf{u})$ *contains no special factor, then* $G_n(\mathbf{u})$ *is an empty graph. In this case the word* $\mathbf{u}$ *is periodic, i.e., there exists a primitive word* w *such that* $\mathbf{u} = w^{\omega}$ *and* $|w| \leq n$*. As proved in [3], since the language of* $\mathbf{u}$ *is closed under reversal, the word* w *is a product of two palindromes. It is easy to see that* $\mathcal{C}_{\mathbf{u}}(n+1) = \mathcal{C}_{\mathbf{u}}(n)$ *and* $2 = \mathcal{P}_{\mathbf{u}}(n+1) + \mathcal{P}_{\mathbf{u}}(n)$*. Therefore* $T_{\mathbf{u}}(n) = 0$.

Lemma 2.8. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal,* $n \in \mathbb{N}$*. Then* $T_{\mathbf{u}}(n) = 0$ *if and only if both of the following conditions are met:*

(1) *The graph obtained from* $G_n(\mathbf{u})$ *by removing loops is a tree;*
(2) *Any n-simple path forming a loop in the graph* $G_n(\mathbf{u})$ *is a palindrome.*

Proof. It is a direct consequence of the proof of Theorem 1.2 in [1] (recalled in this paper as Proposition 2.5). □

Corollary 2.9. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be infinite words with language closed under reversal and* $n \in \mathbb{N}$.

$$\mathcal{L}_{n+1}(\mathbf{v}) \subset \mathcal{L}_{n+1}(\mathbf{u}) \quad \text{and} \quad T_{\mathbf{u}}(n) = 0 \Longrightarrow T_{\mathbf{v}}(n) = 0.$$

Proof. Our assumptions imply that $G_n(\mathbf{v})$ is a subgraph of $G_n(\mathbf{u})$ and $G_n(\mathbf{u})$ meets both conditions in the previous lemma. These conditions are hereditary, i.e., any connected subgraph inherits these conditions as well. □

3. Brlek–Reutenauer conjecture

Brlek and Reutenauer gave in [4] a conjecture relating the defect and the factor and palindromic complexity of infinite words with language closed under reversal.

Conjecture 3.1 (*Brlek–Reutenauer Conjecture*)**.** *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. Then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)\,. \tag{3}$$

It is known from [5] that Conjecture 3.1 holds for rich words.

Theorem 3.2. *Let* $\mathbf{u}$ *be a rich infinite word with the language closed under reversal. Then* (3) *holds.*

Brlek and Reutenauer provided in [4] a result for periodic words.

Theorem 3.3. *Let* $\mathbf{u}$ *be a periodic infinite word. Then* (3) *holds.*

In the sequel, we will prove the following theorem.

Theorem 3.4. *Let* $\mathbf{u}$ *be an infinite word with the language closed under reversal. If* $\mathbf{u}$ *satisfies two assumptions:*

(1) *Both* $D(\mathbf{u})$ *and* $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ *are finite.*
(2) *For any* $M \in \mathbb{N}$ *there exists a factor* $w \in \mathcal{L}(\mathbf{u})$ *such that*
- w *contains all factors of* $\mathbf{u}$ *of length* M,
- ww *is a factor of* $\mathbf{u}$.

Then (3) *holds.*

In order to prove Theorem 3.4, we need to put together several claims. Let us first describe the main ideas of the proof. The assumptions of Theorem 3.4 enable us to construct a periodic word $\mathbf{v}$ with language closed under reversal such that

- $D(\mathbf{u}) = D(\mathbf{v})$ and
- $T_{\mathbf{v}}(n) = T_{\mathbf{u}}(n)$ for all $n \in \mathbb{N}$.

Theorem 3.3 applied to the periodic word $\mathbf{v}$ then concludes the proof.

Let us construct a suitable periodic word. As $D(\mathbf{u})$ is finite, there exists a factor $f \in \mathcal{L}(\mathbf{u})$ such that $D(\mathbf{u}) = D(f)$. Let us denote its length by $H = |f|$. According to the inequality (2), the finiteness of $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ implies that there exists an integer $N \in \mathbb{N}$ such that $T_{\mathbf{u}}(n) = 0$ for all $n \geq N$. Let us put

$$M = \max\{N, H\}. \tag{4}$$

By Assumption (2), there exists a factor w containing all elements of $\mathcal{L}_M(\mathbf{u})$. Let us define

$$\mathbf{v} = w^{\omega}.$$

Claim 3.5. *The word* w *is a concatenation of two palindromes, in particular, the periodic word* w^{ω} *has the language closed under reversal.*

Proof. Since the factor w contains the factor f and the square ww belongs to $\mathcal{L}(\mathbf{u})$, we have $D(f) \leq D(w) \leq D(ww) \leq D(\mathbf{u})$. As the factor f was chosen to satisfy $D(\mathbf{u}) = D(f)$, we may conclude that

$$D(f) = D(w) = D(ww) = D(\mathbf{u}). \tag{5}$$

The factor ww is longer than the factor f and has the same defect as f. Let us denote by p the longest palindromic suffix of $ww \in \mathcal{L}(\mathbf{u})$. According to Corollary 2.3, the palindrome p occurs in ww exactly once and therefore $|p| > |w|$. There exists a proper prefix w' of w such that $ww = w'p$. Let us denote by w'' the suffix of w for which $w = w'w''$. It means that $p = w''w'w''$. As p is a palindrome, we have $\overline{w''} = w''$ and $\overline{w'} = w'$. Hence the word w is a concatenation of two palindromes. □

Claim 3.6. $D(\mathbf{v}) = D(\mathbf{u})$.

Proof. We will use Theorem 6 from [3]. It implies that if w is a product of two palindromes, then $D(w^{\omega}) = D(ww)$. This together with (5) concludes the proof. □

Claim 3.7. $T_{\mathbf{v}}(n) = T_{\mathbf{u}}(n)$ *for all* $n \in \mathbb{N}$.

Proof. Let us first consider $n \leq M - 1$, where M is the constant given by (4). Since w contains all elements of $\mathcal{L}_M(\mathbf{u})$, we have $M \leq |w|$. Since $ww \in \mathcal{L}(\mathbf{u})$, we also have $\mathcal{L}_M(\mathbf{v}) = \mathcal{L}_M(\mathbf{u})$. It implies

$$\mathcal{C}_{\mathbf{u}}(n) = \mathcal{C}_{\mathbf{v}}(n) \quad \text{and} \quad \mathcal{P}_{\mathbf{u}}(n) = \mathcal{P}_{\mathbf{v}}(n) \quad \text{for all } n \leq M.$$

It gives the statement of the claim for all $n \leq M - 1$.

Now we will consider $|w| > n \geq M$. According to the definition of $N \leq M$, it holds that $T_{\mathbf{u}}(n) = 0$. Since $\mathcal{L}_{n+1}(\mathbf{v}) \subset \mathcal{L}_{n+1}(\mathbf{u})$, Corollary 2.9 gives $T_{\mathbf{v}}(n) = 0$ as well.

Finally, we consider $n \geq |w| \geq M$. Since n is longer than or equal to the period of $\mathbf{v}$ and since w is a product of two palindromes, we have $\mathcal{C}_{\mathbf{v}}(n+1) = \mathcal{C}_{\mathbf{v}}(n)$ and $\mathcal{P}_{\mathbf{v}}(n+1) + \mathcal{P}_{\mathbf{v}}(n) = 2$. It implies $T_{\mathbf{v}}(n) = 0$. The value $T_{\mathbf{u}}(n)$ is zero as well, according to the fact that $N \leq M$. □

Proof of Theorem 3.4. It suffices to put together Claims 3.5–3.7 and to realize that Conjecture 3.1 was already proved for periodic words, here stated as Theorem 3.3. □

4. The Brlek–Reutenauer conjecture holds for uniformly recurrent words

In this section we will show that either both sides in the Brlek–Reutenauer equality (3) are infinite or both assumptions of Theorem 3.4 are satisfied for uniformly recurrent words, which results in the main theorem of this paper.

Theorem 4.1. *If* $\mathbf{u}$ *is a uniformly recurrent infinite word with the language closed under reversal, then* (3) *holds.*

In order to prove Theorem 4.1, we will make use of several equivalent characterizations of infinite words with finite defect.

Theorem 4.2. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word with language closed under reversal. Then the following statements are equivalent.*

(1) *The defect of* $\mathbf{u}$ *is finite.*
(2) *There exists an integer* K *such that any complete return word of a palindrome of length at least* K *is a palindrome as well.*
(3) *There exists an integer* H *such that the longest palindromic suffix of any factor* w *of length* $|w| \geq H$ *occurs in* w *exactly once.*
(4) *There exists an integer* N *such that*

$$T_{\mathbf{u}}(n) = 0 \quad \textit{for all } n \geq N.$$

Proof. (1) and (2) are equivalent by Theorem 4.8 from [7]. It follows from the definition of $D(\mathbf{u})$ that (1) and (3) are equivalent. The equivalence of (1) and (4) was stated as Theorem 4.1 in [2]. □

Corollary 4.3. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word with language closed under reversal. Then*

$$D(\mathbf{u}) \ \textit{is finite} \iff \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) \ \textit{is finite}.$$

Thanks to Corollary 4.3, we can focus on uniformly recurrent words $\mathbf{u}$ with finite defect. An important role in the proof of Theorem 4.1 is the presence of squares in $\mathbf{u}$.

Lemma 4.4. *Let* $\mathbf{u}$ *be a uniformly recurrent infinite word with finite defect and with language closed under reversal. Then the set*

$$\{w \in \mathcal{A}^* \mid ww \in \mathcal{L}(\mathbf{u})\}$$

is infinite.

Proof. We shall prove that for any $L \in \mathbb{N}$ there exists a factor w such that $ww \in \mathcal{L}(\mathbf{u})$ and $|w| > L$. Without loss of generality take $L > K$, where K is the constant from the statement (3) of Theorem 4.2. Then any complete return word of a palindrome which is longer than L is a palindrome as well. This implies that $\mathbf{u}$ has infinitely many palindromes. Thus there exists an infinite palindromic branch, i.e., a both-sided infinite word $\dots v_3v_2v_1v_0v_1v_2v_3\dots$, where $v_i \in \mathcal{A}$ for $i = 1, 2, 3, \dots$ and $v_0 \in \mathcal{A} \cup \{\epsilon\}$ such that $v_kv_{k-1}\dots v_0 \dots v_{k-1}v_k \in \mathcal{L}(\mathbf{u})$ for any $k \in \mathbb{N}$. Consider a palindrome $q = v_kv_{k-1}\dots v_0 \dots v_{k-1}v_k$ where $|q| > 3L$. Since $\mathbf{u}$ is uniformly recurrent, there exists an index $i > k$ such that the factor $f = v_iv_{i-1}\dots v_{k+2}v_{k+1}$ is a return word of q. The factor fq is a complete return word of the palindrome q and therefore fq is a palindrome.

At first suppose that the return word f is longer then $|q|$. In this case, $f = qp$ for some palindrome p. Hence the palindromic branch has as its central factor the word $qpqpq$. We can put $w = qp$.

Now suppose that the return word f satisfies $|f| \leq |q|$. In this case there exists an integer $j \geq 2$ and a factor y such that $fq = f^jy$ and $|y| < |f|$. If we put $w = f^i$, with $i = \lfloor \frac{j}{2} \rfloor$, then $ww \in \mathcal{L}(\mathbf{u})$ and $|w| > \frac{1}{3}|q| \geq L$. □

Proof of Theorem 4.1. By Corollary 4.3, the equality $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ holds as soon as one of the sides is infinite. Assume that $D(\mathbf{u}) < +\infty$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty$. Let $M \in \mathbb{N}$ be an arbitrary integer. As $\mathbf{u}$ is uniformly recurrent, there exists an integer R such that any factor longer than R contains all factors of $\mathbf{u}$ of length at most M. According to Lemma 4.4, the set of squares occurring in $\mathbf{u}$ is infinite, thus there exists a factor w longer than R such that ww belongs to the language of $\mathbf{u}$. Its length guarantees that w contains all elements of $\mathcal{L}_M(\mathbf{u})$.

Consequently, Assumptions (1) and (2) of Theorem 3.4 are met and the equality $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ follows. □

5. Open problems

In this section, we summarize which statements concerning defects are known for infinite words which are not necessarily uniformly recurrent.

Let us transform the Brlek–Reutenauer conjecture into a more general question: "For which infinite words $\mathbf{u}$ does the equality

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) \tag{6}$$

hold?"

In our summary of properties related to the above question, let us first recall Proposition 4.6 from [7] which applies in full generality.

Proposition 5.1. *Let* **u** *be an infinite word.*
$D(\mathbf{u}) \geq \#\big\{\{v, \overline{v}\} \mid v \neq \overline{v} \text{ and } v \text{ or } \overline{v} \text{ is a complete return word in } \mathbf{u} \text{ of a palindrome } w\big\}.$

In [7], the set $\{v, \overline{v}\}$ is called an *oddity*.

Observation 5.2. *If an infinite word* **u** *contains finitely many distinct palindromes, then the equality* (6) *holds.*

Proof. It follows from the definition that $D(\mathbf{u}) = +\infty$. Since $\mathcal{P}_{\mathbf{u}}(n) = 0$ for n large enough and $\mathcal{C}_{\mathbf{u}}$ is non-decreasing, we have $T_{\mathbf{u}}(n) \geq 2$ for such indices n. Consequently, $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) = +\infty$. □

Observation 5.3. *Let* **u** *be a periodic word. Then the equality* (6) *holds.*

Proof. Theorem 3.3 states this fact for infinite words with language closed under reversal. In [3] it is shown that periodic words whose language is not closed under reversal contain only finitely many palindromes. Thus, the previous observation implies that the equality is reached for such words, too. □

Observation 5.4. *Let* **u** *be a uniformly recurrent word. Then the equality* (6) *holds.*

Proof. Theorem 4.1 states this fact for infinite words with language closed under reversal. It is well known for uniformly recurrent words whose language is not closed under reversal that it contains only a finite number of palindromes. In such a case, both sides of (6) are infinite. □

From now on, let us limit our considerations to infinite words containing infinitely many palindromes in their language.

Observation 5.5. *The equality* (6) *does not hold in general for infinite words which are not recurrent.*

Proof. The word $\mathbf{u} = ab^{\omega}$ is rich, i.e., $D(\mathbf{u}) = 0$, however $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) = -1$. □

Problem 1. It is an open problem whether the equality (6) holds for recurrent words whose language is not closed under reversal and contains infinitely many palindromes. We have examples for which the equality holds and we have so far no example refuting the equality (6).

Example 5.6. *Let* **u** *be an infinite ternary word satisfying* $\mathbf{u} = \lim_{n\to+\infty} u_n$*, where* $u_0 = a$ *and* $u_{n+1} = u_n b^{n+1} c^{n+1} u_n$*. The word* **u** *is recurrent, however not closed under reversal (it does not contain the factor cb). On one hand,* $D(\mathbf{u}) = +\infty$ *because* b^k *has non-palindromic complete return words for any* $k \geq 1$*, thus the number of oddities is infinite. On the other hand, since the only left extension of a is c and the only right extension of a is b, it is readily seen that the only palindromes of length greater than* 1 *are of the form* b^n *and* c^n*, thus* $\mathcal{P}_{\mathbf{u}}(n) = 2$ *for all* $n \geq 2$*. It is also easy to show that* c^n*,* b^n*, and* $b^{n-1}c$ *are distinct left special factors of length* $n \geq 2$*, therefore* $\mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) \geq 3$ *for all* $n \geq 2$*. This implies that* $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) = +\infty$*.*

In the sequel, let us consider infinite words whose language is closed under reversal and contains infinitely many palindromes. In this case, the sum from the left-hand side of (3) exists by Proposition 2.5, i.e., it is either a nonnegative number or $+\infty$.

Any rich word with language closed under reversal satisfies (6) by Theorem 3.2. For instance, the Rote word **u** - the fixed point of the morphism φ defined by $\varphi(0) = 001$ and $\varphi(1) = 111$, i.e., $\mathbf{u} = \varphi(\mathbf{u})$ - is rich because it satisfies $T_{\mathbf{u}}(n) = 0$ for all $n \in \mathbb{N}$, which is not difficult to show. Therefore, the Rote word is an example of an infinite word which is not uniformly recurrent (it contains blocks of ones of any length) satisfying the equality (6). We have, of course, no counterexample which would refute the Brlek–Reutenauer conjecture.

There exist several equivalent characterizations of words with finite defect.

Theorem 5.7. *Let* **u** *be an infinite word with language closed under reversal and containing infinitely many palindromes. Then the following statements are equivalent.*

(1) *The defect of* **u** *is finite.*
(2) **u** *has only finitely many oddities.*
(3) *There exists an integer H such that the longest palindromic suffix of any factor* w *of length* $|w| \geq H$ *occurs in* w *exactly once.*

Proof. (1) and (3) are equivalent by the definition of defect. (1) implies (2) by Proposition 5.1. The implication (2) ⇒ (1) was proved as Proposition 4.8 in [7] for uniformly recurrent words. However, we will show that the proof works for words with language closed under reversal and containing infinitely many palindromes too.

Assume that $D(\mathbf{u}) = +\infty$ and the number of oddities is finite.

A finite number of oddities means that only finitely many palindromes can have non-palindromic complete return words. Let the longest such palindrome be of length K.

Since the number of palindromes is infinite, there exist infinitely many non-defective positions. Denote by $u^{(n)}$ the prefix of **u** of length n. Then n is a non-defective position if $D(u^{(n-1)}) = D(u^{(n)})$ (such positions correspond to the first occurrences of palindromes).

There exists an integer H such that the prefix of **u** of length H contains all palindromes of length lower than $K+3$. Hence, if $n > H$ is a non-defective position, then the longest palindromic suffix of $u^{(n)}$ is of length greater than $K+2$.

Since both the number of defective and non-defective positions is infinite, we can find an index $k > H$ such that k is a defective and $k+1$ a non-defective position. The longest palindromic suffix p of $u^{(k)}$ occurs at least twice in $u^{(k)}$, thus $u^{(k)}$ ends in a non-palindromic complete return word of p. Since $k+1$ is a non-defective position, it can be easily shown by contradiction that the longest palindromic suffix of $u^{(k+1)}$ is of length equal to or lower than $|p| + 2 \leq K + 2$.

This is a contradiction with the fact that non-defective positions greater than H have their longest palindromic suffix longer than $K+2$. □

For words with language closed under reversal, some implications remain valid. The first one is Proposition 4.3 and the second one is Proposition 4.5 from [2].

Proposition 5.8. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. Suppose that there exists an integer* N *such that for all* $n \geq N$ *the equality* $T_{\mathbf{u}}(n) = 0$ *holds. Then the complete return words of any palindromic factor of length* $n \geq N$ *are palindromes.*

Proposition 5.9. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. If there exists an integer* H *such that for any factor* $f \in \mathcal{L}(\mathbf{u})$ *of length* $|f| \geq H$*, the longest palindromic suffix of* f *is unioccurrent in* f*. Then* $T_{\mathbf{u}}(n) = 0$ *for any* $n \geq H$*.*

The last proposition together with Theorem 5.7 results in the following corollary.

Corollary 5.10. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. Then we have*

$$D(\mathbf{u}) < +\infty \Rightarrow \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty.$$

It is an open question whether the implications in the previous propositions can be reversed.

Problem 2. Let $\mathbf{u}$ be an infinite word with language closed under reversal and containing infinitely many palindromes. Assume that there exists an integer K such that all palindromes of length equal to or greater than K have palindromic complete return words. Does there exist an integer N such that $T_{\mathbf{u}}(n) = 0$ for any $n \geq N$?

Problem 3. Let $\mathbf{u}$ be an infinite word with language closed under reversal and containing infinitely many palindromes. Suppose that there exists an integer N such that for all $n \geq N$ the equality $T_{\mathbf{u}}(n) = 0$ holds. Does there exist also an integer H such that for any factor $f \in \mathcal{L}(\mathbf{u})$ of length $|f| \geq H$ the longest palindromic suffix of f is unioccurrent in f?

We have seen that in the proof of the validity of the Brlek–Reutenauer conjecture for uniformly recurrent words, an important role was played by the presence of long squares in such words. This leads to the last open problem.

Problem 4. Find other classes of infinite words containing for any L a factor w such that $|w| > L$ and ww belongs to the language.

Acknowledgements

We acknowledge financial support by the Czech Science Foundation grant 201/09/0584, by the grants MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic, and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS11/162/OHK4/3T/14.

References

[1] P. Baláži, Z. Masáková, E. Pelantová, Factor versus palindromic complexity of uniformly recurrent infinite words, Theoret. Comput. Sci. 380 (2007) 266–275.
[2] L' Balková, E. Pelantová, Š. Starosta, Infinite words with finite defect, Adv. Appl. Math. (2011) doi:10.1016/j.aam.2010.11.006, the original publication is available at http://www.sciencedirect.com/science/journal/01968858.
[3] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, On the palindromic complexity of infinite words, in: J. Berstel, J. Karhumäki, D. Perrin (Eds.), Combinatorics on Words with Applications, Int. J. Found. Comput. Sci. 15 (2) (2004), 293–306.
[4] S. Brlek, C. Reutenauer, Complexity and palindromic defect of infinite words, Theoret. Comput. Sci. 412 (4–5) (2011) 493–497.
[5] M. Bucci, A. De Luca, A. Glen, L.Q. Zamboni, A connection between palindromic and factor complexity using return words, Adv. in Appl. Math 42 (2009) 60–74.
[6] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theoret. Comput. Sci. 255 (2001) 539–553.
[7] A. Glen, J. Justin, S. Widmer, L.Q. Zamboni, Palindromic richness, European J. Combin 30 (2009) 510–531.

