The Sum of Exponents of Maximal Repetitions in Standard Sturmian Words

Marcin Piątkowski

Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland
marcin.piatkowski@mat.umk.pl

Abstract. A maximal repetition is a non-extendable (with the same period) periodic segment in a string, in which the period repeats at least twice. In this paper we study problems related to the structure of maximal repetitions in standard Sturmian words and present the formulas for the sum of their exponents. Moreover, we show how to compute the sum of exponents of maximal repetitions in any standard Sturmian word in linear time with respect to the (total) size of its compressed representation. The presented formulas and algorithm can be easily modified to obtain the total run length of the word.

Keywords: Sturmian words, repetitions, runs, algorithm

1 Introduction

Problems related to repetitions are fundamental in combinatorics on words and many practical applications: data compression, computational biology, pattern-matching and so on, see for instance [6], [7], [10], [11], [14] and references therein. The most important type of repetitions are maximal repetitions, i.e. non-extendable (with the same period) periodic segments in a string, in which the period repeats at least twice. This paper complements the work [2], where the exact formula for the number of runs in standard Sturmian words was presented. We investigate here the structure of runs in standard Sturmian words in more details to obtain a formula for the sum of their exponents. We show also an algorithm, derived from our formula, which computes the sum of exponents of maximal repetitions in any standard word in linear time with respect to the (total) size of its compressed representation (i.e. the directive sequence).

Throughout the paper we use the standard notions of combinatorics on words. In particular, words are finite sequences over a finite set Σ of letters, called the alphabet. For a word $w = w_1 w_2 \cdots w_n$, by w_i we denote its i-th letter, by $w[i..j]$ the subword $w_i w_{i+1} \cdots w_j$, by $|w|$ its length and by $|w|_a$ the number of letters a occurring in w. The number i is a period of the word w if $w_j = w_{i+j}$ for all i with $i + j \leq |w|$. The minimal period of w is denoted by $period(w)$. We say that a word w is periodic if $period(w) \leq \frac{|w|}{2}$. A word w is said to be *primitive* if w is not of the form z^k, where z is a nonempty word and $k \geq 2$ is a natural number.

A *maximal repetition* (a *run*, in short) in a word w is an interval $\alpha = [i..j]$, such that $w[i..j] = u^k v$ $(k \geq 2)$ is a nonempty periodic subword of w, where u is of the minimal length and v is a proper prefix (possibly empty) of u, that can not be extended (neither $w[i-1..j]$ nor $w[i..j+1]$ is a run with the period $|u|$). The factor v is called the remainder of α and the number $k + \frac{|v|}{|u|}$ is called the exponent of α. The

Marcin Piątkowski: The Sum of Exponents of Maximal Repetitions in Standard Sturmian Words, pp. 48–62.
Proceedings of PSC 2013, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-05330-0 © Czech Technical University in Prague, Czech Republic

sum of exponents of all runs in w is denoted by $\sigma(w)$. Note that the exponent of a run is a rational number, hence the value of $\sigma(w)$ does not have to be integer.

A run α can be properly included as an interval in another run β, but in this case $period(\alpha) < period(\beta)$. The value of the run $\alpha = [i..j]$ is the factor $val(\alpha) = w[i..j]$. When it makes no ambiguity we identify sometimes a run with its value and the period of the run $\alpha = [i..j]$ with the subword $w[i..period(w)]$, called also the *generator* of the repetition. The meaning will always be clear from the context. Observe that two different runs could correspond to the identical subwords, if we disregard their positions. Hence, runs are also called the maximal *positioned* repetitions.

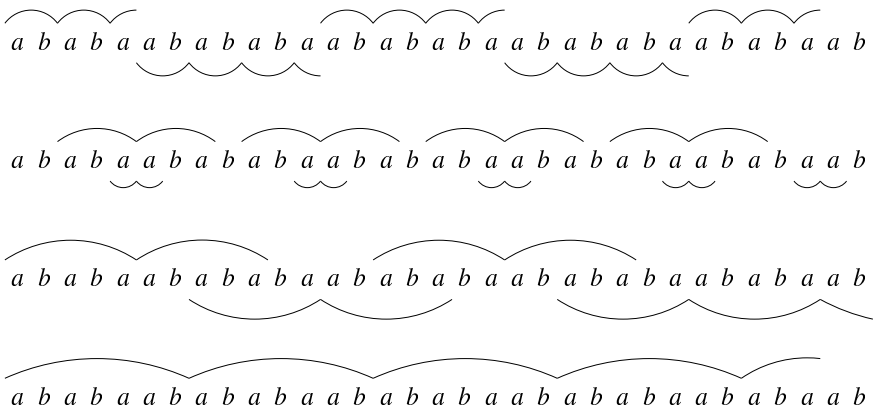

Figure 1. The structure of maximal repetitions for the example binary word.