Proof of the Brlek–Reutenauer Conjecture

Authors: **Ľubomíra Balková, Edita Pelantová, Štěpán Starosta**
Journal: Theoretical Computer Science
Year: 2013
Volume: 475
Pages: 120–125

Theoretical Computer Science 475 (2013) 120–125

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Note

Proof of the Brlek–Reutenauer conjecture

L'. Balková [a,*], E. Pelantová [a], Š. Starosta [b]

[a] Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic

[b] Department of Applied Mathematics, FIT, Czech Technical University in Prague, Thákurova 9, 160 00 Praha 6, Czech Republic

ARTICLE INFO

Article history:
Received 13 August 2012
Accepted 26 December 2012
Communicated by D. Perrin

ABSTRACT

Brlek and Reutenauer conjectured that any infinite word $\mathbf{u}$ with language closed under reversal satisfies the equality $2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ in which $D(\mathbf{u})$ denotes the defect of $\mathbf{u}$ and $T_{\mathbf{u}}(n)$ denotes $\mathcal{C}_{\mathbf{u}}(n+1)-\mathcal{C}_{\mathbf{u}}(n)+2-\mathcal{P}_{\mathbf{u}}(n+1)-\mathcal{P}_{\mathbf{u}}(n)$, where $\mathcal{C}_{\mathbf{u}}$ and $\mathcal{P}_{\mathbf{u}}$ are the factor and palindromic complexity of $\mathbf{u}$, respectively. This conjecture was verified for periodic words by Brlek and Reutenauer themselves. Using their results for periodic words, we have recently proved the conjecture for uniformly recurrent words. In the present article we prove the conjecture in its general version by a new method without exploiting the result for periodic words.

1. Introduction

Brlek and Reutenauer conjectured in [6] a nice equality which combines together the factor complexity $\mathcal{C}_{\mathbf{u}}$, the palindromic complexity $\mathcal{P}_{\mathbf{u}}$, and the palindromic defect $D(\mathbf{u})$ of an infinite word $\mathbf{u}$. It sounds as follows.

Brlek–Reutenauer Conjecture. *If $\mathbf{u}$ is an infinite word with language closed under reversal, then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n),$$

where $T_{\mathbf{u}}(n) = \mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) + 2 - \mathcal{P}_{\mathbf{u}}(n+1) - \mathcal{P}_{\mathbf{u}}(n)$.

Brlek and Reutenauer proved ibidem that their conjecture holds for periodic infinite words. It is known from [7] that the Brlek–Reutenauer conjecture holds for words with zero defect. In [3], we proved the conjecture for uniformly recurrent words. In our proof, we constructed for any uniformly recurrent word $\mathbf{u}$ whose language is closed under reversal a periodic word $\mathbf{v}$ with language closed under reversal such that $D(\mathbf{u}) = D(\mathbf{v})$ and $T_{\mathbf{u}}(n) = T_{\mathbf{v}}(n)$ for any n. Then we used validity of the conjecture for periodic words.

In this paper, we will prove that the Brlek–Reutenauer conjecture holds in full generality without exploiting the result for periodic words. Since both sides of the equality in the Brlek–Reutenauer conjecture are non-negative, validity of the conjecture will be shown if we prove the following two theorems.

Theorem 1. *If $\mathbf{u}$ is an infinite word with language closed under reversal such that both $D(\mathbf{u})$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ are finite, then*

$$2D(\mathbf{u}) = \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n). \qquad (1)$$

[*] Corresponding author. Tel.: +420 224358544.
E-mail address: lubomira.balkova@gmail.com (L. Balková).

0304-3975/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.12.024

Theorem 2. *If* $\mathbf{u}$ *is an infinite word with language closed under reversal, then*

$$D(\mathbf{u}) < +\infty \quad \textit{if and only if} \quad \sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty.$$

In [3], which is devoted mainly to the uniformly recurrent words, we already stated in the section Open problems one part of Theorem 2, namely that $D(\mathbf{u}) < +\infty$ implies $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty$. As pointed out in [4], there is a gap in our proof, and its corrected version can be found in [2]. In order to make the present paper self-sustained so that the reader understands and checks all steps of the proof without having all previous papers at hand, we recall necessary notations and statements together with the proofs of the essential ones.

2. Preliminaries

By $\mathcal{A}$ we denote a finite set of symbols called *letters*; the set $\mathcal{A}$ is therefore called an *alphabet*. A finite string $w = w_0w_1 \dots w_{n-1}$ of letters from $\mathcal{A}$ is said to be a *finite word*, its length is denoted by $|w| = n$. Finite words over $\mathcal{A}$ together with the operation of concatenation and the empty word ϵ as the neutral element form a free monoid $\mathcal{A}^*$. The map

$$w = w_0w_1 \dots w_{n-1} \quad \mapsto \quad \overline{w} = w_{n-1}w_{n-2} \dots w_0$$

is a bijection on $\mathcal{A}^*$, the word $\overline{w}$ is called the *reversal* or the *mirror image* of w. A word w which coincides with its mirror image is a *palindrome*.

Under an *infinite word* we understand an infinite string $\mathbf{u} = u_0u_1u_2 \dots$ of letters from $\mathcal{A}$. A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \epsilon$, then w is said to be a *prefix* of v, if $s = \epsilon$, then w is a *suffix* of v.

The *language* $\mathcal{L}(v)$ of a finite or an infinite word v is the set of all its factors. Factors of v of length n form the set denoted by $\mathcal{L}_n(v)$. We say that the language of an infinite word $\mathbf{u}$ is *closed under reversal* if $\mathcal{L}(\mathbf{u})$ contains with every factor w also its reversal $\overline{w}$.

For any factor $w \in \mathcal{L}(\mathbf{u})$, there exists an index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2} \dots$. Such an index is called an *occurrence* of w in $\mathbf{u}$. If each factor of $\mathbf{u}$ has infinitely many occurrences in $\mathbf{u}$, the infinite word $\mathbf{u}$ is said to be *recurrent*. It is easy to see that if the language of $\mathbf{u}$ is closed under reversal, then $\mathbf{u}$ is recurrent (a proof can be found in [9]). For a recurrent infinite word $\mathbf{u}$, we may define the notion of a *complete return word* of any $w \in \mathcal{L}(\mathbf{u})$. It is a factor $v \in \mathcal{L}(\mathbf{u})$ such that w is a prefix and a suffix of v and w occurs in v exactly twice.

If any factor $w \in \mathcal{L}(\mathbf{u})$ has only finitely many complete return words, then the infinite word $\mathbf{u}$ is called *uniformly recurrent*.

The *factor complexity* of an infinite word $\mathbf{u}$ is the map $\mathcal{C}_{\mathbf{u}} : \mathbb{N} \mapsto \mathbb{N}$ defined by the prescription $\mathcal{C}_{\mathbf{u}}(n) := \#\mathcal{L}_n(\mathbf{u})$. To determine the first difference of the factor complexity, one has to count the possible extensions of factors of length n. A *right extension* of $w \in \mathcal{L}(\mathbf{u})$ is a letter $a \in \mathcal{A}$ such that $wa \in \mathcal{L}(\mathbf{u})$. Of course, any factor of $\mathbf{u}$ has at least one right extension. A factor w is called *right special* if w has at least two right extensions. Similarly, one can define a *left extension* and a *left special* factor. We will deal mainly with recurrent infinite words $\mathbf{u}$. In such a case, any factor of $\mathbf{u}$ has at least one left extension.

In [8] it is shown that any finite word w contains at most $|w| + 1$ distinct palindromes (including the empty word). The *defect* $D(w)$ of a finite word w is the difference between the utmost number of palindromes $|w| + 1$ and the actual number of palindromes contained in w.

In accordance with the terminology introduced in [8], the factor with a unique occurrence in another factor is called *unioccurrent*.

The following corollary gives an insight into the birth of defects.

Corollary 3 *([8]). The defect* $D(w)$ *of a finite word* w *is equal to the number of prefixes* w' *of* w *for which the longest palindromic suffix of* w' *is not unioccurrent in* w'*. In other words, if* b *is a letter and* w *a finite word, then* $D(wb) = D(w) + \delta$*, where* $\delta = 0$ *if the longest palindromic suffix of* wb *occurs exactly once in* wb *and* $\delta = 1$ *otherwise.*

Corollary 3 implies that $D(v) \geq D(w)$ whenever w is a factor of v. It enables to give a reasonable definition of the defect of an infinite word (see [5]).

Definition 4. The defect of an infinite word $\mathbf{u}$ is the number (finite or infinite)

$$D(\mathbf{u}) = \sup\{D(w) \colon w \text{ is a prefix of } \mathbf{u}\}.$$

Let us point out two facts.

(1) If we consider all factors of a finite or an infinite word $\mathbf{u}$, we obtain the same defect, i.e.,

$$D(\mathbf{u}) = \sup\{D(w) \colon w \in \mathcal{L}(\mathbf{u})\}.$$

(2) Any infinite word with finite defect contains infinitely many palindromes.

Using Corollary 3 and Definition 4, we obtain immediately the following corollary.

Corollary 5. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal. The following statements are equivalent.*

(1) *The defect of* $\mathbf{u}$ *is finite.*
(2) *There exists an integer* H *such that the longest palindromic suffix of any prefix* w *of length* $|w| \geq H$ *occurs in* w *exactly once.*

For *the longest palindromic suffix* of a word w we will sometimes use the notation $lps(w)$.

The number of palindromes of a fixed length occurring in an infinite word is measured by the so called *palindromic complexity* $\mathcal{P}_{\mathbf{u}}$, the map which assigns to any non-negative integer n the number

$$\mathcal{P}_{\mathbf{u}}(n) := \#\{w \in \mathcal{L}_n(u) \colon w \text{ is a palindrome}\}.$$

Denote

$$T_{\mathbf{u}}(n) = \mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) + 2 - \mathcal{P}_{\mathbf{u}}(n+1) - \mathcal{P}_{\mathbf{u}}(n).$$

The following proposition is proven in [1] for uniformly recurrent words; however, as also noted in [6], the uniform recurrence is not needed in the proof and it holds for any infinite word with language closed under reversal.

Proposition 6 (*[1]*). *If* $\mathbf{u}$ *is an infinite word with language closed under reversal, then*

$$T_{\mathbf{u}}(n) \geq 0 \quad \text{for all } n \in \mathbb{N}. \tag{2}$$

Let $\mathbf{u}$ be an infinite word with language closed under reversal. Using the proof of Proposition 6, those $n \in \mathbb{N}$ for which $T_{\mathbf{u}}(n) = 0$ can be characterized in the graph language. Before doing that we need to introduce some more notions.

An *n-simple path e* is a factor of $\mathbf{u}$ of length at least $n+1$ such that the only special (right or left) factors of length n occurring in e are its prefix and suffix of length n. If w is the prefix of e of length n and v is the suffix of e of length n, we say that the n-simple path e starts in w and ends in v. We will denote by $G_n(\mathbf{u})$ an undirected graph whose set of vertices is formed by unordered pairs $(w, \overline{w})$ such that $w \in \mathcal{L}_n(\mathbf{u})$ is right or left special. We connect two vertices $(w, \overline{w})$ and $(v, \overline{v})$ by an unordered pair $(e, \overline{e})$ if e or $\overline{e}$ is an n-simple path starting in w or $\overline{w}$ and ending in v or $\overline{v}$. Note that the graph $G_n(\mathbf{u})$ may have multiple edges and loops.

Lemma 7. *If* $\mathbf{u}$ *is an infinite word with language closed under reversal and* $n \in \mathbb{N}$, *then* $T_{\mathbf{u}}(n) = 0$ *if and only if both of the following conditions are met.*

(1) *The graph obtained from* $G_n(\mathbf{u})$ *by removing loops is a tree.*
(2) *Any n-simple path forming a loop in the graph* $G_n(\mathbf{u})$ *is a palindrome.*

Proof. It is a direct consequence of the proof of Theorem 1.2 in [1] (recalled in this paper as Proposition 6). □

3. Proof of Theorem 1

The aim of this section is to prove Theorem 1, i.e., to prove the Brlek–Reutenauer conjecture under the additional assumption that the defect $D(\mathbf{u})$ of an infinite word $\mathbf{u}$ and the sum $\sum_{n=0}^{\infty} T_{\mathbf{u}}(n)$ are finite. As observed in [6], it is easy to prove the "finite analogy" of the conjecture, which deals only with finite words. We will also make use of this result.

Theorem 8 (*[6]*). *For every finite word* w *we have*

$$2D(w) = \sum_{n=0}^{|w|} T_w(n),$$

where $T_w(n) = \mathcal{C}_w(n+1) - \mathcal{C}_w(n) + 2 - \mathcal{P}_w(n+1) - \mathcal{P}_w(n)$ *and the index* w *means that we consider only factors of* w.

It may seem that the Brlek–Reutenauer conjecture for an infinite word $\mathbf{u}$ can be obtained from Theorem 8 by a "limit transition". However, this transition would be far from being kosher. The following lemmas enable us to avoid the incorrectness.

Lemma 9. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal and finite defect. If* q *is its prefix satisfying* $D(\mathbf{u}) = D(q)$, *then for* $H = |q| + 1$ *one has*

$$\mathcal{C}_{\mathbf{u}}(H) - \mathcal{P}_{\mathbf{u}}(H) = 2\#\{x \in \mathcal{L}(\mathbf{u}) \colon x \text{ is a palindrome shorter than } H \text{ which is not contained in } q\}.$$

Proof. Let us define a mapping $f : S \to T$, where

$$S = \{x \in \mathcal{L}(\mathbf{u}) \colon x \notin \mathcal{L}(q),\ |x| < H,\ x = \overline{x}\}$$

and

$$T = \left\{\{w, \overline{w}\} \colon w \in \mathcal{L}_H(\mathbf{u}),\ w \neq \overline{w}\right\}.$$

Let x be a palindrome from S and i be the first occurrence of x in $\mathbf{u}$. Put $w = u_{i+|x|-H} \cdots u_{i+|x|-1}$. It means that w is a factor of $\mathbf{u}$ of length H and x is a suffix of w. Since $H > |x|$, the factor w is not a palindrome — otherwise it contradicts the fact that i is the first occurrence of the palindrome x. We put $f(x) = \{w, \overline{w}\}$.

To show that f is surjective, we consider $w \in \mathcal{L}_H(\mathbf{u})$ such that $w \neq \overline{w}$. Let p be the prefix of $\mathbf{u}$ which ends in the first occurrence of w or $\overline{w}$ in $\mathbf{u}$. Since $|p| \geq H = |w| > |q|$, we have according to Corollary 3 that $D(q) = D(p)$ and consequently, $lps(p)$ is unioccurrent in p, which implies that $lps(p)$ is not a factor of q. Moreover, $lps(p)$ is shorter than H — otherwise it contradicts the choice of the prefix p. We found $x = lps(p) \in S$ such that $f(x) = \{w, \overline{w}\}$, i.e., f is surjective.

To show that f is injective, we consider two palindromes $y, z \in S$ and we denote $f(y) = \{w_y, \overline{w_y}\}$ and $f(z) = \{w_z, \overline{w_z}\}$. From the definition of w_x we know that the palindrome x occurs as a factor of w_x exactly once, namely as its suffix. It means that x equals $lps(w_x)$. Let us suppose that $f(y) = f(z)$. We have to discuss two cases.

(1) Case $w_y = w_z$. It gives $lps(w_y) = lps(w_z)$ and thus $y = z$.
(2) Case $w_y = \overline{w_z}$. It implies that y is a prefix of w_z and z is a prefix of w_y. The fact that y is a prefix of w_z forces the first occurrence of w_y to be strictly smaller than the first occurrence of w_z. Simultaneously, since z is a prefix of w_y, the first occurrence of w_z is strictly smaller than the first occurrence of w_y - a contradiction.

Consequently, the assumption $f(y) = f(z)$ implies $z = y$ and the mapping f is injective as well.

Existence of the bijection f between the finite sets T and S means $\#T = \#S$. Since from the definition of T it follows that $\mathcal{C}_{\mathbf{u}}(H) - \mathcal{P}_{\mathbf{u}}(H) = 2\#T$, the equality stated in the lemma is proven. □

Remark 10. As it was pointed out by Bojan Bašić, Lemma 9 may be stated in a more general form for $H > |q|$, then the equality changes to

$$\mathcal{C}_{\mathbf{u}}(H) - \mathcal{P}_{\mathbf{u}}(H) = 2\#\{x \in \mathcal{L}(\mathbf{u}) : x \notin \mathcal{L}(q),\ |x| < H,\ x = \overline{x}\} - 2(H - |q| - 1).$$

Thanks to him, we added the assumption $H = |q| + 1$ in Lemma 9 necessary for the validity of the statement.

Lemma 11. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal and finite defect. If* q *is its prefix satisfying* $D(\mathbf{u}) = D(q)$, *then for any prefix* p *of* $\mathbf{u}$ *such that* $|p| > |q|$ *the number*

$$\#\{x \in \mathcal{L}(p) : x \text{ is a palindrome of length at most } |q| \text{ which is not contained in } q\} + \sum_{n=|q|+1}^{|p|} \mathcal{P}_p(n)$$

equals $|p| - |q|$.

Proof. At first we will show the equality

$$|p| - |q| = \#\{x \in \mathcal{L}(p) \setminus \mathcal{L}(q) : x = \overline{x}\}. \tag{3}$$

Let us denote by $u^{(i)}$ the prefix of $\mathbf{u}$ of length i. For any palindrome $x \in \mathcal{L}(p) \setminus \mathcal{L}(q)$ we find the minimal index i such that x occurs in $u^{(i)}$. Since $x \in \mathcal{L}(p) \setminus \mathcal{L}(q)$, we have $|q| < i \leq |p|$. Thus we map any element of $\{x \in \mathcal{L}(p) \setminus \mathcal{L}(q) : x = \overline{x}\}$ to an index $i \in \{|q| + 1, |q| + 2, \dots, |p|\}$.

Let us look at the details of this mapping. The minimality of i guarantees that x is unioccurrent in $u^{(i)}$. Palindromicity of x gives that $x = lps(u^{(i)})$. It implies that no two different palindromes are mapped to the same index i, i.e., the mapping is injective.

Since $D(q) = D(\mathbf{u})$, according to Corollary 3, $lps(u^{(i)})$ is unioccurrent in $u^{(i)}$ and thus $lps(u^{(i)}) \notin \mathcal{L}(q)$. Thus any index i such that $|q| < i \leq |p|$ has its preimage $x = lps(u^{(i)})$. Therefore the mapping is a bijection and its domain and range have the same cardinality as stated in (3).

To finish the proof, we split elements of $\{x \in \mathcal{L}(p) \setminus \mathcal{L}(q) \ : \ x = \overline{x}\}$ into two disjoint parts: elements of length smaller than or equal to $|q|$ and elements of length greater than $|q|$. Since

$$\#\{x \in \mathcal{L}(p) \setminus \mathcal{L}(q) \ : \ x = \overline{x},\ |x| > |q|\} = \#\{x \in \mathcal{L}(p) \ : \ x = \overline{x},\ |x| > |q|\} = \sum_{n=|q|+1}^{|p|} \mathcal{P}_p(n)\,,$$

the statement of Lemma 11 is proven. □

Now we can complete the proof of Theorem 1.

Proof of Theorem 1. Finiteness of defect means that there exists a constant $L \in \mathbb{N}$ such that $D(\mathbf{u}) = D(q)$ for any prefix q of $\mathbf{u}$ which is longer than or of length equal to L. On the other hand, finiteness of the sum $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n)$ together with the fact $0 \leq T_{\mathbf{u}}(n) \in \mathbb{Z}$ for any $n \in \mathbb{N}$ implies that there exists a constant $M \in \mathbb{N}$ such that $T_{\mathbf{u}}(n) = 0$ for any $n > M$. Let us fix an integer $H > \max\{L, M\}$ and denote by q the prefix of $\mathbf{u}$ of length $|q| = H - 1$. Consequently,

$$T_{\mathbf{u}}(n) = 0 \quad \text{for any } n \geq H \quad \text{and} \quad D(\mathbf{u}) = D(q).$$

In order to show the equality (1), it thus remains to show $2D(q) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n)$.

Let us consider a prefix p of $\mathbf{u}$ containing all factors of length H. In this case p is longer than q, thus it holds by Corollary 3 that $D(q) = D(p)$. Using Theorem 8, we have

$$2D(p) = \sum_{n=0}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_p(n) + \sum_{n=H}^{|p|} T_p(n) = \sum_{n=0}^{H-1} T_{\mathbf{u}}(n) + \sum_{n=H}^{|p|} T_p(n),$$

where the last equality is due to the fact that p contains all factors of length H. It remains to prove that $\sum_{n=H}^{|p|} T_p(n) = 0$. Let us rewrite the sum by definition.

$$\begin{aligned}\sum_{n=H}^{|p|} T_p(n) &= \sum_{n=H}^{|p|} \big(\mathcal{C}_p(n+1) - \mathcal{C}_p(n) + 2 - \mathcal{P}_p(n+1) - \mathcal{P}_p(n)\big)\\ &= -\mathcal{C}_p(H) + 2(|p| - H + 1) - 2\sum_{n=H}^{|p|} \mathcal{P}_p(n) + \mathcal{P}_p(H)\\ &= -\mathcal{C}_{\mathbf{u}}(H) + 2(|p| - H + 1) - 2\sum_{n=H}^{|p|} \mathcal{P}_p(n) + \mathcal{P}_{\mathbf{u}}(H), \qquad (4)\end{aligned}$$

where in the last equality we again used the fact that p contains all factors of length H. This fact also allows us to rewrite the set $\{x \in \mathcal{L}(p) : x \notin \mathcal{L}(q), x = \overline{x}, |x| \leq |q|\}$ from Lemma 11 as $\{x \in \mathcal{L}(\mathbf{u}) : x \notin \mathcal{L}(q), x = \overline{x}, |x| < H\}$. Denote the cardinality of this set by B.

In this notation, Lemmas 9 and 11 say

$$\mathcal{C}_{\mathbf{u}}(H) - \mathcal{P}_{\mathbf{u}}(H) = 2B \quad \text{and} \quad B + \sum_{n=H}^{|p|} \mathcal{P}_p(n) = |p| - H + 1.$$

This implies that the last expression in (4) is zero as desired. □

4. Proof of Theorem 2

If an infinite word $\mathbf{u}$ is periodic with language closed under reversal, then $D(\mathbf{u}) < +\infty$ and $\sum_{n=0}^{+\infty} T_{\mathbf{u}}(n) < +\infty$, as shown in [6]. Consequently, we will limit our considerations in the sequel to aperiodic words.

Proposition 12. *If $\mathbf{u}$ is an aperiodic infinite word with language closed under reversal and N is an integer, then $T_{\mathbf{u}}(n) = 0$ for all $n \geq N$ if and only if for any factor w such that $|w| \geq N$, any factor longer than w beginning in w or $\overline{w}$ and ending in w or $\overline{w}$, with no other occurrences of w or $\overline{w}$, is a palindrome.*

Proof. ($\Leftarrow$): Let us show for any $n \geq N$ that the assumptions of Lemma 7 are satisfied. We have to show two properties of $G_n(\mathbf{u})$ for any $n \geq N$.

(1) Any loop in $G_n(\mathbf{u})$ is a palindrome.
Since any loop e in $G_n(\mathbf{u})$ at a vertex $(w, \overline{w})$ is a word longer than w beginning in a special factor w or $\overline{w}$ and ending in w or $\overline{w}$, with no other occurrences of w or $\overline{w}$, the loop e is a palindrome by the assumption.

(2) The graph obtained from $G_n(\mathbf{u})$ by removing loops is a tree.
Or equivalently, we have to show that in $G_n(\mathbf{u})$ there exists a unique path between any two different vertices $(w', \overline{w'})$ and $(w'', \overline{w''})$. Let p be a factor of $\mathbf{u}$ such that w' or $\overline{w'}$ is its prefix, w'' or $\overline{w''}$ is its suffix and p has no other occurrences of w', $\overline{w'}$, w'', $\overline{w''}$. Let v be a factor starting in p, ending in w' or $\overline{w'}$ and containing no other occurrences of w' or $\overline{w'}$. By the assumption the factor v is a palindrome, thus $\overline{p}$ is a suffix of v. It is then a direct consequence of the construction of v that the next factor with the same properties as p, i.e., representing a path in the undirected graph $G_n(\mathbf{u})$ between w' and w'', which occurs in $\mathbf{u}$ after p, is $\overline{p}$. This shows that there is only one such path.

Consequently, Lemma 7 implies that $T_n(\mathbf{u}) = 0$ for any $n \geq N$.
($\Rightarrow$): First we prove an auxiliary claim.

Claim. *If $\mathbf{u}$ is an aperiodic infinite word with language closed under reversal and N is an integer such that $T_{\mathbf{u}}(n) = 0$ for all $n \geq N$, then for any w such that $|w| \geq N$ and any factor v longer than w beginning in w and ending in w or $\overline{w}$, with no other occurrences of w or $\overline{w}$, there exists a letter $a \in \mathcal{A}$ such that v has prefix wa and suffix $a\overline{w}$.*

It is clear that repeated application of the previous claim to factors w of length gradually increased by one gives the proof of implication ($\Rightarrow$) of Proposition 12.

We split the proof of the auxiliary claim into two cases.

- Case 1: Assume that w is a special factor.
  If v does not contain any other special factor of length $n = |w|$ except for w and $\overline{w}$, then v is a loop in the graph $G_n(\mathbf{u})$ and according to Lemma 7, the factor v is a palindrome. Necessarily, v begins in wa for some letter a and ends in $a\overline{w}$.
  Suppose now that $v = v_0v_1 \cdots v_m$ contains a special factor $z \neq w, \overline{w}$ of length n at the position i, i.e., $z = v_iv_{i+1} \cdots v_{n+i-1}$. Without loss of generality, we consider the smallest index i with this property. The pair $(z, \overline{z})$ is a vertex in the graph $G_n(\mathbf{u})$ and a prefix of v, say e, corresponds to an edge in $G_n(\mathbf{u})$ starting in $(w, \overline{w})$ and ending in $(z, \overline{z})$. Since the graph $G_n(\mathbf{u})$ is a tree, the word v which corresponds to a walk from $(w, \overline{w})$ to the same vertex $(w, \overline{w})$ has a suffix f representing an edge in $G_n(\mathbf{u})$ connecting again vertices $(z, \overline{z})$ and $(w, \overline{w})$. It means that the suffix f starts in z or $\overline{z}$ and ends in w or $\overline{w}$. Since $G_n(\mathbf{u})$ has no multiple edges connecting distinct vertices, necessarily $f = \overline{e}$, which already gives the claim.

- Case 2: Assume w is not a special factor.

  It means that there exists a unique letter a such that wa belongs to the language of $\mathbf{u}$. As the language is closed under reversal, the factor $\overline{w}$ has a unique left extension, namely a. If v starts in w and ends in $\overline{w}$, then the claim is proven.

  It remains to exclude that v begins and ends in a non-palindromic factor w. Suppose this situation happens. In this case, there exists a unique q such that wq is a right special factor and it is the shortest right special factor having the prefix w. The factor wq has only one occurrence of the factor w — otherwise we can find a shorter prolongation of w which is right special. Since w is a suffix of v, we deduce that $|wq| < |v|$. Because wq is the shortest right special factor with prefix w, the factor vq belongs to the language and its prefix and suffix wq is a special factor. According to already proven Case 1, we have $wq = \overline{wq} = \overline{q}\,\overline{w}$. It means together with the inequality $|wq| < |v|$ that $\overline{w}$ is contained in v as well — a contradiction. □

The proof of the implication (⇒) of Proposition 12 is taken from [10], where we showed a more general statement for an infinite word whose language is closed under a larger group of symmetries.

Corollary 13. *Let* $\mathbf{u}$ *be an aperiodic infinite word with language closed under reversal and let N be an integer. If* $T_{\mathbf{u}}(n) = 0$ *for all* $n \geq N$*, then the occurrences of* w *and* $\overline{w}$ *in* $\mathbf{u}$ *alternate for any factor* w *of* $\mathbf{u}$ *of length at least N.*

The following lemma builds a bridge between Corollary 5 and Proposition 12.

Lemma 14. *Let* $\mathbf{u}$ *be an aperiodic infinite word with language closed under reversal. There exists* $H \in \mathbb{N}$ *such that the longest palindromic suffix of any prefix* w *of* $\mathbf{u}$ *of length* $|w| \geq H$ *occurs in* w *exactly once if and only if there exists* $N \in \mathbb{N}$ *such that for any factor* w *with* $|w| \geq N$*, any factor longer than* w *beginning in* w *or* $\overline{w}$ *and ending in* w *or* $\overline{w}$*, with no other occurrences of* w *or* $\overline{w}$*, is a palindrome.*

Proof. (⇒): We will show that N may be set equal to H. Let us proceed by contradiction. Suppose there exists a factor $w \in \mathcal{L}(\mathbf{u})$ such that $|w| \geq H$ and there exists a non-palindromic factor of $\mathbf{u}$ longer than w beginning in w or $\overline{w}$ and ending in w or $\overline{w}$, with no other occurrences of w or $\overline{w}$. Let us find the first non-palindromic factor of the above form in $\mathbf{u}$ and let us denote it as r. Let p be the prefix of $\mathbf{u}$ ending in the first occurrence of r in $\mathbf{u}$, i.e., $p = tr$ for some word t and r is unioccurrent in p. Denote by s the longest palindromic suffix of p. By the assumption, s is unioccurrent in p. No matter how long the suffix s is, we will obtain a contradiction.

(1) If $|s| \leq |w|$, then we have a contradiction to the unioccurrence of s.
(2) If $|r| > |s| > |w|$, then we can find at least 3 occurrences of w or $\overline{w}$ in r which is a contradiction to the form of r.
(3) The equality $|r| = |s|$ contradicts the fact that we supposed r to be non-palindromic.
(4) Finally, if $|r| < |s|$, then there is an occurrence of the mirror image of r which is a non-palindromic factor having the same properties as r which occurs before r and contradicts the choice of p.

(⇐): Take a prefix containing all factors of length N. Set H equal to its length. Let us show that any prefix p of length greater than or equal to H has $lps(p)$ of length greater than or equal to N. Consider a suffix of p of length N, say w. Either w is a palindrome, then $lps(p)$ is of length greater than or equal to N. Or w is not a palindrome, then we find a suffix of p beginning in $\overline{w}$ and containing exactly two occurrences of w or $\overline{w}$. Such a suffix exists since all factors of length N are contained in p. By assumptions, such a suffix is a palindrome, hence $lps(p)$ is longer than N.

Any prefix p of $\mathbf{u}$ of length greater than or equal to H has $lps(p)$ unioccurrent. Assume there are more occurrences of $lps(p)$ in p and consider its suffix v starting in the last-but-one occurrence of $lps(p)$. Since the length of $lps(p)$ is greater than or equal to N, the factor v is a palindrome by assumptions, which contradicts the choice of $lps(p)$. □

Proof of Theorem 2. For periodic words, the statement was shown in [6]. If $\mathbf{u}$ is aperiodic, then the statement is a direct consequence of Lemma 14, Corollary 5, and Proposition 12. □

Acknowledgments

This work was supported by the Czech Science Foundation grant GAČR 201/09/0584 and by the grant MSM6840770039 of the Ministry of Education, Youth, and Sports of the Czech Republic.

We would like to thank Bojan Bašić for his careful reading, useful remarks and in particular for his correction of a lemma.

References

[1] P. Baláži, Z. Masáková, E. Pelantová, Factor versus palindromic complexity of uniformly recurrent infinite words, Theoret. Comput. Sci. 380 (2007) 266–275.
[2] L'. Balková, E. Pelantová, Š. Starosta, Corrigendum: "On Brlek–Reutenauer conjecture", Theoret. Comput. Sci. 465 (2012) 73–74.
[3] L'. Balková, E. Pelantová, Š. Starosta, On Brlek-Reutenauer conjecture, Theoret. Comput. Sci. 412 (2011) 5649–5655.
[4] B. Bašić, A note on the paper On Brlek-Reutenauer conjecture, Theoret. Comput. Sci. 448 (2012) 94–96.
[5] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, On the palindromic complexity of infinite words, Internat. J. Found. Comput. 15 (2004) 293–306.
[6] S. Brlek, C. Reutenauer, Complexity and palindromic defect of infinite words, Theoret. Comput. Sci. 412 (2011) 493–497.
[7] M. Bucci, A. De Luca, A. Glen, L.Q. Zamboni, A connection between palindromic and factor complexity using return words, Adv. in Appl. Math. 42 (2009) 60–74.
[8] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theoret. Comput. Sci. 255 (2001) 539–553.
[9] A. Glen, J. Justin, S. Widmer, L.Q. Zamboni, Palindromic richness, European J. Combin. 30 (2009) 510–531.
[10] E. Pelantová, Š Starosta, Palindromic richness and Coxeter groups, preprint available at http://arxiv.org/abs/1108.3042.

Infinite Words with Well Distributed Occurrences

Authors:	**L'ubomíra Balková, Michelangelo Bucci, Alessandro De Luca, Svetlana Puzynina**
Journal:	In: J. Karhumäki, A. Lepistö, L. Zamboni (Eds.), Combinatorics on Words, LNCS, Springer
Year:	2013
Volume:	8079
Pages:	46–57

Infinite Words with Well Distributed Occurrences

Ľubomíra Balková[1], Michelangelo Bucci[2], Alessandro De Luca[3], and Svetlana Puzynina[2,4]

[1] Department of Mathematics, FNSPE, Czech Technical University in Prague,
Trojanova 13, 120 00 Praha 2, Czech Republic
lubomira.balkova@gmail.com
[2] Department of Mathematics, University of Turku, FI-20014 Turku, Finland
{michelangelo.bucci, svepuz}@utu.fi
[3] DIETI, Università degli Studi di Napoli Federico II
via Claudio, 21, 80125 Napoli, Italy
alessandro.deluca@unina.it
[4] Sobolev Institute of Mathematics, Russia

Abstract. In this paper we introduce the *well distributed occurrences (WDO)* combinatorial property for infinite words, which guarantees good behavior (no lattice structure) in some related pseudorandom number generators. An infinite word u on a d-ary alphabet has the WDO property if, for each factor w of u, positive integer m, and vector $\mathbf{v} \in \mathbb{Z}_m^d$, there is an occurrence of w such that the Parikh vector of the prefix of u preceding such occurrence is congruent to $\mathbf{v}$ modulo m. We prove that Sturmian words, and more generally Arnoux-Rauzy words and some morphic images of them, have the WDO property.

Introduction

The combinatorial problem studied in this paper comes from random number generation. Pseudorandom number generators aim to produce random numbers using a deterministic process. No wonder they suffer from many defects. The most usual ones – linear congruential generators – are known to produce periodic sequences having a defect called lattice structure. Guimond et al. [2] proved that when two linear congruential generators are combined using infinite words coding certain classes of quasicrystals or, equivalently, of cut-and-project sets, the resulting sequence is aperiodic and has no lattice structure.

We have found a combinatorial condition – *well distributed occurrences*, or WDO for short – that guarantees absence of lattice structure if two arbitrary generators having the same output alphabet are combined using an infinite word having the WDO property. The WDO property for an infinite word u over an alphabet A means that for any integer m and any factor w of u, the set of Parikh vectors modulo m of prefixes of u preceeding the occurrences of w coincides with $\{0, 1, \ldots, m-1\}^{|A|}$ (see Definition 2.1). In other words, among Parikh vectors modulo m of such prefixes one has all possible vectors. Besides giving generators without lattice structure, the WDO property is an interesting combinatorial property of infinite words itself.

We have proved first that Sturmian words have well distributed occurrences, and then we have shown this property for Arnoux-Rauzy words. The proof for Sturmian words is based on different ideas than the one for Arnoux-Rauzy words, therefore we will provide in the sequel both of them.

In the next section, we deal with pseudorandom number generation, thus establishing the motivation for our work. Next, in Section 2, we give the basic combinatorial definitions needed for our main results, including the WDO property. Finally, in the last two sections, we prove that the property holds for Sturmian and Arnoux-Rauzy words, respectively.

1 Motivation in Pseudorandom Number Generation

For the sake of our discussion, any infinite sequence of integers can be understood as a *pseudorandom number generator (PRNG)*; see also [2].

Let $X = (x_n)_{n\in\mathbb{N}}$ and $Y = (y_n)_{n\in\mathbb{N}}$ be two PRNGs with the same output $M \subset \mathbb{N}$ and the same period $m \in \mathbb{N}$, and let $u = u_0u_1u_2\ldots$ be a binary infinite word, i.e., an infinite sequence over $\{0, 1\}$.

The PRNG

$$Z = (z_n)_{n\in\mathbb{N}} \tag{1}$$

based on u is obtained by the following algorithm:

1. Read step by step the letters of u.
2. When you read 0 for the i-th time, copy the i-th symbol from X to the end of the constructed sequence Z.
3. When you read 1 for the i-th time, copy the i-th symbol from Y to the end of the constructed sequence Z.

Of course, it is possible to generalize this construction – using infinite words over a multiliteral alphabet, one can combine more than two PRNGs.

1.1 Lattice Structure

Let $X = (x_n)_{n\in\mathbb{N}}$ be a PRNG whose output is a finite set $M \subset \mathbb{N}$. We say that X has the *lattice structure* if there exists $t \in \mathbb{N}$ such that

$$\{(x_i, x_{i+1}, \ldots, x_{i+t-1}) \mid i \in \mathbb{N}\}$$

is covered by a family of parallel equidistant hyperplanes and at the same time, this family does not cover the whole lattice

$$M^t = \{(a_1, a_2, \ldots, a_t) \mid a_i \in M \text{ for all } i \in \{1, \ldots, t\}\}.$$

It is known that all linear congruential generators have the lattice structure. Recall that a *linear congruential generator* $(x_n)_{n\in\mathbb{N}}$ is given by $a, m, c \in \mathbb{N}$ and defined by the recurrence relation $x_{n+1} = ax_n + c \mod m$. Let us mention a famous example of a PRNG with a striking lattice structure. For $t = 3$, the set of triples of RANDU, i.e., $\{(x_i, x_{i+1}, x_{i+2}) \mid i \in \mathbb{N}\}$ is covered by only 15 equidistant hyperplanes, see Figure 1.

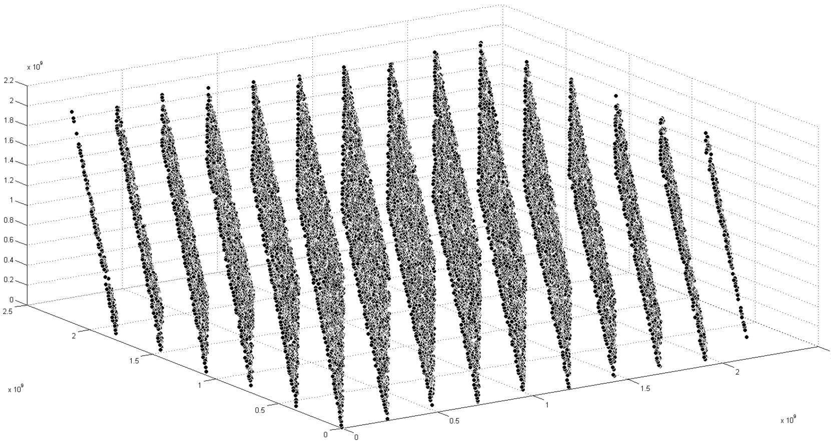


Fig. 1. The triples of RANDU – the linear congruential generator with $a = (2^{16} + 3), m = 2^{31}, c = 0$ – are covered by as few as 15 parallel equidistant planes.

1.2 Combinatorial Condition on Absence of the Lattice Structure

Guimond et al. in [2] have shown that PRNGs based on infinite words coding a certain class of cut-and-project sets have no lattice structure. A crucial part of their proof is the following lemma.

Lemma 1.1. *Let Z be the PRNG from* (1) *based on an aperiodic infinite word. If there exist for any $a, b \in M$ and for any $\ell \in \mathbb{N}$ an ℓ-tuple z such that both za and zb are $(\ell+1)$-tuples of the sequence Z, then Z does not have the lattice structure.*

We have found the following combinatorial condition on binary infinite words guaranteeing that the assumptions of the previous lemma are met: we say that a binary aperiodic infinite word u over the alphabet $\{0,1\}$ has *well distributed occurrences* (or has *the WDO property*) if u satisfies for any $m \in \mathbb{N}$ and any factor w of u the following condition. If we denote $i_0, i_1, \ldots$ the occurrences of w in u, then

$$\left\{ \left(|u_0 u_1 \cdots u_{i_j - 1}|_0, |u_0 u_1 \cdots u_{i_j - 1}|_1 \right) \bmod m \mid j \in \mathbb{N} \right\} = \mathbb{Z}_m^2 \,,$$

where mod$m$ is applied elementwise.

See the next section for the definition of aperiodicity, factor occurrences, and the WDO property for general alphabets.

The WDO property for binary words thus ensures no lattice structure for PRNGs defined in (1).

Theorem 1.2. *Let Z be the PRNG from* (1) *based on a binary aperiodic infinite word having the WDO property. Then Z has no lattice structure.*

We omit the proof of this theorem for the sake of brevity.

Moreover, we have shown that the class of infinite words satisfying the WDO property for binary words is larger than the class described in [2] (see Section 3).

2 Combinatorics on Words and the WDO Property

By A we denote a finite set of symbols called *letters*; the set A is therefore called an *alphabet.* A finite string $w = w_1 w_2 \dots w_n$ of letters from A is said to be a *finite word*, its length is denoted by $|w| = n$ and $|w|_a$ denotes the number of occurrences of $a \in A$ contained in w. The empty word, a neutral element for concatenation of finite words, is denoted ε and it is of zero length.

Under an *infinite word* we understand an infinite sequence $u = u_0 u_1 u_2 \dots$ of letters from A. A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \varepsilon$, then w is said to be a *prefix* of v; if $s = \varepsilon$, then w is a *suffix* of v. The set of factors and prefixes of v are denoted by $\mathrm{Fact}(v)$ and $\mathrm{Pref}(v)$, respectively. If $v = ps$ for finite words v, p, s, then we write $p = vs^{-1}$ and $s = p^{-1}v$.

An infinite word u over the alphabet A is called *eventually periodic* if it is of the form $u = vw^\omega$, where v, w are finite words over A and ω denotes an infinite repetition. An infinite word is called *aperiodic* if it is not eventually periodic.

For any factor w of an infinite word u, every index i such that w is a prefix of the infinite word $u_i u_{i+1} u_{i+2} \dots$ is called an *occurrence* of w in u.