Example 1.
Let $w = ababaababababaababababaababababaab$ be a binary word.
There are 5 runs with the period a and the exponent 2:

$$w[5..6] = a^2, \quad w[12..13] = a^2, \quad w[19..20] = a^2, \quad w[26..27] = a^2, \quad w[31..32] = a^2,$$

5 runs with the period ab and the exponents respectively $2\frac{1}{2}$ and $3\frac{1}{2}$:

$$w[1..5] = (ab)^2a, \qquad w[6..12] = (ab)^3a, \qquad w[13..19] = (ab)^3a,$$
$$w[20..26] = (ab)^3a, \qquad w[27..31] = (ab)^2a,$$

4 runs with the period aba and the exponent 2:

$$w[3..8] = (aba)^2, \quad w[10..15] = (aba)^2, \quad w[17..22] = (aba)^2, \quad w[24..29] = (aba)^2,$$

4 runs with the period $ababa$ and the exponents respectively 2 and $2\frac{2}{5}$:

$$w[1..10] = (ababa)^2, \qquad w[8..17] = (ababa)^2,$$
$$w[15..24] = (ababa)^2, \qquad w[22..33] = (ababa)^2ab,$$

and 1 run with the period $ababaab$ and the exponent $4\frac{3}{7}$: $w[1..31] = (ababaab)^4aba$. Altogether we have 19 runs and sum of their exponents equals $49\frac{23}{70} \approx 49.3286$, see Figure 1 for comparison.

In 1999 Kolpakov and Kucherov showed that the number of runs in a word is linear with respect to its length (see [13]). The stronger property of runs is that the sum of their exponents is also linear with respect to the length of the word. Kolpakov and Kucherov conjectured that for all w we have $\sigma(w) \leq 2 \cdot |w|$. In 2012 Crochemore with coauthors contradicted this conjecture and showed that the upper bound for $\sigma(w)$ is $2.035 \cdot |w| \leq \sigma(w) \leq 4.1 \cdot |w|$. In this paper we investigate this problem in very special class of strings – the standard Sturmian words. We present compact formulas for the sum of runs exponents and an algorithm for its efficient computation.

Recently a new measure of a string periodicity was proposed by Glen and Simpson (see [12]). The *total run length* (TRL) of a word w is the sum of the lengths of all runs in w. Since this notion is similar to the sum of exponents of maximal repetitions, our formulas and algorithm could be easily adopted to compute also the total run length of any standard Sturmian word.

The paper is organized as follows. In section 2 we introduce the definition of standard Sturmian words and some of their basic properties. Next, in section 3 we study the structure of repetitions in standard Sturmian words and present a few facts necessary in further investigation. Finally, we show an prove the formulas for the sum of exponents of maximal repetitions together with an algorithm for its fast computation. Some useful applets related to problems considered in this paper can be found on the web site:

`http://www.mat.umk.pl/~martinp/stringology/applets/`

2 Standard Sturmian words

Standard Sturmian words (standard words in short) are one of the most investigated class of strings in combinatorics on words, see for instance [1], [3], [4], [5], [15], [17], [18] and references therein. They have very compact representations in terms of sequences of integers, which has many algorithmic consequences.

The *directive sequence* is the integer sequence: $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$, where $\gamma_0 \geq 0$ and $\gamma_i > 0$ for $i = 1, 2, \ldots, n$. The standard word corresponding to γ, denoted by $\mathrm{Sw}(\gamma)$, is described by the recurrences of the form:

$$x_{-1} = b, \qquad x_0 = a, \qquad \ldots, \qquad x_n = (x_{n-1})^{\gamma_{n-1}} x_{n-2}, \qquad x_{n+1} = (x_n)^{\gamma_n} x_{n-1}, \quad (1)$$

where $\mathrm{Sw}(\gamma) = x_{n+1}$. For simplicity we denote $q_i = |x_i|$.