The *factor complexity* of an infinite word u is a map $\mathcal{C}_u : \mathbb{N} \mapsto \mathbb{N}$ defined by $\mathcal{C}_u(n) :=$ the number of factors of length n contained in u. The factor complexity of eventually periodic words is bounded, while the factor complexity of an aperiodic word u satisfies $\mathcal{C}_u(n) \geq n + 1$ for all $n \in \mathbb{N}$. A *right extension* of a factor w of u over the alphabet A is any letter $a \in A$ such that wa is a factor u. Of course, any factor of u has at least one right extension. A factor w is called *right special* if w has at least two right extensions. Similarly, one can define a *left extension* and a *left special* factor. A factor is *bispecial* if it is both right and left special. An aperiodic word contains right special factors of any length.

The *Parikh vector* of a finite word w over an alphabet $\{0, 1, \dots, d-1\}$ is defined as $(|w|_0, |w|_1, \dots, |w|_{d-1})$. For a finite or infinite word $u = u_0 u_1 u_2 \dots$, we denote by $\mathrm{Pref}_n u$ the prefix of length n of u, i.e., $\mathrm{Pref}_n u = u_0 u_1 \dots u_{n-1}$.

Let us generalize the combinatorial condition on infinite words that guarantees no lattice structure for pseudorandom number generators from binary to multiliteral alphabets.

Definition 2.1 (The WDO property). *We say that an aperiodic infinite word u over the alphabet $\{0, 1, \dots, d-1\}$ has* well distributed occurrences *(or has* the WDO property*) if u satisfies for any $m \in \mathbb{N}$ and any factor w of u the following condition. If we denote $i_0, i_1, \dots$ the occurrences of w in u, then*

$$\left\{\left(|u_0 u_1 \cdots u_{i_j - 1}|_0, \dots, |u_0 u_1 \cdots u_{i_j - 1}|_{d-1}\right) \bmod m \mid j \in \mathbb{N}\right\} = \mathbb{Z}_m^d \, ;$$

that is, the Parikh vectors of $\mathrm{Pref}_{i_j}(u)$ for $j \in \mathbb{N}$, when reduced modulo m, give the whole $\mathbb{Z}_m^d$.

We define the WDO property for aperiodic words since it clearly never holds for periodic ones.

With the above notation, it is easy to see that if a recurrent infinite word u has the WDO property, then for every vector $\mathbf{v} \in \mathbb{Z}_m^d$ there are infinitely many values of j such that the Parikh vector of $\mathrm{Pref}_{i_j}(u)$ is congruent to $\mathbf{v}$ modulo m.

Example 2.2. The Thue-Morse word $t = 0110100110010110\cdots$, which is a fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 10$, does not satisfy the WDO property. Indeed, take $m = 2$ and $w = 00$, then w occurs only in odd positions i_j so that $(|t_0 \cdots t_{i_j-1}|_0 + |t_0 \cdots t_{i_j-1}|_1) = i_j$ is odd. Thus, e.g., $(|t_0 \cdots t_{i_j-1}|_0, |t_0 \cdots t_{i_j-1}|_1) \bmod 2 \neq (0,0)$, and hence $\{(|t_0 \cdots t_{i_j-1}|_0, |t_0 \cdots t_{i_j-1}|_1) \bmod 2 \mid j \in \mathbb{N}\} \neq \mathbb{Z}_2^2$.

Example 2.3. We say that an infinite word u over an alphabet A, $|A| = d$, is *universal* if it contains all finite words over A as its factors. It is easy to see that any universal word satisfies the WDO property. Indeed, for any word $w \in A^*$ and any m there exists a finite word v such that if we denote $i_0, i_1, \ldots, i_k$ the occurrences of w in v, then

$$\left\{\left(|\mathrm{Pref}_{i_j} v|_0, \ldots, |\mathrm{Pref}_{i_j} v|_{d-1}\right) \bmod m \mid j \in \{0, 1, \ldots, k\}\right\} = \mathbb{Z}_m^d\,.$$

Since u is universal, v is a factor of u. Denoting by i an occurrence of v in u, one gets that the positions $i + i_j$ are occurrences of w in u. Hence

$$\begin{aligned}&\left\{\left(|\mathrm{Pref}_{i+i_j} u|_0, \ldots, |\mathrm{Pref}_{i+i_j} u|_{d-1}\right) \bmod m \mid j \in \{0, 1, \ldots, k\}\right\} = \\ &= (|\mathrm{Pref}_i u|_0, \ldots, |\mathrm{Pref}_i u|_{d-1}) + \\ &+ \left\{\left(|\mathrm{Pref}_{i_j} v|_0, \ldots, |\mathrm{Pref}_{i_j} v|_{d-1}\right) \bmod m \mid j \in \{0, 1, \ldots, k\}\right\} = \mathbb{Z}_m^d\,.\end{aligned}$$

Therefore, u satisfies the WDO property.

3 Sturmian Words

In this section, we show that Sturmian words have well distributed occurrences.

Definition 3.1. *An aperiodic infinite word u is called* Sturmian *if its factor complexity satisfies $\mathcal{C}_u(n) = n + 1$ for all $n \in \mathbb{N}$.*

So, Sturmian words are by definition binary and they have the lowest possible factor complexity among aperiodic infinite words. Sturmian words admit various types of characterizations of geometric and combinatorial nature. One of such characterizations is via irrational rotations on the unit circle. In [4] Hedlund and Morse showed that each Sturmian word may be realized measure-theoretically by an irrational rotation on the circle. That is, every Sturmian word is obtained by coding the symbolic orbit of a point on the circle of circumference one under a rotation R_α by an irrational angle[5] α, $0 < \alpha < 1$, where the circle is partitioned into two complementary intervals, one of length α and the other of length $1 - \alpha$. And conversely each such coding gives rise to a Sturmian word.

[5] Measured by arc length (thus equivalent to $2\pi\alpha$ radians).

Definition 3.2. *The* rotation *by angle α is the mapping R_α from $[0,1)$ (identified with the unit circle) to itself defined by $R_\alpha(x) = \{x+\alpha\}$, where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of x. Considering a partition of $[0,1)$ into $I_0 = [0, 1-\alpha)$, $I_1 = [1-\alpha, 1)$, define a word*

$$s_{\alpha,\rho}(n) = \begin{cases} 0, & \textit{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_0, \\ 1, & \textit{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_1. \end{cases}$$

One can also define $I'_0 = (0, 1-\alpha]$, $I'_1 = (1-\alpha, 1]$, the corresponding word is denoted by $s'_{\alpha,\rho}$.

For more information on Sturmian words we refer to [3, Chapter 2].

Theorem 3.3. *Let u be a Sturmian word on $\{0,1\}$. Then u has Property WDO.*

Proof. In the proof we use the definition of Sturmian word via rotation. The main idea is controlling the number of 1's modulo m by taking circle of length m, and controlling the length taking the rotation by $m\alpha$.

For the proof we will use an equivalent reformulation of the theorem:

Let u be a Sturmian word on $\{0,1\}$, for any natural number m and any factor w of u let us denote $i_0, i_1, \ldots$ the occurrences of w in u. Then

$$\big\{\big(i_j, |u_0 u_1 \cdots u_{i_j - 1}|_1\big) \bmod m \mid j \in \mathbb{N}\big\} = \{0, 1, ..., m-1\}^2.$$

That is, we will control the number of 1's and the length instead the number of 0's.

Since a Sturmian word can be defined via rotations by an irrational angle on a unit circle, without loss of generality we may assume that $u = s_{\alpha,\rho}$ for some $0 < \alpha < 1$, $0 \leq \rho < 1$, α irrational (see Definition 3.2). Equivalently, we can consider m copies of the circle connected into one circle of length m with m intervals $I_1^i = [i - \alpha, i)$ of length α corresponding to 1. The Sturmian word is obtained by rotation by α on this circle of length m (see Fig. 2).

Namely, we define the rotation $R_{\alpha,m}$ as the mapping from $[0,m)$ (identified with the circle of length m) to itself defined by $R_{\alpha,m}(x) = \{x + \alpha\}_m$, where $\{x\}_m = x - \lfloor x/m \rfloor m$ and for $m = 1$ coincides with the fractional part of x. A partition of $[0,m)$ into $2m$ intervals $I_0^i = [i, i+1-\alpha)$, $I_1^i = [i+1-\alpha, i+1)$, $i = 0, \ldots, m-1$ defines the Sturmian word $u = s_{\alpha,\rho}$:

$$s_{\alpha,\rho}(n) = \begin{cases} 0, & \text{if } R_{\alpha,m}^n(\rho) = \{\rho + n\alpha\} \in I_0^i \text{ for some } i = 0, \ldots, m-1, \\ 1, & \text{if } R_{\alpha,m}^n(\rho) = \{\rho + n\alpha\} \in I_1^i \text{ for some } i = 0, \ldots, m-1. \end{cases}$$

It is well known that any factor $w = w_0 \cdots w_{k-1}$ of u corresponds to an interval I_w in $[0,1)$, so that whenever you start rotating from the interval I_w, you obtain w. Namely, $x \in I_w$ if and only if $x \in I_{w_0}, R_\alpha(x) \in I_{w_1}, \ldots, R_\alpha^{|w|-1}(x) \in I_{w_{|w|-1}}$.

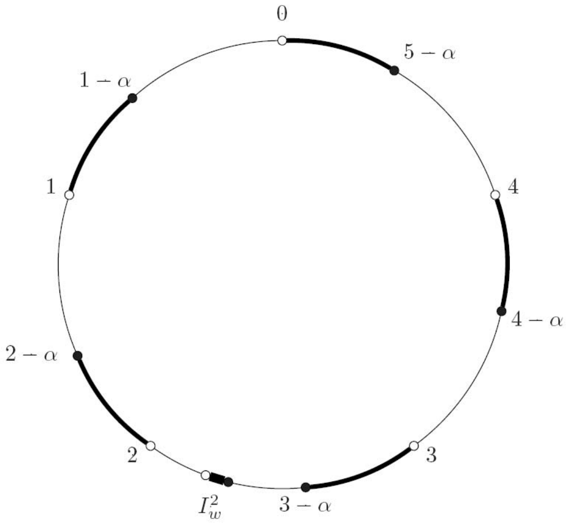


Fig. 2. Illustration to the proof of Theorem 3.3: the example for $m = 5$.

Similarly, we can define m intervals corresponding to w in $[0, m)$ (circle of length m), so that if $I_w = [x_1, x_2)$, then $I_w^i = [x_1 + i, x_2 + i)$, $i = 0, \dots, m - 1$.

Fix a factor w of u, take arbitrary $(j, i) \in \{0, 1, \dots, m - 1\}^2$. Now we will organize (j, i) among the occurrences of w, i.e., find l such that $u_l \dots u_{l+|w|-1} = w$, $l \mod m = j$ and $|\mathrm{Pref}_l u|_1 \mod m = i$.

Consider rotation $R_{m\alpha,m}(x)$ by $m\alpha$ instead of rotation by α, and start m-rotating from $j\alpha + \rho$. Formally, $R_{m\alpha,m}(x) = \{x + m\alpha\}_m$, where, as above, $\{x\}_m = x - [x/m]m$. This rotation will put us to positions $mk + j$, $k \in \mathbb{N}$ in the Sturmian word: for $a \in \{0, 1\}$ one has $s_{\alpha,\rho}(mk + j) = a$ if $R^k_{m\alpha,m}(j\alpha + \rho) = \{j\alpha + \rho + km\alpha\}_m \in I_a^i$ for some $i = 0, \dots, m - 1$.

Remark that the points in the orbit of an m-rotation of a point on the m-circle are dense, and hence the rotation comes infinitely often to each interval. So pick k when $j\alpha + mk\alpha + \rho \in I_w^i \subset [i, i + 1)$ (and actually there exist infinitely many such k). Then the length l of the corresponding prefix is equal to $km + j$, and the number of 1's in it is $i + mp$, where p is the number of complete circles you made, i.e., $p = [(j\alpha + mk\alpha + \rho)/m]$. □

Remark 3.4. In the next section we will show that Arnoux-Rauzy words [1], which are natural extensions of Sturmian words to larger alphabets, also satisfy the WDO property. Note that the proof above cannot be generalized to Arnoux-Rauzy words, because it is based on the geometric interpretation of Sturmian words via rotations, while this interpretation does not extend to Arnoux-Rauzy words.

4 Arnoux-Rauzy Words

4.1 Basic Definitions

Definition 4.1. *Let A be a finite alphabet. The* reversal operator *is the operator* $\sim: A^* \mapsto A^*$ *defined by recurrence in the following way:*

$$\tilde{\varepsilon} = \varepsilon, \quad \widetilde{va} = a\tilde{v}$$

for all $v \in A^*$ and $a \in A$. The fixed points of the reversal operator are called palindromes*.*

Definition 4.2. *Let $u \in A^*$ be a finite word over the alphabet A. We define the* right palindromic closure *of u, and we denote it by $u^{(+)}$ as the shortest palindrome that has u as a prefix. It is readily verified that if p is the longest palindromic suffix of $u = vp$, then $u^{(+)} = vp\tilde{v}$.*

Definition 4.3. *We call the* iterated (right) palindromic closure operator *the operator ψ recurrently defined by the following rules:*

$$\psi(\varepsilon) = \varepsilon, \quad \psi(va) = (\psi(v)a)^{(+)}$$

for all $v \in A^*$ and $a \in A$. The definition of ψ may be extended to infinite words u over A as $\psi(u) = \lim_n \psi(\mathrm{Pref}_n\, u)$, i.e., $\psi(u)$ is the infinite word having $\psi(\mathrm{Pref}_n\, u)$ as its prefix for every $n \in \mathbb{N}$.

Definition 4.4. *Let Δ be an infinite word on the alphabet A such that every letter occurs infinitely often in Δ. The word $c = \psi(\Delta)$ is then called a* characteristic (or standard) Arnoux-Rauzy word *and Δ is called the* directive sequence *of c. An infinite word u is called an Arnoux-Rauzy word if it has the same set of factors as a (unique) characteristic Arnoux-Rauzy word, which is called the characteristic word of u. The directive sequence of an Arnoux-Rauzy word is the directive sequence of its characteristic word.*

Let us also recall the following well-known characterization:

Theorem 4.5. *Let u be an aperiodic infinite word over the alphabet A. Then u is a standard Arnoux-Rauzy word if and only if the following hold:*

1. $\mathrm{Fact}(u)$ *is closed under reversal (that is, if v is a factor of u so is $\tilde{v}$).*
2. *Every left special factor of u is also a prefix.*
3. *If v is a right special factor of u then va is a factor of u for every $a \in A$.*

From the preceding theorem, it can be easily verified that the bispecial factors of a standard Arnoux-Rauzy correspond to its palindromic prefixes (including the empty word), and hence to the iterated palindromic closure of the prefixes of its directive sequence. That is, if

$$\varepsilon = b_0, b_1, b_2, \dots$$

is the sequence, ordered by length, of bispecial factors of the standard Arnoux-Rauzy word u, $\Delta = \Delta_0\Delta_1 \cdots$ its directive sequence (with $\Delta_i \in A$ for every i), we have $b_{i+1} = (b_i\Delta_i)^{(+)}$.

A direct consequence of this, together with the preceding definitions, is the following statement, which will be used in the sequel.

Lemma 4.6. *Let u be a characteristic Arnoux-Rauzy word and let Δ and $(b_i)_{i\geq 0}$ be defined as above. If Δ_i does not occur in b_i, then $b_{i+1} = b_i\Delta_i b_i$. Otherwise let $j < i$ be the largest integer such that $\Delta_j = \Delta_i$. Then $b_{i+1} = b_i b_j^{-1} b_i$.*

4.2 Parikh Vectors and Arnoux-Rauzy Factors

Where no confusion arises, given an Arnoux-Rauzy word u, we will denote by

$$\varepsilon = b_0, b_1, \dots, b_n, \dots$$

the sequence of bispecial factors of u ordered by length and we will set for any $i \in \mathbb{N}$, B_i as the Parikh vector of b_i.

Remark 4.7. By the pigeonhole principle, it is clear that for every $m \in \mathbb{N}$ there exists an integer $N \in \mathbb{N}$ such that, for every $i \geq N$, the set $\{j > i \mid B_j \equiv_m B_i\}$ is infinite. Where no confusion arises and with a slight abuse of notation, fixed m, we will always denote by N the smallest of such integers.

Lemma 4.8. *Let u be a characteristic Arnoux-Rauzy word and let $m \in \mathbb{N}$. Let*

$$\alpha_1 B_{j_1} + \cdots + \alpha_k B_{j_k} \equiv_m \bar{\mathbf{v}} \in \mathbb{Z}_m^d$$

be a linear combination of Parikh vectors such that $\sum_{i=1}^{k} \alpha_i = 0$, with $j_i \geq N$ and $\alpha_i \in \mathbb{Z}$ for all $i \in \{1, \dots k\}$. Then, for any $\ell \in \mathbb{N}$, there exists a prefix v of u such that the Parikh vector of v is congruent to $\bar{\mathbf{v}}$ modulo m and vb_ℓ is also a prefix of u.

Proof. Without loss of generality, we can assume $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$, hence there exists k' such that

$$\alpha_1 \geq \alpha_{k'} \geq 0 \geq \alpha_{k'+1} \geq \alpha_k.$$

We will prove the result by induction on $\beta = \sum_{j=1}^{k'} \alpha_j$. If $\beta = 0$, trivially, we can take $v = \varepsilon$ and the statement is clearly verified. Let us assume the statement true for all $0 \leq \beta < M$ and let us prove it for $\beta = M$. By the remark preceding this lemma, for every ℓ we can choose $i' > j' > \ell$ such that $B_{j_1} \equiv_m B_{i'}$ and $B_{j_k} \equiv_m B_{j'}$. Since every bispecial factor is a prefix and suffix of all the bigger ones, in particular we have that $b_{j'}$ is a suffix of $b_{i'}$, and b_ℓ is a prefix of $b_{j'}$; this implies that $b_{i'}b_{j'}^{-1}b_\ell$ is actually a prefix of $b_{i'}$. By assumption, the Parikh vector of $b_{i'}b_{j'}^{-1}$ is clearly $B_{i'} - B_{j'} \equiv_m B_{j_1} - B_{j_k}$. Since $\alpha_1 \geq 1$ implies $\alpha_k \leq -1$, we have, by induction hypothesis, that there exists a prefix v of u such that the Parikh vector of v is congruent modulo m to

$$(\alpha_1 - 1)B_{j_1} + \cdots + (\alpha_k + 1)B_{j_k}$$

and $vb_{i'}$ is a prefix of u. Hence $vb_{i'}b_{j'}^{-1}b_\ell$ is also a prefix of u and, by simple computation, the Parikh vector of $vb_{i'}b_{j'}^{-1}$ is congruent modulo m to $\bar{\mathbf{v}}$. □

Definition 4.9. *Let $n \in \mathbb{Z}$. We will say that an integer linear combination of integer vectors is a n-*combination *if the sum of all the coefficients equals n.*

Lemma 4.10. *Let u be a characteristic Arnoux-Rauzy word and let $n \in \mathbb{N}$. Every n-combination of Parikh vectors of bispecial factors can be expressed as a n-combination of Parikh vectors of arbitrarily large bispecials. In particular, for every $K, M \in \mathbb{N}$, it is possible to find a finite number of integers $\alpha_1, \ldots, \alpha_k$ such that $B_K = \alpha_1 B_{j_1} + \cdots + \alpha_k B_{j_k}$ with $j_i > M$ for every i and $\alpha_1 + \cdots + \alpha_k = 1$.*

Proof. A direct consequence of Lemma 4.6 is that for every i such that Δ_i appears in b_i, we have $B_{i+1} = 2B_i - B_j$, where $j < i$ is the largest such that $\Delta_j = \Delta_i$. This in turn (since every letter in Δ appears infinitely many times from the definition of Arnoux-Rauzy word) implies that *for every* non-negative integer j, there exists a positive k such that $B_j = 2B_{j+k} - B_{j+k+1}$, that is, we can substitute each Parikh vector of a bispecial with a 1-combination of Parikh vectors of strictly larger bispecials. Simply iterating the process, we obtain the statement. □

In the following we will assume the set A to be a finite alphabet of cardinality d. For every set $X \subseteq A^*$ of finite words, we will denote by $\mathrm{PV}(X) \subseteq \mathbb{Z}^d$ the set of Parikh vectors of elements of X and for every $m \in \mathbb{N}$ we will denote by $\mathrm{PV}_m(X) \subseteq \mathbb{Z}_m^d$ the set of elements of $\mathrm{PV}(X)$ reduced modulo m.

Let u be an infinite word over A and let v be a factor of u. We denote by $S_v(u)$ the set of all prefixes of u followed by an occurrence of v. In other words,

$$S_v(u) = \{p \in \mathrm{Pref}(u) \mid pv \in \mathrm{Pref}(u)\}.$$

Definition 4.11. *For any set of finite words $X \subseteq A^*$, we will say that u* has the property $\mathcal{P}_X$ *(or, for short, that u has $\mathcal{P}_X$) if, for every $m \in \mathbb{N}$ and for every $v \in X$ we have that*

$$\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d.$$

That is to say, for every vector $\mathbf{v} \in \mathbb{Z}_m^d$ there exists a word $w \in S_v(u)$ such that the Parikh vector of w is congruent to $\mathbf{v}$ modulo m.

With this notation, an infinite word u has the WDO property if and only if it has property $\mathcal{P}_{\mathrm{Fact}(u)}$.

Proposition 4.12. *Let u be a characteristic Arnoux-Rauzy word over the d-letter alphabet A. Then u has the property $\mathcal{P}_{\mathrm{Pref}(u)}$.*

Proof. Let us fix an arbitrary $m \in \mathbb{N}$. We want to show that, for every $v \in \mathrm{Pref}(u)$, $\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d$. Let then $\bar{\mathbf{v}} \in \mathbb{Z}^d$ and ℓ be the smallest number such that v is a prefix of b_ℓ. Let $i_1 < i_2 < \cdots < i_d$ be such that Δ_{i_j} does not appear in b_{i_j}, where Δ is the directive word of u. Without loss of generality, we can rearrange the letters so that each Δ_{i_j} is lexicographically smaller than $\Delta_{i_{j+1}}$. With this assumption if, for every j, we set $\bar{\mathbf{v}}_j$ as the Parikh vector of b_{i_j+1}, which, by the first part of Lemma 4.6, equals $b_{i_j}\Delta_{i_j}b_{i_j}$, we can find $j-1$ positive integers $\mu_1, \dots, \mu_{j-1}$ such that $\bar{\mathbf{v}}_j = (\mu_1, \mu_2, \dots, \mu_{j-1}, 1, 0, \dots, 0)$. It is easy to show, then, that the set $V = \{\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_d\}$ generates $\mathbb{Z}^d$, hence there exists an integer n such that $\bar{\mathbf{v}}$ can be expressed as an n-combination of elements of V (which are Parikh vectors of bispecial factors of u). Trivially, then, $\bar{\mathbf{v}} = \bar{\mathbf{v}} - n\bar{\mathbf{0}} = \bar{\mathbf{v}} - nB_0$; thus, it is possible to express $\bar{\mathbf{v}}$ as a 0-combination of Parikh vectors of (by the previous Lemma 4.10) arbitrarily large bispecial factors of u. By Lemma 4.8, then there exists a prefix p of u with Parikh vector $\bar{\mathbf{p}}$ such that $\bar{\mathbf{p}} \equiv_m \bar{\mathbf{v}}$ and pb_ℓ is a prefix of u. Since we picked ℓ such that v is a prefix of b_ℓ, we have that $p \in S_v(u)$. From the arbitrariness of v, $\bar{\mathbf{v}}$ and m, we obtain the statement. □

As a corollary of Proposition 4.12, we obtain the main result of this section.

Theorem 4.13. *Let u be an Arnoux-Rauzy word over the d-letter alphabet A. Then u has the property $\mathcal{P}_{\mathrm{Fact}(u)}$.*

Proof. Let m be a positive integer and let c be the characteristic word of u. Let v be a factor of u and xvy be the smallest bispecial containing v. By Proposition 4.12, we have that $\mathrm{PV}_m(S_{xv}(c)) = \mathbb{Z}_m^d$ and, since the set is finite, we can find a prefix p of c such that $\mathrm{PV}_m(S_{xv}(p)) = \mathbb{Z}_m^d$. Let w be a prefix of u such that wp is a prefix of u. If $\bar{\mathbf{x}}$ and $\bar{\mathbf{w}}$ are the Parikh vectors of, respectively, x and w, it is easy to see that

$$\bar{\mathbf{w}} + \bar{\mathbf{x}} + \mathrm{PV}(S_{xv}(p)) \subseteq \bar{\mathbf{w}} + \mathrm{PV}(S_v(p)) \subseteq \mathrm{PV}(S_v(u))$$

Since we have chosen p such that $\mathrm{PV}_m(S_{xv}(p)) = \mathbb{Z}_m^d$, we clearly obtain that $\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d$ and hence, by the arbitrariness of v and m, the statement. □

Remark 4.14. Actually, Theorem 4.13 implies Theorem 3.3.

Remark 4.15. Note the following simple method of obtaining words satisfying the WDO property. Take a word u with the WDO property over an alphabet $\{0, 1, \dots, d-1\}$, $d > 2$, apply a morphism $\varphi : d-1 \mapsto 0, i \mapsto i$ for $i = 0, \dots, d-2$, i. e., φ joins two letters into one. It is straightforward that $\varphi(u)$ has WDO property. So, taking Arnoux-Rauzy words and joining some letters, we obtain other words than Sturmian and Arnoux-Rauzy satisfying the WDO property.

Acknowledgements

The first author was supported by the Czech Science Foundation grant GAČR 13-03538S, and thanks L'Oréal Czech Republic for the Fellowship Women in Science. The third author was partially supported by the Italian Ministry of Education (MIUR), under the PRIN 2010–11 project "Automi e Linguaggi Formali: Aspetti Matematici e Applicativi". The fourth author is supported in part by the Academy of Finland under grant 251371 and by Russian Foundation of Basic Research (grants 12-01-00089 and 12-01-00448).

We would like to acknowledge statistical testing of the pseudorandom number generators based on Sturmian and Arnoux-Rauzy words made by Jiří Hladký. He has shown using the Diehard and U01 tests that not only the lattice structure is absent, but also other important properties of PRNGs are improved when LCGs are combined using infinite words having the WDO property.

References

1. P. Arnoux, G. Rauzy, *Représentation géométrique de suites de complexité* $2n+1$, Bull. Soc. Math. France **119** (1991), 199–215.
2. L.-S. Guimond, Jan Patera, Jiří Patera, *Statistical properties and implementation of aperiodic pseudorandom number generators*, Applied Numerical Mathematics **46(3-4)** (2003), 295–318.
3. M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications 90, Cambridge University Press, 2002.
4. M. Morse and G.A. Hedlund, *Symbolic Dynamics II: Sturmian trajectories*, Amer. J. Math. **62 (1)** (1940), 1–42.

Pseudorandom Number Generators Based on Infinite Words

Authors: **Ľubomíra Balková, Michelangelo Bucci, Alessandro De Luca, Jiří Hladký, Svetlana Puzynina**
Journal: submitted to Mathematics of Computation
Year: 2014

MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

APERIODIC PSEUDORANDOM NUMBER GENERATORS BASED ON INFINITE WORDS

ĽUBOMÍRA BALKOVÁ, MICHELANGELO BUCCI, ALESSANDRO DE LUCA,
JIŘÍ HLADKÝ, AND SVETLANA PUZYNINA

Abstract. In this paper we study how certain families of aperiodic infinite words can be used to produce aperiodic pseudorandom number generators (PRNGs) with good statistical behavior. We introduce the *well distributed occurrences* (WELLDOC) combinatorial property for infinite words, which guarantees absence of the lattice structure defect in related pseudorandom number generators. An infinite word u on a d-ary alphabet has the WELLDOC property if, for each factor w of u, positive integer m, and vector $\mathbf{v} \in \mathbb{Z}_m^d$, there is an occurrence of w such that the Parikh vector of the prefix of u preceding such occurrence is congruent to $\mathbf{v}$ modulo m. (The Parikh vector of a finite word v over an alphabet $\mathcal{A}$ has its i-th component equal to the number of occurrences of the i-th letter of $\mathcal{A}$ in v.) We prove that Sturmian words, and more generally Arnoux-Rauzy words and some morphic images of them, have the WELLDOC property. Using the TestU01 [12] and PractRand [6] statistical tests, we moreover show that not only the lattice structure is absent, but also other important properties of PRNGs are improved when linear congruential generators are combined using infinite words having the WELLDOC property.

Introduction

Pseudorandom number generators aim to produce random numbers using a deterministic process. No wonder they suffer from many defects. The most usual ones – linear congruential generators – are known to produce periodic sequences with a defect called the lattice structure. Guimond et al. [15] proved that when two linear congruential generators are combined using infinite words coding certain classes of quasicrystals or, equivalently, of cut-and-project sets, the resulting sequence is aperiodic and has no lattice structure. For some other related results concerning aperiodic pseudorandom generators we refer to [13, 14]. We mention that although the lattice structure is considered as a defect of a random number generator, it can be useful in some applications for approximation of the uniform distribution [10].

We have found a combinatorial condition – *well distributed occurrences*, or WELLDOC for short – that also guarantees absence of the lattice structure in related pseudorandom generators. The WELLDOC property for an infinite word u over an alphabet $\mathcal{A}$ means that for any integer m and any factor w of u, the set of Parikh vectors modulo m of prefixes of u preceding the occurrences of w coincides with $\mathbb{Z}_m^{|\mathcal{A}|}$ (see Definition 2.1). In other words, among Parikh vectors modulo m of such prefixes one has all possible vectors. Besides giving generators without lattice

2010 *Mathematics Subject Classification.* 68R15.

structure, the WELLDOC property is an interesting combinatorial property of infinite words itself. We prove that the WELLDOC property holds for the family of Sturmian words, and more generally for Arnoux-Rauzy words.

Sturmian words constitute a well studied family of infinite aperiodic words. Let u be an infinite word, i. e., an infinite sequence of elements from a finite set called an alphabet. The *(factor) complexity* function counts the number of distinct factors of u of length n. A fundamental result of Morse and Hedlund [18] states that a word u is eventually periodic if and only if for some n its complexity is less than or equal to n. Infinite words of complexity $n + 1$ for all n are called *Sturmian words,* and hence they are aperiodic words of the smallest complexity. The most studied Sturmian word is the so-called Fibonacci word

$$0100101001001010010100100101001 0 \dots$$

fixed by the morphism $0 \mapsto 01$ and $1 \mapsto 0$. (See Section 2 for formal definitions.) The first systematic study of Sturmian words was given by Morse and Hedlund in [19]. Such sequences arise naturally in many contexts, and admit various types of characterizations of geometric and combinatorial nature (see, e.g., [16]).

Arnoux-Rauzy words were introduced in [1] as natural extensions of Sturmian words to multiliteral alphabets (see Definition 4.4). Despite the fact that they were introduced as generalizations of Sturmian words, Arnoux-Rauzy words display a much more complex behavior. In particular, we have two different proofs of the WELLDOC property for Sturmian words, and only one of them can be generalized to Arnoux-Rauzy words. In the sequel we provide both of them.

An infinite word with the WELLDOC property is then used to combine two linear congruential generators and form an infinite aperiodic sequence with good statistical behavior. Using the TestU01 [12] and PractRand [6] statistical tests, we have moreover shown that not only the lattice structure is absent, but also other important properties of PRNGs are improved when linear congruential generators are combined using infinite words having the WELLDOC property.

The paper is organized as follows. In the next section, we give some background on pseudorandom number generation. Next, in Section 2, we give the basic combinatorial definitions needed for our main results, including the WELLDOC property, and we prove that the WELLDOC property of u guarantees absence of the lattice structure of the PRNG based on u. In Sections 3 and 4, we prove that the property holds for Sturmian and Arnoux-Rauzy words. Finally, in Section 5, we present results of empirical tests of PRNGs based on words having the WELLDOC property.

A preliminary version of this paper [2], using the acronym *WDO* instead of WELLDOC, was presented at the WORDS 2013 conference.

1. Pseudorandom Number Generators and Lattice Structure

For the sake of our discussion, any infinite sequence of integers can be understood as a *pseudorandom number generator (PRNG)*; see also [15]. The generators the most widely used in the past – linear congruential generators – are known to suffer from a defect called the lattice structure (they possess it already from dimension 2 as shown in [17]).

Let $Z = (Z_n)_{n\in\mathbb{N}}$ be a PRNG whose output is a finite set $M \subset \mathbb{N}$. We say that Z has the *lattice structure* if there exists $t \in \mathbb{N}$ such that the set

$$\{(Z_i, Z_{i+1}, \dots, Z_{i+t-1}) \bigm| i \in \mathbb{N}\}$$

is covered by a family of parallel equidistant hyperplanes and at the same time, this family does not cover the whole lattice

$$M^t = \{(A_1, A_2, \ldots, A_t) \mid A_i \in M \text{ for all } i \in \{1, \ldots, t\}\}.$$

Recall that a *linear congruential generator* (LCG) $(Z_n)_{n\in\mathbb{N}}$ is given by parameters $a, m, c \in \mathbb{N}$ and defined by the recurrence relation $Z_{n+1} = aZ_n + c \mod m$. Let us mention a famous example of a LCG whose lattice structure is striking. For $t = 3$, the set of triples of RANDU, i.e., $\{(Z_i, Z_{i+1}, Z_{i+2}) \mid i \in \mathbb{N}\}$ is covered by only 15 parallel equidistant hyperplanes, see Figure 1.

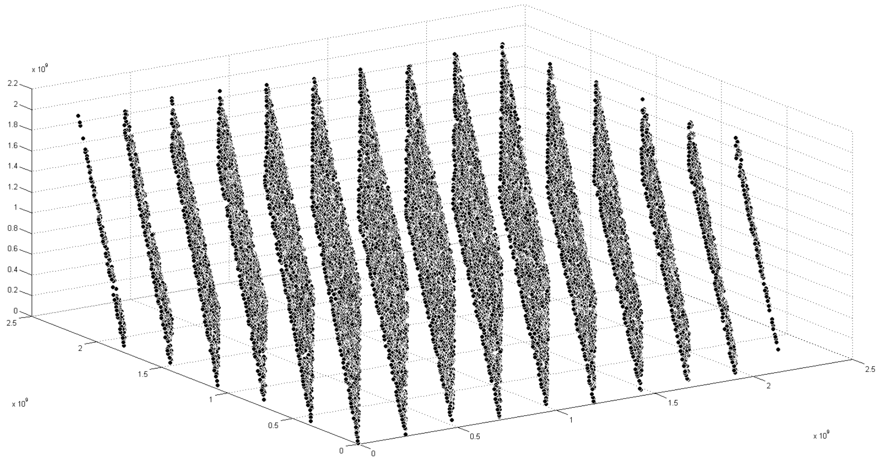


Figure 1. The triples of RANDU – the LCG with $a = (2^{16} + 3), m = 2^{31}, c = 0$ – are covered by as few as 15 parallel equidistant planes.

In the paper of Guimond et al. [15], a restricted version of the following sufficient condition for the absence of the lattice structure is formulated.

Proposition 1.1. *Let Z be a PRNG whose output is a finite set $M \subset \mathbb{N}$ containing at least two elements. Assume there exists for any $A, B \in M$ and for any $\ell \in \mathbb{N}$ an ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ such that both $(A_1, A_2, \ldots, A_\ell, A)$ and $(A_1, A_2, \ldots, A_\ell, B)$ are $(\ell + 1)$-tuples of the generator Z. Then Z does not have the lattice structure.*

Remark 1.2. Proposition 1.1 can be reformulated in terms of combinatorics on words (see Section 2) as follows: Let Z be a PRNG whose output is a finite set $M \subset \mathbb{N}$ containing at least two elements. If for any $A, B \in M$ and any length ℓ Z has a right special factor of length ℓ with right extensions A and B, then Z does not have the lattice structure.

Since Proposition 1.1 is formulated for a restricted class of generators in [15] (see Lemma 2.3 ibidem), we will provide its proof. However, we point out that all ideas of the proof are taken from [15]. We start with an auxiliary lemma.

Let us denote $\lambda = \gcd\{A - B \mid A, B \in M\}$.

Lemma 1.3. *Let Z be a PRNG satisfying all assumptions of Proposition 1.1. Let $\bar{\mathbf{n}}$ be the unit normal vector of a family of parallel equidistant hyperplanes covering all t-tuples of Z. Assume $\bar{\mathbf{e}}_i$ (the i-th vector of the canonical basis of the Euclidean space $\mathbb{R}^t$) is not orthogonal to $\bar{\mathbf{n}}$. Then the distance d_i of adjacent hyperplanes in the family along $\bar{\mathbf{e}}_i$ is of the form λ/k for some $k \in \mathbb{N}$.*

Remark 1.4. The distance d_i of adjacent hyperplanes W_0, W_1 along $\bar{\mathbf{e}}_i$ means $|x_i - y_i|$ for any $\bar{\mathbf{x}} \in W_0$ and $\bar{\mathbf{y}} \in W_1$, where the j-th components of $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ satisfy $x_j = y_j$ for all $j \in \{1, \dots, t\}, j \neq i$. This is a well defined term because the hyperplanes in the family are of the form $W_j \equiv \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} = \alpha + jd,\ j \in \mathbb{Z}$, where d is the distance of adjacent hyperplanes in the family and $\cdot$ denotes the standard scalar product. Thus, without loss of generality, consider the adjacent hyperplanes

$$W_0 \equiv \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} = \alpha \quad \text{and} \quad W_1 \equiv \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} = \alpha + d.$$

Then for any $\bar{\mathbf{x}} \in W_0$ and $\bar{\mathbf{y}} = \bar{\mathbf{x}} + s\bar{\mathbf{e}}_i$ from W_1, we have

$$\begin{aligned} \bar{\mathbf{y}} \cdot \bar{\mathbf{n}} &= \alpha + d = \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} + d, \\ \bar{\mathbf{y}} \cdot \bar{\mathbf{n}} &= \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} + s\bar{\mathbf{e}}_i \cdot \bar{\mathbf{n}} = \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} + sn_i, \end{aligned}$$

where n_i is the i-th component of $\bar{\mathbf{n}}$. Consequently, $d_i = |s| = \left|\frac{d}{n_i}\right|$ and is the same for any choice of $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ which differ only in their i-th component and belong to adjacent hyperplanes.

Proof of Lemma 1.3. Let us start with a useful observation. Let $\bar{\mathbf{z}}$ belong to a hyperplane W of the family in question.

(1) If $\bar{\mathbf{e}}_j$ is orthogonal to $\bar{\mathbf{n}}$, then we may change the j-th component of $\bar{\mathbf{z}}$ in an arbitrary way and the resulting vector will belong to the same hyperplane, i.e., if $W \equiv \bar{\mathbf{x}} \cdot \bar{\mathbf{n}} = \alpha$, then clearly $(\bar{\mathbf{z}} + \beta\bar{\mathbf{e}}_j) \cdot \bar{\mathbf{n}} = \bar{\mathbf{z}} \cdot \bar{\mathbf{n}} = \alpha$ for any $\beta \in \mathbb{R}$, thus $\bar{\mathbf{z}} + \beta\bar{\mathbf{e}}_j$ belongs to W.

(2) If $\bar{\mathbf{e}}_j$ is not orthogonal to $\bar{\mathbf{n}}$ and the distance d_j of adjacent hyperplanes along $\bar{\mathbf{e}}_i$ in the family is of the form λ/k for some $k \in \mathbb{N}$, then $\bar{\mathbf{z}} + r\lambda\bar{\mathbf{e}}_j$ belongs to the family for any $r \in \mathbb{Z}$. This follows from a repeated application of the fact that if $\bar{\mathbf{z}}$ belongs to a hyperplane W, then $\bar{\mathbf{z}} + \frac{\lambda}{k}\bar{\mathbf{e}}_j$ belongs to an adjacent hyperplane of W.

Let us proceed by contradiction, i.e., we assume that there exists $i \in \{1, \dots, t\}$ such that $\bar{\mathbf{e}}_i$ is not orthogonal to $\bar{\mathbf{n}}$ and the distance along $\bar{\mathbf{e}}_i$ of adjacent hyperplanes of the family in question is not of the form $\lambda/k,\ k \in \mathbb{N}$. Take the largest of such indices and denote it by ℓ. Choose $A, B \in M$ arbitrarily. According to assumptions, there exists an $(\ell-1)$-tuple $(A_1, A_2, \dots, A_{\ell-1})$ such that both $(A_1, A_2, \dots, A_{\ell-1}, A)$ and $(A_1, A_2, \dots, A_{\ell-1}, B)$ are ℓ-tuples of Z. It is therefore possible to find two t-tuples of Z such that the first one is of the form $(A_1, A_2, \dots, A_{\ell-1}, A, A_{\ell+1}, \dots, A_t)$ and the second one of the form $(A_1, A_2, \dots, A_{\ell-1}, B, \hat{A}_{\ell+1}, \dots, \hat{A}_t)$. These two t-tuples – considered as vectors in $\mathbb{R}^t$ – belong by the assumption of Lemma 1.3 to some hyperplanes in the family. Since all vectors $\bar{\mathbf{e}}_j,\ j \in \{\ell+1, \dots, t\}$ are either orthogonal to $\bar{\mathbf{n}}$ or the distance of adjacent hyperplanes along $\bar{\mathbf{e}}_j$ is of the form λ/k for some $k \in \mathbb{N}$, we can change the last $t-\ell$ coordinates $\hat{A}_{\ell+1}, \dots, \hat{A}_t$ of the second vector to arbitrary values from M (we transform them into $A_{\ell+1}, \dots, A_t$) and it will still belong to a hyperplane in the family. This is a consequence of the observation at the beginning of this proof. Hence, both vectors $(A_1, A_2, \dots, A_{\ell-1}, A, A_{\ell+1}, \dots, A_t)$ and $(A_1, A_2, \dots, A_{\ell-1}, B, A_{\ell+1}, \dots, A_t)$ belong to some hyperplanes of the family.

Their distance along $\bar{\mathbf{e}}_\ell$ equals $|A-B|$, i.e., d_ℓ divides $A-B$. Since A, B have been chosen arbitrarily, it follows that d_ℓ divides λ, i.e., $\lambda = kd_\ell$ for some $k \in \mathbb{N}$, which is a contradiction with the choice of $\bar{\mathbf{e}}_\ell$. □

Proof of Proposition 1.1. Let $\bar{\mathbf{n}}$ be the unit normal vector of a family of parallel equidistant hyperplanes covering all t-tuples of Z. Suppose without loss of generality that $\bar{\mathbf{e}}_1, \dots, \bar{\mathbf{e}}_\ell$ are not orthogonal to $\bar{\mathbf{n}}$ and $\bar{\mathbf{e}}_{\ell+1}, \dots, \bar{\mathbf{e}}_t$ are orthogonal to $\bar{\mathbf{n}}$. Let $\bar{\mathbf{z}} = (Z_n, Z_{n+1}, \dots, Z_{n+t-1})$ be a t-tuple of Z, thus $\bar{\mathbf{z}}$ belongs to one of the hyperplanes. Take any vector $\bar{\mathbf{y}} \in M^t$ and let us show that it belongs to a hyperplane in the family.