Example 2.
Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have $\mathrm{Sw}(\gamma) = x_5$, where:

$$\begin{aligned}
x_{-1} &= b && & q_{-1} &= 1\\
x_0 &= a && & q_0 &= 1\\
x_1 &= (x_0)^1 \cdot x_{-1} &&= a \cdot b & q_1 &= 2\\
x_2 &= (x_1)^2 \cdot x_0 &&= ab \cdot ab \cdot a & q_2 &= 5\\
x_3 &= (x_2)^1 \cdot x_1 &&= ababa \cdot ab & q_3 &= 7\\
x_4 &= (x_3)^3 \cdot x_2 &&= ababaab \cdot ababaab \cdot ababaab \cdot ababa & q_4 &= 26\\
x_5 &= (x_4)^1 \cdot x_3 &&= ababaababababaababababaababababa \cdot ababaab & q_5 &= 33
\end{aligned}$$

The sequence of words $\{x_i\}_{i=0}^{n+1}$ is called the standard sequence. Every word occurring in a standard sequence is a standard word, and every standard word occurs in some standard sequence. We assume that the standard word given by the empty directive sequence is a and $\mathrm{Sw}(0) = b$.

Observe that for even $n > 0$ the standard word x_n has the suffix ba, and for odd $n > 0$ it has the suffix ab. Moreover, for $\gamma_0 > 0$ we have standard words starting with the letter a and for $\gamma_0 = 0$ we have standard words starting with the letter b. In fact the word $\mathrm{Sw}(0, \gamma_1, \ldots, \gamma_n)$ can be obtained from $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ by switching the letters a and b. Without loss of generality we consider here standard words starting with the letter a, therefore we assume $\gamma_0 > 0$. Words starting with the letter b can be considered similarly.

Remark 3.
The special kind of standard words are well known Fibonacci words. They are formed by repeated concatenation in the same way that the Fibonacci numbers are formed by repeated addition. By definition Fibonacci words are standard words given by directive sequences of the form $\gamma = (1, 1, \ldots, 1)$ (n-th Fibonacci word F_n corresponds to a sequence of n ones).

The number $N = |\mathrm{Sw}(\gamma)|$ is the (real) size of the word, while $(n + 1) = |\gamma|$ can be thought as its compressed size. Observe that, by the definition of standard words, N is exponential with respect to n. Moreover, each directive sequence corresponds to a *grammar-based compression*, which consists in describing a given word by a context-free grammar G generating this (single) word. The size of the grammar G is the total length of all productions of G. In our case the size of the considered grammar is proportional to the length of the directive sequence.

2.1 Morphic reduction of standard words

The recurrent definition of standard words from equation (1) leads to their simple characterization by a composition of morphisms. Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a directive sequence. We associate with γ a sequence of morphisms $\{h_i\}_{i=0}^{n}$, defined as:

$$h_i : \begin{cases} a \longrightarrow a^{\gamma_i} b \\ b \longrightarrow a \end{cases} \qquad \text{for } 0 \leq i \leq n. \tag{2}$$

The following fact describes another simple method of standard word generation. It can be proven by a simple induction, see [2] for more details.

Lemma 4 (see [2]).
For $0 \leq i \leq n$ the morphism h_i transforms a standard word into another standard word, and we have:

$$\begin{aligned} \mathrm{Sw}(\gamma_n) &= h_n(a), \\ \mathrm{Sw}(\gamma_i, \gamma_{i+1}, \ldots, \gamma_n) &= h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \gamma_{i+2}, \ldots, \gamma_n)\Big). \end{aligned}$$

As a direct corollary to Lemma 4 we have that for $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$:

$$\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n) = h_0 \circ h_1 \circ \cdots \circ h_n(a). \tag{3}$$

Moreover, the inverse morphism h_i^{-1} can be seen as a reduction of a standard word $w^{(i)} = \mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$ to $w^{(i+1)} = \mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n)$.