(1) Any vector from M^t which differs from $\bar{\mathbf{z}}$ only in the first ℓ components belongs to a hyperplane of the family. This comes from Lemma 1.3 because when we change for $i \in \{1, \dots, \ell\}$ the i-th component of $\bar{\mathbf{z}}$ by $d_i = \frac{\lambda}{k}$, then we jump on the adjacent parallel hyperplane. So, any transformation of the i-th component of $\bar{\mathbf{z}}$ into another value from M means a finite number of jumps from one hyperplane onto another. Hence, we may transform $\bar{\mathbf{z}}$ so that it has the first ℓ components equal to $\bar{\mathbf{y}}$ and the obtained vector $\bar{\mathbf{x}}$ belongs to a hyperplane in the family.

(2) Any vector from M^t which differs from $\bar{\mathbf{x}}$ only in the last $t-\ell$ components belongs to the same hyperplane as $\bar{\mathbf{x}}$. This comes from the orthogonality $\bar{\mathbf{e}}_i \perp \bar{\mathbf{n}}$ for $i > \ell$ (the argument is the same as in the proof of Lemma 1.3). Since $\bar{\mathbf{y}}$ differs from $\bar{\mathbf{x}}$ only in the last $t-\ell$ components, $\bar{\mathbf{y}}$ belongs to a hyperplane in the family.

□

2. Combinatorics on Words and the WELLDOC Property

2.1. **Backgrounds on Combinatorics on Words.** In the following, $\mathcal{A}$ denotes a finite set of symbols called *letters*; the set $\mathcal{A}$ is therefore called an *alphabet*. A *finite word* is a finite string $w = w_1w_2\dots w_n$ of letters from $\mathcal{A}$; its length is denoted by $|w| = n$ and $|w|_a$ denotes the number of occurrences of $a \in \mathcal{A}$ in w. The empty word, a neutral element for concatenation of finite words, is denoted ε and it is of zero length. The set of all finite words over the alphabet $\mathcal{A}$ is denoted by $\mathcal{A}^*$.

Under an *infinite word* we understand an infinite sequence $u = u_0u_1u_2\dots$ of letters from $\mathcal{A}$. A finite word w is a *factor* of a word v (finite or infinite) if there exist words p and s such that $v = pws$. If $p = \varepsilon$, then w is said to be a *prefix* of v; if $s = \varepsilon$, then w is a *suffix* of v. The set of factors and prefixes of v are denoted by $\mathrm{Fact}(v)$ and $\mathrm{Pref}(v)$, respectively. If $v = ps$ for finite words v, p, s, then we write $p = vs^{-1}$ and $s = p^{-1}v$.

An infinite word u over the alphabet $\mathcal{A}$ is called *eventually periodic* if it is of the form $u = vw^\omega$, where v, w are finite words over $\mathcal{A}$ and ω denotes an infinite repetition. An infinite word is called *aperiodic* if it is not eventually periodic.

For any factor w of an infinite word u, every index i such that w is a prefix of the infinite word $u_iu_{i+1}u_{i+2}\dots$ is called an *occurrence* of w in u. An infinite word u is *recurrent* if each of its factors has infinitely many occurrences in u.

The *factor complexity* of an infinite word u is a map $\mathcal{C}_u : \mathbb{N} \mapsto \mathbb{N}$ defined by $\mathcal{C}_u(n) :=$ the number of factors of length n contained in u. The factor complexity of eventually periodic words is bounded, while the factor complexity of an aperiodic word u satisfies $\mathcal{C}_u(n) \geq n+1$ for all $n \in \mathbb{N}$. A *right extension* of a factor w of u

over the alphabet $\mathcal{A}$ is any letter $a \in \mathcal{A}$ such that wa is a factor of u. Of course, any factor of u has at least one right extension. A factor w is called *right special* if w has at least two right extensions. Similarly, one can define a *left extension* and a *left special* factor. A factor is *bispecial* if it is both right and left special. An aperiodic word contains right special factors of any length.

The *Parikh vector* of a finite word w over an alphabet $\{0, 1, \ldots, d-1\}$ is defined as $(|w|_0, |w|_1, \ldots, |w|_{d-1})$. For a finite or infinite word $u = u_0u_1u_2\ldots$, $\operatorname{Pref}_n u$ will denote the prefix of length n of u, i.e., $\operatorname{Pref}_n u = u_0u_1\ldots u_{n-1}$.

In some of the examples we consider are morphic words. A *morphism* is a function $\varphi : \mathcal{A}^* \to \mathcal{B}^*$ such that $\varphi(\varepsilon) = \varepsilon$ and $\varphi(wv) = \varphi(w)\varphi(v)$, for all $w, v \in \mathcal{A}^*$. Clearly, a morphism is completely defined by the images of the letters in the domain. A morphism is *prolongable* on $a \in \mathcal{A}$, if $|\varphi(a)| \geq 2$ and a is a prefix of $\varphi(a)$. If φ is prolongable on a, then $\varphi^n(a)$ is a proper prefix of $\varphi^{n+1}(a)$, for all $n \in \mathbb{N}$. Therefore, the sequence $(\varphi^n(a))_{n\geq 0}$ of words defines an infinite word u that is a fixed point of φ. Such a word u is a (pure) *morphic* word.

Let us introduce a combinatorial condition on infinite words that – as we will see later – guarantees no lattice structure for the associated PRNGs.

Definition 2.1 (The WELLDOC property)**.** We say that an aperiodic infinite word u over the alphabet $\{0, 1, \ldots, d-1\}$ has *well distributed occurrences* (or has *the WELLDOC property*) if for any $m \in \mathbb{N}$ and any factor w of u the word u satisfies the following condition. If $i_0, i_1, \ldots$ denote the occurrences of w in u, then

$$\big\{\big(|\operatorname{Pref}_{i_j} u|_0, \ldots, |\operatorname{Pref}_{i_j} u|_{d-1}\big) \bmod m \mid j \in \mathbb{N}\big\} = \mathbb{Z}_m^d\,;$$

that is, the Parikh vectors of $\operatorname{Pref}_{i_j} u$ for $j \in \mathbb{N}$, when reduced modulo m, give the whole set $\mathbb{Z}_m^d$.

We define the WELLDOC property for aperiodic words since it clearly never holds for periodic ones. It is easy to see that if a recurrent infinite word u has the WELLDOC property, then for every vector $\mathbf{v} \in \mathbb{Z}_m^d$ there are infinitely many values of j such that the Parikh vector of $\operatorname{Pref}_{i_j} u$ is congruent to $\mathbf{v}$ modulo m.

Example 2.2. The Thue-Morse word

$$u = 0110100110010110100101100110100 1\cdots,$$

which is a fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 10$, does not satisfy the WELLDOC property. Indeed, take $m = 2$ and $w = 00$, then w occurs only in odd positions i_j so that $(|\operatorname{Pref}_{i_j} u|_0 + |\operatorname{Pref}_{i_j} u|_1) = i_j$ is odd. Thus, e.g.,

$$(|\operatorname{Pref}_{i_j} u|_0, |\operatorname{Pref}_{i_j} u|_1) \bmod 2 \neq (0, 0),$$

and hence

$$\{(|\operatorname{Pref}_{i_j} u|_0, |\operatorname{Pref}_{i_j} u|_1) \bmod 2 \mid j \in \mathbb{N}\} \neq \mathbb{Z}_2^2.$$

Example 2.3. We say that an infinite word u over an alphabet $\mathcal{A}$, $|\mathcal{A}| = d$, is *universal* if it contains all finite words over $\mathcal{A}$ as its factors. It is easy to see that any universal word satisfies the WELLDOC property. Indeed, for any word $w \in \mathcal{A}^*$ and any m there exists a finite word v such that if $i_0, i_1, \ldots, i_k$ denote the occurrences of w in v, then

$$\big\{\big(|\operatorname{Pref}_{i_j} v|_0, \ldots, |\operatorname{Pref}_{i_j} v|_{d-1}\big) \bmod m \mid j \in \{0, 1, \ldots, k\}\big\} = \mathbb{Z}_m^d\,.$$

Since u is universal, v is a factor of u. Denoting by i an occurrence of v in u, one gets that the positions $i+i_j$ are occurrences of w in u. Hence

$$\begin{aligned}&\left\{\left(|\mathrm{Pref}_{i+i_j}u|_0,\dots,|\mathrm{Pref}_{i+i_j}u|_{d-1}\right) \bmod m \mid j\in\{0,1,\dots,k\}\right\}=\\&=\left(|\mathrm{Pref}_{i}u|_0,\dots,|\mathrm{Pref}_{i}u|_{d-1}\right)+\\&+\left\{\left(|\mathrm{Pref}_{i_j}v|_0,\dots,|\mathrm{Pref}_{i_j}v|_{d-1}\right) \bmod m \mid j\in\{0,1,\dots,k\}\right\}=\mathbb{Z}_m^d\,.\end{aligned}$$

Therefore, u satisfies the WELLDOC property.

2.2. **Combination of PRNGs.** In order to eliminate the lattice structure, it helps to combine PRNGs in a smart way. Such a method was introduced in [14]. Let $X=(X_n)_{n\in\mathbb{N}}$ and $Y=(Y_n)_{n\in\mathbb{N}}$ be two PRNGs with the same output $M\subset\mathbb{N}$ and the same period $m\in\mathbb{N}$, and let $u=u_0u_1u_2\dots$ be a binary infinite word over the alphabet $\{0,1\}$.

The PRNG $Z=(Z_n)_{n\in\mathbb{N}}$ based on u is obtained by the following algorithm:

1. Read step by step the letters of u.
2. When you read 0 for the i-th time, copy the i-th symbol from X to the end of the constructed sequence Z.
3. When you read 1 for the i-th time, copy the i-th symbol from Y to the end of the constructed sequence Z.

This construction can be generalized for non-binary alphabets: Using infinite words over a multiliteral alphabet, one can combine more than two PRNGs. Remark that following terminology from [3], the sequence Z is obtained as a *shuffle* of the sequences X and Y with the steering word u.

In order to distinguish between generators and infinite words used for their combination, we always denote generators with capital letters $X,Y,Z,\dots$ and words with lower-case letters u,v,w (the same convention is applied for their outputs: $A,B,\dots$ for output values of generators (elements of M), $a,b,\dots$ for letters of words). Finite sequences of successive elements $\bar{\mathbf{x}}=(X_i,X_{i+1},\dots,X_{i+t-1})$ of a PRNG X are called t-tuples, or vectors, while in the case of an infinite word u, we call $u_iu_{i+1}\dots u_{i+t-1}$ a factor of length t.

2.3. **The WELLDOC Property and Absence of the Lattice Structure.** Guimond et al. in [15] have shown that PRNGs based on infinite words coding a certain class of cut-and-project sets have no lattice structure. In the sequel, we will generalize their result and find larger classes of words guaranteeing no lattice structure for associated generators. We focus on the binary alphabet, although everything works for multiliteral words as well (and for combination of more generators therefore), since the proofs become more technical in non-binary case.

Theorem 2.4. *Let Z be the PRNG based on a binary infinite word u with the WELLDOC property. Then Z has no lattice structure.*

Proof. According to Proposition 1.1, it suffices to check that its assumptions are met. Let $A,B\in M$ and $\ell\in\mathbb{N}$. Assume $A=X_i$ and $B=Y_j$, where $X=(X_n)_{n\in\mathbb{N}}$ and $Y=(Y_n)_{n\in\mathbb{N}}$ are the two combined PRNGs with the same output $M\subset\mathbb{N}$ and the same period $m\in\mathbb{N}$. Consider a right special factor w of u of length ℓ, i.e., both words $w0$ and $w1$ are factors of u (such a factor w exists since u is an aperiodic word because of the WELLDOC property). By Definition 2.1, it is possible to find an occurrence i_k of $w0$ in u such that

$$|\,\mathrm{Pref}_{i_k}u|_0=i-|w|_0-1 \bmod m,\qquad |\,\mathrm{Pref}_{i_k}u|_1=j-|w|_1-1 \bmod m.$$

Reading the word $w0$ at the occurrence i_k, the corresponding ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of the generator Z consists of symbols

$$X_{(i-|w|_0) \bmod m}, \ldots, X_{(i-1) \bmod m} \text{ and } Y_{(j-|w|_1) \bmod m}, \ldots, Y_{(j-1) \bmod m}.$$

When reading 0 after w, the symbol $X_i = A$ from the first generator follows $(A_1, A_2, \ldots, A_\ell)$.

Again, by Definition 2.1, there exists an occurrence i_s of $w1$ in u such that

$$|\operatorname{Pref}_{i_s} u|_0 = i - |w|_0 - 1 \bmod m, \qquad |\operatorname{Pref}_{i_s} u|_1 = j - |w|_1 - 1 \bmod m.$$

When reading the word w at the occurrence i_s, the same ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of Z as previously occurs. This time, however, $(A_1, A_2, \ldots, A_\ell)$ is followed by B because we read $w1$ and $Y_j = B$. Thus, we have found an ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of Z followed in Z by both A and B. □

Remark 2.5. The WELLDOC property is sufficient, but not necessary for absence of the lattice structure. For example, consider a modified Fibonacci word $\hat{u}$ where the letter 2 is inserted after each letter, i.e., $\hat{u} = 0212020212021202\ldots$. It is easy to verify that $\hat{u}$ does not have well distributed occurrences. However, we will show the following: Let Z be the PRNG combining three generators $X = (X_n)_{n\in\mathbb{N}}, Y = (Y_n)_{n\in\mathbb{N}}$ and $V = (V_n)_{n\in\mathbb{N}}$ with the same output $M \subset \mathbb{N}$ and the same period $m \in \mathbb{N}$ according to the modified Fibonacci word $\hat{u}$. Then Z has no lattice structure.

It suffices to verify assumptions of Proposition 1.1. Let $A, B \in M$ and $\ell \in \mathbb{N}$, ℓ an even number (the proof is analogous for odd ℓ). Assume $A = X_i$ and $B = Y_j$. Consider a right special factor w of the Fibonacci word u of length $\ell/2$. Since u has the WELLDOC property, there exists an occurrence i_k of $w0$ in u such that

$$|\operatorname{Pref}_{i_k} u|_0 = i - |w|_0 - 1 \bmod m, \qquad |\operatorname{Pref}_{i_k} u|_1 = j - |w|_1 - 1 \bmod m.$$

Then if we insert the letter 2 after each letter of w, we obtain a right special factor $\hat{w}$ of the modified Fibonacci word $\hat{u}$ of length ℓ. It holds then that

$$\begin{aligned}
|\operatorname{Pref}_{2i_k} \hat{u}|_0 &= i - |w|_0 - 1 \bmod m = i - |\hat{w}|_0 - 1 \bmod m,\\
|\operatorname{Pref}_{2i_k} \hat{u}|_1 &= j - |w|_1 - 1 \bmod m = j - |\hat{w}|_1 - 1 \bmod m,\\
|\operatorname{Pref}_{2i_k} \hat{u}|_2 &= i - |w|_0 - 1 + j - |w|_1 - 1 \bmod m = i + j - |\hat{w}|_2 - 2 \bmod m.
\end{aligned}$$

When reading the word $\hat{w}0$ at the occurrence $2i_k$, the corresponding ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of the generator Z is followed by the symbol $X_i = A$ from the first generator.

Again, by the WELLDOC property of u, there exists an occurrence i_s of $w1$ in u such that

$$|\operatorname{Pref}_{i_s} u|_0 = i - |w|_0 - 1 \bmod m, \qquad |\operatorname{Pref}_{i_s} u|_1 = j - |w|_1 - 1 \bmod m.$$

It holds then that

$$\begin{aligned}
|\operatorname{Pref}_{2i_s} \hat{u}|_0 &= i - |w|_0 - 1 \bmod m = i - |\hat{w}|_0 - 1 \bmod m,\\
|\operatorname{Pref}_{2i_s} \hat{u}|_1 &= j - |w|_1 - 1 \bmod m = j - |\hat{w}|_1 - 1 \bmod m,\\
|\operatorname{Pref}_{2i_s} \hat{u}|_2 &= i - |w|_0 - 1 + j - |w|_1 - 1 \bmod m = i + j - |\hat{w}|_2 - 2 \bmod m.
\end{aligned}$$

When reading the word $\hat{w}$ at the occurrence $2i_s$, the same ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of Z as previously occurs. This time, however, $(A_1, A_2, \ldots, A_\ell)$ is followed by B because we read $\hat{w}1$ and $Y_j = B$. Thus, we have found an ℓ-tuple $(A_1, A_2, \ldots, A_\ell)$ of Z followed in Z by both A and B. Therefore Z has no lattice structure.

Remark 2.6. In the proof of Theorem 2.4, the modulus m from the WELLDOC property is set to be equal to the period of the combined generators. Therefore, if we require absence of the lattice structure for a PRNG obtained when combining PRNGs with a fixed period $\hat{m}$, then it is sufficient to use an infinite word u that satisfies the WELLDOC property for the modulus $m = \hat{m}$. This means for instance that the Thue-Morse word is not completely out of the game, but it cannot be used to combine periodic PRNGs with the period being a power of 2.

We have formulated a combinatorial condition – well distributed occurrences – guaranteeing no lattice structure of the associated generator. It is now important to find classes of words satisfying such a condition.

3. Sturmian Words

In this section we show that Sturmian words have well distributed occurrences.

Definition 3.1. An aperiodic infinite word u is called *Sturmian* if its factor complexity satisfies $\mathcal{C}_u(n) = n + 1$ for all $n \in \mathbb{N}$.

So, Sturmian words are by definition binary and they have the lowest possible factor complexity among aperiodic infinite words. Sturmian words admit various types of characterizations of geometric and combinatorial nature. One of such characterizations is via irrational rotations on the unit circle. In [19] Hedlund and Morse showed that each Sturmian word may be realized measure-theoretically by an irrational rotation on the circle. That is, every Sturmian word is obtained by coding the symbolic orbit of a point on the circle of circumference one under a rotation R_α by an irrational angle[1] α, $0 < \alpha < 1$, where the circle is partitioned into two complementary intervals, one of length α and the other of length $1 - \alpha$. Conversely, each such coding gives rise to a Sturmian word.

Definition 3.2. The *rotation* by angle α is the mapping R_α from $[0, 1)$ (identified with the unit circle) to itself defined by $R_\alpha(x) = \{x + \alpha\}$, where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of x. Considering a partition of $[0, 1)$ into $I_0 = [0, 1 - \alpha)$, $I_1 = [1 - \alpha, 1)$, define a word

$$s_{\alpha,\rho}(n) = \begin{cases} 0 & \text{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_0, \\ 1 & \text{if } R_\alpha^n(\rho) = \{\rho + n\alpha\} \in I_1. \end{cases}$$

One can also define $I'_0 = (0, 1 - \alpha]$, $I'_1 = (1 - \alpha, 1]$, the corresponding word is denoted by $s'_{\alpha,\rho}$.

Remark that some but not all Sturmian words are morphic. In fact, it is known that a characteristic Sturmian word (i.e., $\rho = \alpha$) is morphic if and only if the continuous fraction expansion of α is periodic. For more information on Sturmian words we refer to [16, Chapter 2].

Theorem 3.3. *Let u be a Sturmian word. Then u has the WELLDOC property.*

Proof. In the proof we use the definition of Sturmian word via rotation. The main idea is controlling the number of 1's modulo m by taking circle of length m, and controlling the length taking the rotation by $m\alpha$.

For the proof we will use an equivalent reformulation of the theorem:

[1] Measured by arc length (thus equivalent to $2\pi\alpha$ radians).

Let u be a Sturmian word on $\{0,1\}$, for any natural number m and any factor w of u let us denote $i_0, i_1, \ldots$ the occurrences of w in u. Then

$$\left\{\left(i_j, |\operatorname{Pref}_{i_j} u|_1\right) \bmod m \mid j \in \mathbb{N}\right\} = \mathbb{Z}_m^2.$$

That is, we control the number of 1's and the length instead of the number of 0's.

Since a Sturmian word can be defined via rotations by an irrational angle on a unit circle, without loss of generality we may assume that $u = s_{\alpha,\rho}$ for some $0 < \alpha < 1$, $0 \leq \rho < 1$, α irrational (see Definition 3.2). Equivalently, we can consider m copies of the circle connected into one circle of length m with m intervals I_1^i of length α corresponding to 1. The Sturmian word is obtained by rotation by α on this circle of length m (see Fig. 2).

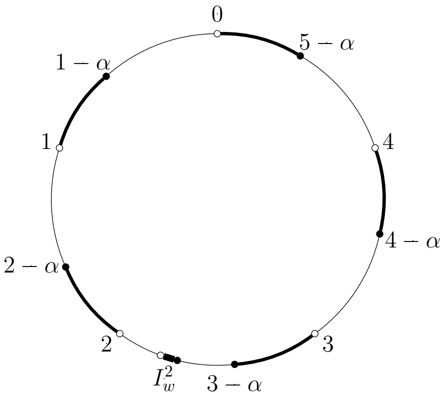


Figure 2. Illustration to the proof of Theorem 3.3: the example for $m = 5$.

Namely, we define the rotation $R_{\alpha,m}$ as the mapping from $[0,m)$ (identified with the circle of length m) to itself defined by $R_{\alpha,m}(x) = \{x+\alpha\}_m$, where $\{x\}_m = x - \lfloor x/m \rfloor m$ and for $m = 1$ coincides with the fractional part of x. A partition of $[0,m)$ into $2m$ intervals $I_0^i = [i, i+1-\alpha)$, $I_1^i = [i+1-\alpha, i+1)$, $i = 0, \ldots, m-1$ defines the Sturmian word $u = s_{\alpha,\rho}$:

$$s_{\alpha,\rho}(n) = \begin{cases} 0 & \text{if } R_{\alpha,m}^n(\rho) = \{\rho + n\alpha\}_m \in I_0^i \text{ for some } \ i = 0, \ldots, m-1, \\ 1 & \text{if } R_{\alpha,m}^n(\rho) = \{\rho + n\alpha\}_m \in I_1^i \text{ for some } \ i = 0, \ldots, m-1. \end{cases}$$

It is well known that any factor $w = w_0 \cdots w_{k-1}$ of u corresponds to an interval I_w in $[0,1)$, so that whenever you start rotating from the interval I_w, you obtain w. Namely, $x \in I_w$ if and only if $x \in I_{w_0}, R_\alpha(x) \in I_{w_1}, \ldots, R_\alpha^{|w|-1}(x) \in I_{w_{|w|-1}}$.

Similarly, we can define m intervals corresponding to w in $[0,m)$ (circle of length m), so that if $I_w = [x_1, x_2)$, then $I_w^i = [x_1 + i, x_2 + i)$, $i = 0, \ldots, m-1$.

Fix a factor w of u, take arbitrary $(j,i) \in \mathbb{Z}_m^2$. Now let us organize (j,i) among the occurrences of w, i.e., find l such that $u_l \ldots u_{l+|w|-1} = w$, $l \mod m = j$ and $|\mathrm{Pref}_l u|_1 \mod m = i$:

Consider rotation $R_{m\alpha,m}(x)$ by $m\alpha$ instead of rotation by α, and start m-rotating from $j\alpha + \rho$. Formally, $R_{m\alpha,m}(x) = \{x + m\alpha\}_m$, where, as above, $\{x\}_m = x - [x/m]m$. This rotation will put us to positions $mk+j$, $k \in \mathbb{N}$, in the Sturmian word: for $a \in \{0, 1\}$ one has $s_{\alpha,\rho}(mk + j) = a$ if $R^k_{m\alpha,m}(j\alpha + \rho) = \{j\alpha + \rho + km\alpha\}_m \in I^i_a$ for some $i = 0, \ldots, m - 1$.

Remark that the points in the orbit of an m-rotation of a point on the m-circle are dense, and hence the rotation comes infinitely often to each interval. So pick k when $j\alpha + mk\alpha + \rho \in I^i_w \subset [i, i + 1)$ (and actually there exist infinitely many such k). Then the length l of the corresponding prefix is equal to $km + j$, and the number of 1's in it is $i + mp$, where p is the number of complete circles you made, i.e., $p = [(j\alpha + mk\alpha + \rho)/m]$.

□

4. Arnoux-Rauzy Words

In this section we show that Arnoux-Rauzy words [1], which are natural extensions of Sturmian words to larger alphabets, also satisfy the WELLDOC property. Note that the proof for Sturmian words cannot be generalized to Arnoux-Rauzy words, because it is based on the geometric interpretation of Sturmian words via rotations, while this interpretation does not extend to Arnoux-Rauzy words.

4.1. Basic Definitions.

The definitions and results we remind in this subsection are well-known and mostly taken from [1, 9] and generalize the ones given for binary words in [5].

Definition 4.1. Let $\mathcal{A}$ be a finite alphabet. The *reversal operator* is the operator $\sim$: $\mathcal{A}^* \mapsto \mathcal{A}^*$ defined by recurrence in the following way:

$$\tilde{\varepsilon} = \varepsilon, \quad \widetilde{va} = a\tilde{v}$$

for all $v \in \mathcal{A}^*$ and $a \in \mathcal{A}$. The fixed points of the reversal operator are called *palindromes.*

Definition 4.2. Let $v \in \mathcal{A}^*$ be a finite word over the alphabet $\mathcal{A}$. The *right palindromic closure* of v, denoted by $v^{(+)}$, is the shortest palindrome that has v as a prefix. It is readily verified that if p is the longest palindromic suffix of $v = wp$, then $v^{(+)} = wp\tilde{w}$.

Definition 4.3. We call the *iterated (right) palindromic closure operator* the operator ψ recurrently defined by the following rules:

$$\psi(\varepsilon) = \varepsilon, \quad \psi(va) = (\psi(v)a)^{(+)}$$

for all $v \in \mathcal{A}^*$ and $a \in \mathcal{A}$. The definition of ψ may be extended to infinite words u over $\mathcal{A}$ as $\psi(u) = \lim_n \psi(\mathrm{Pref}_n\, u)$, i.e., $\psi(u)$ is the infinite word having $\psi(\mathrm{Pref}_n\, u)$ as its prefix for every $n \in \mathbb{N}$.

Definition 4.4. Let Δ be an infinite word on the alphabet $\mathcal{A}$ such that every letter occurs infinitely often in Δ. The word $c = \psi(\Delta)$ is then called a *characteristic (or standard) Arnoux-Rauzy word* and Δ is called the *directive sequence* of c. An infinite word u is called an Arnoux-Rauzy word if it has the same set of factors as a (unique) characteristic Arnoux-Rauzy word, which is called the characteristic word of u. The directive sequence of an Arnoux-Rauzy word is the directive sequence of its characteristic word.

Let us also recall the following well-known characterization (see e.g. [9]):

Theorem 4.5. *Let u be an aperiodic infinite word over the alphabet $\mathcal{A}$. Then u is a standard Arnoux-Rauzy word if and only if the following hold:*

(1) $\mathrm{Fact}(u)$ *is closed under reversal (that is, if v is a factor of u so is $\tilde{v}$).*
(2) *Every left special factor of u is also a prefix.*
(3) *If v is a right special factor of u then va is a factor of u for every $a \in \mathcal{A}$.*

From the preceding theorem, it can be easily verified that the bispecial factors of a standard Arnoux-Rauzy correspond to its palindromic prefixes (including the empty word), and hence to the iterated palindromic closure of the prefixes of its directive sequence. That is, if

$$\varepsilon = b_0, b_1, b_2, \ldots$$

is the sequence, ordered by length, of bispecial factors of the standard Arnoux-Rauzy word u, $\Delta = \Delta_0\Delta_1\cdots$ its directive sequence (with $\Delta_i \in \mathcal{A}$ for every i), we have $b_{i+1} = (b_i\Delta_i)^{(+)}$.

A direct consequence of this, together with the preceding definitions, is the following statement, which will be used in the sequel.

Lemma 4.6. *Let u be a characteristic Arnoux-Rauzy word and let Δ and $(b_i)_{i\geq 0}$ be defined as above. If Δ_i does not occur in b_i, then $b_{i+1} = b_i\Delta_i b_i$. Otherwise let $j < i$ be the largest integer such that $\Delta_j = \Delta_i$. Then $b_{i+1} = b_i b_j^{-1} b_i$.*

4.2. **Parikh Vectors and Arnoux-Rauzy Factors.** Where no confusion arises, given an Arnoux-Rauzy word u, we will denote by

$$\varepsilon = b_0, b_1, \ldots, b_n, \ldots$$

the sequence of bispecial factors of u ordered by length and we will denote for any $i \in \mathbb{N}$, $\bar{\mathbf{b}}_i$ the Parikh vector of b_i.

Remark 4.7. By the pigeonhole principle, it is clear that for every $m \in \mathbb{N}$ there exists an integer $N \in \mathbb{N}$ such that, for every $i \geq N$, the set $\{j > i \mid \bar{\mathbf{b}}_j \equiv_m \bar{\mathbf{b}}_i\}$ is infinite. Where no confusion arises and with a slight abuse of notation, fixed m, we will always denote by N the smallest of such integers.

Lemma 4.8. *Let u be a characteristic Arnoux-Rauzy word and let $m \in \mathbb{N}$. Let*

$$\alpha_1\bar{\mathbf{b}}_{j_1} + \cdots + \alpha_k\bar{\mathbf{b}}_{j_k} \equiv_m \bar{\mathbf{v}} \in \mathbb{Z}_m^d$$

be a linear combination of Parikh vectors such that $\sum_{i=1}^{k}\alpha_i = 0$, with $j_i \geq N$ and $\alpha_i \in \mathbb{Z}$ for all $i \in \{1, \ldots k\}$. Then, for any $\ell \in \mathbb{N}$, there exists a prefix v of u such that the Parikh vector of v is congruent to $\bar{\mathbf{v}}$ modulo m and vb_ℓ is also a prefix of u.

Proof. Without loss of generality, we can assume $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$, hence there exists k' such that

$$\alpha_1 \geq \alpha_{k'} \geq 0 \geq \alpha_{k'+1} \geq \alpha_k.$$

We will prove the result by induction on $\beta = \sum_{j=1}^{k'}\alpha_j$. If $\beta = 0$, trivially, we can take $v = \varepsilon$ and the statement is clearly verified. Let us assume the statement true for all $0 \leq \beta < n$ and let us prove it for $\beta = n$. By the remark preceding this lemma, for every ℓ we can choose $i' > j' > \ell$ such that $\bar{\mathbf{b}}_{j_1} \equiv_m \bar{\mathbf{b}}_{i'}$ and $\bar{\mathbf{b}}_{j_k} \equiv_m \bar{\mathbf{b}}_{j'}$. Since every bispecial factor is a prefix and suffix of all the bigger ones, in particular we have that $b_{j'}$ is a suffix of $b_{i'}$, and b_ℓ is a prefix of $b_{j'}$; this implies that $b_{i'}b_{j'}^{-1}b_\ell$ is actually a prefix of $b_{i'}$. By assumption, the Parikh vector of $b_{i'}b_{j'}^{-1}$ is clearly $\bar{\mathbf{b}}_{i'} - \bar{\mathbf{b}}_{j'} \equiv_m \bar{\mathbf{b}}_{j_1} - \bar{\mathbf{b}}_{j_k}$. Since $\alpha_1 \geq 1$ implies $\alpha_k \leq -1$, we have, by induction hypothesis, that there exists a prefix w of u such that the Parikh vector of w is congruent modulo m to

$$(\alpha_1 - 1)\bar{\mathbf{b}}_{j_1} + \cdots + (\alpha_k + 1)\bar{\mathbf{b}}_{j_k}$$

and $wb_{i'}$ is a prefix of u. Hence $wb_{i'}b_{j'}^{-1}b_\ell$ is also a prefix of u and, by simple computation, the Parikh vector of $v = wb_{i'}b_{j'}^{-1}$ is congruent modulo m to $\bar{\mathbf{v}} = \alpha_1\bar{\mathbf{b}}_{j_1} + \cdots + \alpha_k\bar{\mathbf{b}}_{j_k}$. □

Definition 4.9. Let $n \in \mathbb{Z}$. We will say that an integer linear combination of integer vectors is a *n-combination* if the sum of all the coefficients equals n.

Lemma 4.10. *Let u be a characteristic Arnoux-Rauzy word and let $n \in \mathbb{N}$. Every n-combination of Parikh vectors of bispecial factors can be expressed as an n-combination of Parikh vectors of arbitrarily large bispecials. In particular, for every $K, L \in \mathbb{N}$, it is possible to find a finite number of integers $\alpha_1, \dots, \alpha_k$ such that $\bar{\mathbf{b}}_K = \alpha_1\bar{\mathbf{b}}_{j_1} + \cdots + \alpha_k\bar{\mathbf{b}}_{j_k}$ with $j_i > L$ for every i and $\alpha_1 + \cdots + \alpha_k = 1$.*

Proof. A direct consequence of Lemma 4.6 is that for every i such that Δ_i appears in b_i, we have $\bar{\mathbf{b}}_{i+1} = 2\bar{\mathbf{b}}_i - \bar{\mathbf{b}}_j$, where $j < i$ is the largest such that $\Delta_j = \Delta_i$. This in turn (since every letter in Δ appears infinitely many times from the definition of Arnoux-Rauzy word) implies that *for every* non-negative integer j, there exists a positive k such that $\bar{\mathbf{b}}_j = 2\bar{\mathbf{b}}_{j+k} - \bar{\mathbf{b}}_{j+k+1}$, that is, we can substitute each Parikh vector of a bispecial with a 1-combination of Parikh vectors of strictly larger bispecials. Simply iterating the process, we obtain the statement. □

In the following we will assume the set $\mathcal{A}$ to be a finite alphabet of cardinality d. For every set $X \subseteq \mathcal{A}^*$ of finite words, we will denote by $\mathrm{PV}(X) \subseteq \mathbb{Z}^d$ the set of Parikh vectors of elements of X and for every $m \in \mathbb{N}$ we will denote by $\mathrm{PV}_m(X) \subseteq \mathbb{Z}_m^d$ the set of elements of $\mathrm{PV}(X)$ reduced modulo m.

For an infinite word u over $\mathcal{A}$, and a factor v of u, let $S_v(u)$ denote the set of all prefixes of u followed by an occurrence of v. In other words,

$$S_v(u) = \{p \in \mathrm{Pref}(u) \mid pv \in \mathrm{Pref}(u)\}.$$

Definition 4.11. For any set of finite words $X \subseteq \mathcal{A}^*$, we will say that u *has the property $\mathcal{P}_X$* (or, for short, that u has $\mathcal{P}_X$) if, for every $m \in \mathbb{N}$ and for every $v \in X$ we have that

$$\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d.$$

That is to say, for every vector $\bar{\mathbf{w}} \in \mathbb{Z}_m^d$ there exists a word $w \in S_v(u)$ such that the Parikh vector of w is congruent to $\bar{\mathbf{w}}$ modulo m.

With this notation, an infinite word u has the WELLDOC property if and only if it has the property $\mathcal{P}_{\mathrm{Fact}(u)}$.

Proposition 4.12. *Let u be a characteristic Arnoux-Rauzy word over the d-letter alphabet $\mathcal{A}$. Then u has the property $\mathcal{P}_{\mathrm{Pref}(u)}$.*

Proof. Let us fix an arbitrary $m \in \mathbb{N}$. We want to show that, for every $v \in \mathrm{Pref}(u)$, $\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d$. Let then $\bar{\mathbf{v}} \in \mathbb{Z}^d$ and ℓ be the smallest number such that v is a prefix of b_ℓ. Let $i_1 < i_2 < \cdots < i_d$ be such that Δ_{i_j} does not appear in b_{i_j}, where Δ

is the directive word of u. Without loss of generality, we can rearrange the letters so that each Δ_{i_j} is lexicographically smaller than $\Delta_{i_{j+1}}$. With this assumption if, for every j, we set $\bar{\mathbf{v}}_j = \bar{\mathbf{b}}_{i_j+1}$, i.e., equal to the Parikh vector of b_{i_j+1}, which, by the first part of Lemma 4.6, equals $b_{i_j}\Delta_{i_j}b_{i_j}$, we can find $j-1$ positive integers $\mu_1,\dots,\mu_{j-1}$ such that $\bar{\mathbf{v}}_j = (\mu_1,\mu_2,\dots,\mu_{j-1},1,0,\dots,0)$. It is easy to show, then, that the set $V = \{\bar{\mathbf{v}}_1,\dots,\bar{\mathbf{v}}_d\}$ generates $\mathbb{Z}^d$, hence there exists an integer n such that $\bar{\mathbf{v}}$ can be expressed as an n-combination of elements of V (which are Parikh vectors of bispecial factors of u). Trivially, then, $\bar{\mathbf{v}} = \bar{\mathbf{v}} - n\bar{\mathbf{0}} = \bar{\mathbf{v}} - n\bar{\mathbf{b}}_0$; thus, it is possible to express $\bar{\mathbf{v}}$ as a 0-combination of Parikh vectors of (by the previous Lemma 4.10) arbitrarily large bispecial factors of u. By Lemma 4.8, then there exists a prefix p of u whose Parikh vector $\bar{\mathbf{p}}$ satisfies $\bar{\mathbf{p}} \equiv_m \bar{\mathbf{v}}$ and pb_ℓ is a prefix of u. Since we picked ℓ such that v is a prefix of b_ℓ, we have that $p \in S_v(u)$. From the arbitrariness of v, $\bar{\mathbf{v}}$ and m, we obtain the statement. □

As a corollary of Proposition 4.12, we obtain the main result of this section.

Theorem 4.13. *Let u be an Arnoux-Rauzy word over the d-letter alphabet $\mathcal{A}$. Then u has the property $\mathcal{P}_{\mathrm{Fact}(u)}$, or equivalently, u has the WELLDOC property.*

Proof. Let m be a positive integer and let c be the characteristic word of u. Let v be a factor of u and xvy be the shortest bispecial containing v. By Proposition 4.12, we have that $\mathrm{PV}_m(S_{xv}(c)) = \mathbb{Z}_m^d$ and, since the set is finite, we can find a prefix p of c such that $\mathrm{PV}_m(S_{xv}(p)) = \mathbb{Z}_m^d$. Let w be a prefix of u such that wp is a prefix of u. If $\bar{\mathbf{x}}$ and $\bar{\mathbf{w}}$ are the Parikh vectors of, respectively, x and w, it is easy to see that

$$\bar{\mathbf{w}} + \bar{\mathbf{x}} + \mathrm{PV}(S_{xv}(p)) \subseteq \bar{\mathbf{w}} + \mathrm{PV}(S_v(p)) \subseteq \mathrm{PV}(S_v(u))$$

Since we have chosen p such that $\mathrm{PV}_m(S_{xv}(p)) = \mathbb{Z}_m^d$, we clearly obtain that $\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d$ and hence, by the arbitrariness of v and m, the statement. □

Remark 4.14. Now we introduce a simple method of obtaining words satisfying the WELLDOC property. Take a word u with the WELLDOC property over an alphabet $\{0,1,\dots,d-1\}$, $d > 2$, apply a morphism $\varphi : d-1 \mapsto 0, i \mapsto i$ for $i = 0,\dots,d-2$, i.e., φ joins two letters into one. It is straightforward that $\varphi(u)$ has the WELLDOC property. So, taking Arnoux-Rauzy words and joining some letters, we obtain other words than Sturmian and Arnoux-Rauzy satisfying the WELLDOC property.

Remark 4.15. Now we introduce another class of morphisms preserving the WELLDOC property. Recall that the *adjacency matrix* Φ of a morphism $\varphi : \mathcal{A} \to \mathcal{A}$, with $\mathcal{A} = \{0,1,\dots,d-1\}$, is defined by $\Phi_{i,j} = |\varphi(j-1)|_{i-1}$ for $1 \leq i,j \leq d$. By definition, it follows that if $\bar{\mathbf{v}}$ is the Parikh vector of $v \in \mathcal{A}^*$, then $\Phi\bar{\mathbf{v}}$ is the Parikh vector of $\varphi(v)$.

Let us show that if $\det\Phi = \pm 1$ and u has the WELLDOC property, then so does $\varphi(u)$. Indeed, let w be any factor of $\varphi(u)$, and suppose $xwy = \varphi(v)$ for some $v \in \mathrm{Fact}(u)$ and $x,y \in \mathcal{A}^*$. We then have $S_w(\varphi(u)) \supseteq \varphi(S_v(u))x$, so that, writing $\bar{\mathbf{x}}$ for the Parikh vector of x, we have for any $m > 0$

$$\mathrm{PV}_m(S_w(\varphi(u))) \supseteq \Phi \cdot \mathrm{PV}_m(S_v(u)) + \bar{\mathbf{x}} \bmod m\,.$$

Since u has the WELLDOC property, $\mathrm{PV}_m(S_v(u)) = \mathbb{Z}_m^d$. As $\det\Phi = \pm 1$, Φ is invertible (even modulo m), so that $\Phi\cdot\mathbb{Z}_m^d + \bar{\mathbf{x}} \bmod m = \mathbb{Z}_m^d$. Hence $\mathrm{PV}_m(S_w(\varphi(u))) =$

$\mathbb{Z}_m^d$, showing that $\varphi(u)$ has the WELLDOC property by the arbitrariness of w and m.

5. Statistical Tests of PRNGs

In the previous part, we have explained that PRNGs based on infinite words with well distributed occurrences have no lattice structure. In this sequel we demonstrate this by empirical statistical tests. We have chosen to use LCGs as underlying generators *explicitly* for their known weaknesses. We will show how mixing based on aperiodic infinite words will cope with these weaknesses and whether statistical tests will show any significant improvements.

5.1. **Computer Generation of Morphic Words.** Any real computer is a finite state machine and hence it can generate only finite prefixes of infinite words. From practical point of view it is important to find algorithms that are efficient both in memory footprint and CPU time. In [20] an efficient algorithm for generating the Fibonacci word was introduced: The prefix of length n is generated in $O(\log(n))$ space and $O(n)$ time. We generalize this method for any Sturmian and Arnoux-Rauzy word being a fixed point of a morphism φ. The main ingredient is that we consider φ^n instead of φ; we precompute and store in the memory $\varphi^n(a)$ for any $a \in \mathcal{A}$. The runtime to generate 10^{10} letters of the Fibonacci and the Tribonacci word is summarized in Table 1. There are the following observations we would like to point out:

(1) There is no need to store the first n letters in memory to generate the $(n+1)$-th letter. Letters are generated on the fly and only nodes of the traversal tree are kept in the memory. Memory consumption needed to generate the first 10^{10} letters is shown in Table 1. The algorithm also supports leap frogging, generation can be started at any position in the word. The consequence is that the algorithm can be easily parallelized to produce multiple streams [11].

(2) Using the method from [20] together with our improvement for generation of Sturmian and Arnoux-Rauzy words, the speed of generation of their prefixes is much higher than the speed of generation of LCGs output values. For example, generation of 10^{10} 32-bit values using a LCG modulo 2^{64} takes 14.3 seconds on our machine. Compare it to 0.5 seconds for generation of 10^{10} letters of a fixed point of a morphism with the same hardware. Thus, using a fixed point to combine LCGs causes only a negligible runtime penalty.

(3) The speed of generation can be further improved by using a higher initial memory footprint and CPU that can effectively copy such larger chunks of memory (size of L1 data cache is a limiting factor). Thus the new method scales nicely and can benefit form the future CPUs with higher L1 caches. The only requirement is to precompute $\varphi^n(a)$, $a \in \mathcal{A}$, for larger n. Our program does this automatically based on the limit on the initial memory consumption provided by the user.

5.2. **Testing PRNGs Based on Sturmian and Arnoux-Rauzy words.** We will present results for PRNGs based on:

- the Fibonacci word (as an example of a Sturmian word), i.e., the fixed point of the morphism $0 \mapsto 01, 1 \mapsto 0$,

Word	Fibonacci	Tribonacci
φ morphism rule	115s / 336 Bytes	107s / 256 Bytes
φ^n morphism rule	0.41s / 32 Bytes	0.36s / 32 Bytes

Table 1. The comparison of time in seconds and memory consumption to hold the traversal tree state needed to generate the first 10^{10} letters of the Fibonacci and the Tribonacci word using the original [20] (1st line) and the new algorithm (2nd line). The iteration n in the φ^n rule was chosen so that the length of $\varphi^n(a)$ does not exceed 4096 bytes for any $a \in \mathcal{A}$. The measurement was done on Intel Core i7-3520M CPU running at 2.90GHz.

- the modified Fibonacci word – Fibonacci2 – with the letter 2 inserted after each letter (see Remark 2.5),
- the Tribonacci word (as the simplest example of a ternary Arnoux-Rauzy word), i.e., the fixed point of $0 \mapsto 01, 1 \mapsto 02, 2 \mapsto 0$.

We have implemented PRNGs for more morphic Sturmian words and ternary Arnoux-Rauzy words. Since the results are similar, we present in the sequel only the above three representatives. Our program generating PRNGs based on morphic words is available online, together with a description [4].

Remark that we included the modified Fibonacci word that does not have the WELLDOC property, but at the same time it guarantees no lattice structure for the arising generator. The reason for including it is that we would like to illustrate that such a word leads to worse results in testing than the Fibonacci word.

5.2.1. *Combining LCGs.* Instead of combining plain LCGs, we will execute some modifications before their combination. Those modifications turn out to be useful according to the known weaknesses of LCGs.

We have chosen LCGs with the period m in range from $2^{47} - 115$ to 2^{64}, but we use only their upper 32 bits as the output since the statistical tests require 32-bit sequences as the input. Their output is thus in all cases $M = \{0, 1, \dots, 2^{32} - 1\}$.

We use two batteries of random tests – TestU01 BigCrush and PractRand. They operate differently. The first one includes 160 statistical tests, many of them tailored to the specific classes of PRNGs. It is a reputable test, however its drawback is that it works with a fixed amount of data and discards the least significant bit (for some tests even two bits) of the 32-bit numbers being tested. The second battery consists of three different tests where one is adapted on short range correlations, one reveals long range violations, and the last one is a variation on the classical Gap test. Details can be found in [7, 8]. Moreover, the PractRand battery applies automatically various filters on the input data. For our purpose the lowbit filter is interesting – it is passing various number of the least significant bits to the statistical tests. As we have already mentioned, the LCGs with $m = 2^\ell$ have a much shorter period than the LCG itself. Therefore the lowbit filter is useful to check whether this weakness disappears when LCGs are combined according to an infinite word. The PractRand tests are able to treat very long input sequences, up to a few exabytes. To control the runtime we have limited the length of input sequences to 16TB.

The first column of Table 2 shows the list of tested LCGs. The BigCrush column shows how many tests of the TestU01 BigCrush battery failed. The PractRand column gives the $\log_2$ of sample datasize in Bytes for which the results of the PractRand tests started to be "very suspicious" (p-values smaller than 10^{-5}). One LCG did not show any failures in the PractRand tests which is denoted as > 44 – the meaning is that the PractRand test has passed successfully 16TB of input data and the test was stopped there. The last column provides time in seconds to generate the first 10^{10} 32-bit sequences of output on Intel i7-3520M CPU running at 2.90GHz.

Generator	Legend	BigCrush	PractRand	Time 10^{10}
$\mathrm{LCG}(2^{47} - 115, 71971110957370, 0)$	L47-115	14	40	281
$\mathrm{LCG}(2^{63} - 25, 2307085864, 0)$	L63-25	2	>44	277
$\mathrm{LCG}(2^{59}, 13^{13}, 0)$	L59	19	27	14.1
$\mathrm{LCG}(2^{63}, 5^{19}, 1)$	L63	19	33	14.4
$\mathrm{LCG}(2^{64}, 2862933555777941757, 1)$	L64_28	18	35	14.0
$\mathrm{LCG}(2^{64}, 3202034522624059733, 1)$	L64_32	14	34	14.1
$\mathrm{LCG}(2^{64}, 3935559000370003845, 1)$	L64_39	13	33	14.0

Table 2. List of the used LCGs with parameters $\mathrm{LCG}(m, a, c)$. Results in the BigCrush (number of failed tests) and in the PractRand ($\log_2$ of sample size for which the test started to fail) battery of statistical tests. Time in seconds to generate the first 10^{10} 32-bit words of output on Intel i7-3520M CPU running at 2.90GHz.

From Table 2 it can be seen that the LCGs with $m \in \{2^{47} - 115, 2^{63} - 25\}$ have the best statistical properties from the chosen LCGs. At the same time, these LCGs are 20 times slower than the other LCGs used. This is because we have used 128-bit integer arithmetic to compute their internal state and because explicit modulo operation cannot be avoided. As the CPU used does not have the 128-bit integer arithmetic, it has to be implemented in software (in this case via GCC's `__int128` type) which is much slower than the 64-bit arithmetic wired on CPU.

5.2.2. *Results in Statistical Tests.* We will present results for the PRNGs based on the Fibonacci, Fibonacci2 and Tribonacci word using the different combinations of LCGs from Table 2. It includes also the situations where the instances of the same LCG are used. Each instance has its own state. The LCGs were seeded with the value 1. The PRNGs were warmed up by generating 10^9 values before statistical tests started. Since the relative frequency of the letters in the aperiodic words differ a lot (for example for the Fibonacci word the ratio of zeroes to ones is given by $\tau = \frac{1+\sqrt{5}}{2}$), the warming procedure will guarantee that the state of instances of LCGs will differ even when the same LCGs are used. Even more importantly, the distance between the LCGs is growing as the new output of PRNGs is generated.

Summary of results is in Table 3. The BigCrush column is using the following notation: the first number indicates how many tests from the BigCrush battery have clearly failed and the optional second number in parenthesis denotes how many tests have suspiciously low p-value in the range from 10^{-6} to 10^{-4}. The PractRand column gives the $\log_2$ of sample datasize in Bytes for which the results of the PractRand tests started to be “very suspicious” (p-values smaller than 10^{-5}). The maximum sample data size used was 16TB $\doteq 2^{44}$B. The Time column gives runtime in seconds to generate the first 10^{10} 32-bit words of output on Intel i7-3520M CPU running at 2.90GHz. The source code of the testing programs is in [4].

Word	Group	0	1	2	BigCrush	PractRand	Time 10^{10}
Fib	A	L64_28	L64_28		0	41	30.2
		L64_32	L64_28		0(1)	41	29.3
		L64_39	L64_28		0 (2)	41	31
		L64_28	L64_32		0	41	30.2
		L64_32	L64_32		0	41	30.1
		L64_39	L64_32		0	41	30.1
		L64_28	L64_39		0	42	30.2
		L64_32	L64_39		0	40	30.5
		L64_39	L64_39		0	42	30.1
	B	L47-115	L47-115		1(1)	>44	302
		L63-25	L63-25		0(1)	>44	299
		L59	L59		0(1)	34	28.7
		L63	L63		0	40	29.8
	C	L63-25	L59		0	38	198
		L59	L63-25		0(1)	35	134
		L63-25	L64_39		0	>44	199
		L64_39	L63-25		0	41	135
		L59	L64_39		0	35	30.4
		L64_39	L59		0	37	31.3
Fib2	A	L64_28	L64_28	L64_28	0	40	28.4
		L64_39	L64_28	L64_28	0(2)	40	27.9
		L64_39	L64_32	L64_28	0	39	27.5
		L64_28	L64_39	L64_28	0	40	27.3
		L64_32	L64_39	L64_28	0	40	27.5
		L64_39	L64_39	L64_28	0	40	27.4
		L64_39	L64_28	L64_32	0	40	27.3
		L64_28	L64_39	L64_32	0	40	27.9
		L64_28	L64_28	L64_39	0(1)	40	27.4
		L64_32	L64_28	L64_39	0	39	27.7
		L64_39	L64_28	L64_39	0	40	27.3
		L64_28	L64_32	L64_39	0	40	27.3
		L64_28	L64_39	L64_39	0	40	27.3
		L64_39	L64_39	L64_39	0	40	27.4
	B	L47-115	L47-115	L47-115	0(2)	>44	297.0
		L63-25	L63-25	L63-25	0(2)	>44	293.0
		L59	L59	L59	0(1)	32	27.4
		L63	L63	L63	0	38	27.3
	C	L63-25	L59	L64_39	0(1)	39	113.0
		L63-25	L64_39	L59	0	32	113.0
		L59	L63-25	L64_39	0	38	81.1
		L59	L64_39	L63-25	0	39	158.3
		L64_39	L63-25	L59	0	31	81.0

Continued on the next page

Table 3 – Continued from the previous page

Word	Group	0	1	2	BigCrush	PractRand	Time 10^{10}
		L64_39	L59	L63-25	0	42	159.0
Trib	A	L64_28	L64_28	L64_28	0(2)	42	27.2
		L64_39	L64_28	L64_28	0	43	27.1
		L64_39	L64_32	L64_28	0(1)	42	28.0
		L64_28	L64_39	L64_28	0(1)	42	28.1
		L64_32	L64_39	L64_28	0	42	27.1
		L64_39	L64_39	L64_28	0(1)	42	27.2
		L64_39	L64_28	L64_32	0	43	27.1
		L64_28	L64_39	L64_32	0(1)	42	27.1
		L64_28	L64_28	L64_39	0	42	28.0
		L64_32	L64_28	L64_39	0	42	27.2
		L64_39	L64_28	L64_39	0(1)	43	27.1
		L64_28	L64_32	L64_39	0	43	27.1
		L64_28	L64_39	L64_39	0(2)	42	27.3
		L64_39	L64_39	L64_39	0	43	27.1
	B	L47-115	L47-115	L47-115	1	>44	299.0
		L63-25	L63-25	L63-25	0(1)	>44	298.0
		L59	L59	L59	0	35	27.2
		L63	L63	L63	0(1)	41	27.2
	C	L63-25	L59	L64_39	0(1)	39	172.0
		L63-25	L64_39	L59	0(1)	41	173.0
		L59	L63-25	L64_39	0	35	106.0
		L59	L64_39	L63-25	0	34	70.5
		L64_39	L63-25	L59	0	41	107.0
		L64_39	L59	L63-25	0(1)	40	74.3

Table 3. Summary of results of statistical tests for PRNGs based on the Fibonacci, Fibonacci2 and Tribonacci word and different combinations of LCGs from Table 2.

We can make the following observations based on the results in statistical tests:

(1) The quality of LCGs has improved substantially when we combined them according to infinite words with the WELLDOC property. This can be seen in the TestU01 BigCrush results. While for LCGs 13 to 19 tests have clearly failed (the only exception is the generator L63-25 with two failures – see Table 2), almost all of the BigCrush tests passed. The worst result was to have one BigCrush test failed for the Tribonacci combination and one for the Fibonacci combination of L47-115 generators. The likely reason is that the generator L47-115 has the shortest period of all tested LCGs.

(2) The results of the PractRand battery confirm the above findings. For instance, in the case of LCGs with modulo 2^{64}, the test started to find irregularities in the distribution of the least significant bit of tested PRNGs output at around 2TB sample size. Compare it with the sample size of 8GB to 32GB when fast plain LCGs started to fail the test. The PractRand battery applies different filters on the input stream and all failures appeared for `Low1/32` filter where only the least significant bit of the PRNG output

is used. It corresponds to a known weakness of power-of-2 modulo LCGs: lower bits of the output have significantly smaller period than the LCG itself. The quality of the PRNGs can be therefore further improved by combining LCGs that do not show flaws for the least significant bits or by using for example just 16 upper bits of the LCGs output.

(3) The quality of the PRNG is linked to the quality of the underlying LCG. When looking at the group B in Table 3, we observe that the PractRand results of the arising PRNGs are closely related to the succes of LCGs from Table 2 in the PractRand tests.

(4) Another interesting observation is that using the instances of the same LCG (with only sufficiently distinct seeds) produces as good results as combination of different LCGs (multipliers and shifts are different, but the modulus is the same). It is just important to make sure that starting states of the LCGs are far apart enough. Refer to the group A in Table 3.

(5) The lower quality LCG dictates the quality of resulting PRNG. When mixing LCGs with different quality, use better ones as replacement for more frequent letters in the aperiodic word.

Please refer to the group C in Table 3. For example for the Fibonacci word compare first two rows in the group C - the order of LCGs is merely swapped but the difference in the sample size for which PractRand starts to fail is 8×. This is even more significant for the Tribonacci based generators where the difference between the worst and best PractRand results when reordering the underlying LCGs is given by factor 128×.

(6) On the other hand, results from the group A in Table 3 demonstrate that when using generators of similar quality (same modulus, similar deficiencies), the order in which generators are used to substitute the letters of the infinite word does not influence the quality of the resulting generator.

(7) We can also see that the modified Fibonacci word (see Remark 2.5) does not produce better results than the Fibonacci word. Clearly, a regular structure of 2's on every other position does not help to produce a better random sequence even if we mix now three LCGs instead of two as in the case of the Fibonacci word.

(8) Results for the Tribonacci word are better than for the Fibonacci word. (We have observed this fact for all ternary Arnoux-Rauzy words in comparison to Sturmian words.) It seems therefore that mixing three LCGs is better than using just two LCGs, assuming that an infinite word with the WELLDOC property is used for mixing. We expect naturally that the better chosen LCGs (or even some other modern fast linear PRNGs, e.g. *mt19937* or nonlinear PRNGs based on the AES cipher) we combine according to an infinite word with the WELLDOC property, the better their results in statistical tests will be.

(9) We have also tested LCGs with $m = 2^{31} - 1$. It has revealed that if the underlying generators have poor statistical properties, then the PRNG will not be able to mask it. In particular, you cannot expect that PRNGs – despite their infinite aperiodic nature – will fix the short period problem. Once the period of the underlying LCG is exhausted, statistical tests will find irregularities in the output of the PRNG.

In conclusion, we summarize the main results from the user point of view:

- Using different instances of the same LCG to form a new generator based on the infinite word with the WELLDOC property gives a generator with improved statistical properties.
- The introduced method of generation of morphic words is very fast and supports parallel processing.
- The period of underlying generators has to be large enough – much larger than the number of needed values.
- When using different types of the underlying LCGs to form a PRNG, close attention has to be paid to the right order of the combined LCGs. The generator with the worst properties should be used to replace the least frequent letter of the aperiodic word. Moreover, statistical properties of the resulting PRNG are ruled by the deficiencies of the worst used generator.
- We have used the LCGs only for study reasons. Instead of LCGs, the modern generators (of user choice) could be used as underlying PRNG to obtain better results. We have done testing with two instances (respectively three for the Tribonacci and other Arnoux-Rauzy words) of Mersenne twister 19937 as the underlying generator. The newly constructed generator has passed all the empirical tests on randomness we have executed (in contrary to Mersenne twister 19937 itself which is failing two tests from TestU01's BigCrush battery). For the practical usage Arnoux-Rauzy (AR) words are very appealing since there is an infinite number of AR words and we have implementation in place to create the AR words based on user input (it can be sought of as the seed). Thus, we recommend to create new PRNGs based on one's favorite modern PRNGs and the custom AR word.

6. Open problems and future research

Concerning the combinatorial part of our paper, one of the interesting open questions there is finding large families of infinite words satisfying the WELLDOC property. For example, which morphic words have the WELLDOC property? Also, it seems to be meaningful to study a weaker WELLDOC property where in Definition 2.1 instead of every $m \in \mathbb{N}$ we consider only a particular m. For instance, one can search for words satisfying such a modified WELLDOC condition for $m = 2$, $m = 2^\ell$ etc. Another question to be asked is how to construct words with the WELLDOC property over larger alphabets using words with such a property over smaller alphabets. Regarding statistical tests, it remains to explain why PRNGs based on infinite words with the WELLDOC property succeed in tests and to compare their results with other comparably fast generators.

Acknowledgements

The first author was supported by the Czech Science Foundation grant GAČR 13-03538S, and thanks L'Oréal Czech Republic for the Fellowship Women in Science. The third author was partially supported by the Italian Ministry of Education (MIUR), under the PRIN 2010–11 project "Automi e Linguaggi Formali: Aspetti Matematici e Applicativi". The fifth author was supported in part by the Academy of Finland under grant 251371 and by Russian Foundation of Basic Research (grants 12-01-00089 and 12-01-00448).

References

1. P. Arnoux, G. Rauzy, *Représentation géométrique de suites de complexité* $2n+1$, Bull. Soc. Math. France **119** (1991), 199–215.
2. L. Balková, M. Bucci, A. De Luca, S. Puzynina, *Infinite Words with Well Distributed Occurrences*. In: J. Karhumäki, A. Lepistö, L. Zamboni (Eds.), *Combinatorics on Words*, LNCS **8079** (2013), 46–57, Springer.
3. E. Charlier, T. Kamae, S. Puzynina, L. Zamboni, *Self-shuffling infinite words*, in preraration. Preliminary version: *Self-shuffling words,* ICALP 2013, Part II, LNCS **7966** (2013), 113–124, arXiv:1302.3844.
4. J. Hladký, *Random number generators based on the aperiodic infinite words*, https://github.com/jirka-h/aprng
5. A. de Luca, *Sturmian words: structure, combinatorics, and their arithmetics*, Theoret. Comput. Sci. **183** (1997), 45–82.
6. Ch. Doty-Humphrey, *Practically Random: C++ library of statistical tests for RNGs*, https://sourceforge.net/projects/pracrand
7. Ch. Doty-Humphrey, *Practically Random: Specific tests in PractRand*, http://pracrand.sourceforge.net/Tests_engines.txt
8. Ch. Doty-Humphrey, J. Hladký *Practically Random: Discussion of testing results*, http://sourceforge.net/p/pracrand/discussion/366935/thread/a2eaad12
9. X. Droubay, J. Justin, G. Pirillo, *Episturmian words and some constructions by de Luca and Rauzy*, Theoret. Comput. Sci. **255** (2001), 539–553.
10. P. L'Ecuyer. *Random number generation.* In J. E. Gentle, W. Haerdle, and Y. Mori, editors, Handbook of Computational Statistics, 35–71. Springer-Verlag, Berlin, second edition, 2012.
11. P. L'Ecuyer, B. Oreshkin, and R. Simard, *Random numbers for parallel computers: Requirements and methods* (2014) http://www.iro.umontreal.ca/ lecuyer/myftp/papers/parallel-rng-imacs.pdf
12. P. L'Ecuyer, R. Simard, *TestU01: A C library for empirical testing of random number generators*, ACM Trans. Math. Softw. **33(4)** (2007).
13. L.-S. Guimond, Jiří Patera, *Proving the deterministic period breaking of linear congruential generators using two tile quasicrystals*, Math. Comput. **71(237)** (2002), 319–332.
14. L.-S. Guimond, Jan Patera, Jiří Patera, *Combining random number generators using cut-and-project sequences*, Czechoslovak Journal of Physics **51** (2001), 305–311.
15. L.-S. Guimond, Jan Patera, Jiří Patera, *Statistical properties and implementation of aperiodic pseudorandom number generators*, Applied Numerical Mathematics **46(3-4)** (2003), 295–318.
16. M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications 90, Cambridge University Press, 2002.
17. G. Marsaglia, *Random numbers fall mainly in the planes*,Proc. Natl. Acad. Sci. **61 (1)** (1968), 25–28.
18. M. Morse, G. A. Hedlund, *Symbolic dynamics*, Amer. J. Math. **60** (1938), 815–866.
19. M. Morse, G. A. Hedlund, *Symbolic dynamics II: Sturmian trajectories*, Amer. J. Math. **62 (1)** (1940), 1–42.
20. J. Patera, *Generating the Fibonacci chain in* $O(\log n)$ *space and* $O(n)$ *time*, Phys. Part. Nuclei **33** (2002), 118–122.

Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic
E-mail address: lubomira.balkova@gmail.com

Department of Mathematics, University of Turku, FI-20014 Turku, Finland
E-mail address: michelangelo.bucci@utu.fi

DIETI, Università degli Studi di Napoli Federico II, via Claudio, 21, 80125 Napoli, Italy
E-mail address: alessandro.deluca@unina.it

Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic
E-mail address: hladky.jiri@gmail.com

Sobolev Institute of Mathematics, Russia
Current address: Department of Mathematics, University of Turku, FI-20014 Turku, Finland
E-mail address: svepuz@utu.fi