Recall that $|w|_a$ denotes the number of occurrences of the letter a in the word w. In the rest of this paper, for $\gamma = (\gamma_0, \ldots, \gamma_n)$ and $0 \leq k \leq n$, we use the following notation:

$$N_\gamma(k) \;=\; |\mathrm{Sw}(\gamma_k, \gamma_{k+1}, \ldots, \gamma_n)|_a, \tag{4}$$

which enables us to simplify the formulas for the sum of runs exponents. Observe that equations (2) and (4) imply:

$$N_\gamma(k) \;=\; \gamma_k \cdot N_\gamma(k+1) + N_\gamma(k+2). \tag{5}$$

Example 5.
Consider a directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have (compare with Example 2):

$$\begin{aligned}
\mathrm{Sw}(1,2,1,3,1) &= ababaababababaababababaababababaab & N_\gamma(0) &= 19,\\
\mathrm{Sw}(2,1,3,1) &= aabaaabaaabaaabaaba & N_\gamma(1) &= 14,\\
\mathrm{Sw}(1,3,1) &= abababaab & N_\gamma(2) &= 5,\\
\mathrm{Sw}(3,1) &= aaaba & N_\gamma(3) &= 4,\\
\mathrm{Sw}(1) &= ab & N_\gamma(4) &= 1,\\
\mathrm{Sw}(\varepsilon) &= a & N_\gamma(5) &= 1.
\end{aligned}$$

As a straightforward corollary to equations (2), (4) and (5) we have:

Corollary 6.
The number of letters b in a word $\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$ equals $N_\gamma(i+1)$.

2.2 The m-partition of a standard word

The concept of the m-partition of a standard word is crucial in the maximal repetitions structure investigation. It allows us to divide the set of all runs in a standard word to disjoint sets depending on the length of their periods and simplify the considered problems. The following fact is a direct consequence of the recurrent definition of standard words.

Proposition 7.
Every standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ can be represented as a sequence of concatenated words x_m and x_{m-1}, and has the form:

$$(i)\;\; x_m^{\alpha_1}\, x_{m-1}\, x_m^{\alpha_2}\, x_{m-1} \cdots x_m^{\alpha_s}\, x_{m-1}\, x_m \qquad \text{or} \qquad (ii)\;\; x_m^{\beta_1}\, x_{m-1}\, x_m^{\beta_2}\, x_{m-1} \cdots x_m^{\beta_s}\, x_{m-1},$$

where $\alpha_k,\ \beta_k \in \{\gamma_m,\ \gamma_m + 1\}$, $0 \leq m \leq n$, and x_m are as in equation (1).

Such a decomposition of a standard word w is called the *m-partition* of w. The block x_m is called the *repeating block* and x_{m-1} – the *single block*. Recall that for $m > 0$ the last two letters of x_m are ab for an odd m and ba for an even m. Therefore the m-partition of $x_{n+1} = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is of the form (i) if m has the same parity as $(n+1)$, and of the form (ii) otherwise (see Example 9 and Figure 2).

Note that the 0-partition of a standard word is its decomposition into letters. Moreover, Proposition 7, Lemma 4 and equation (3) imply the following fact.

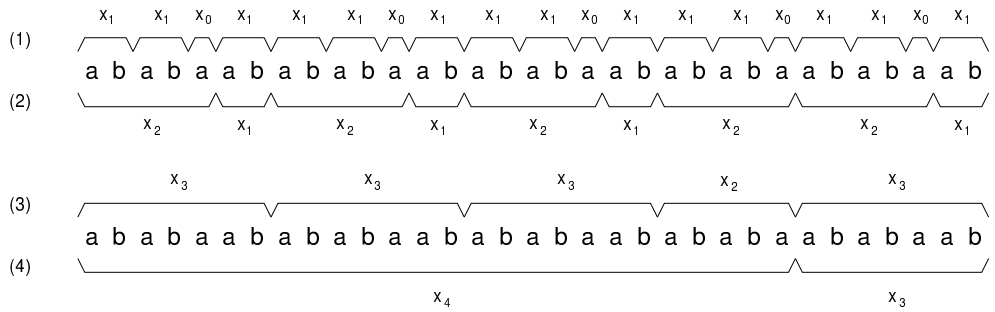

Figure 2. The m-partition of the word $\mathrm{Sw}(1,2,1,3,1)$ for $1 \leq m \leq 4$.

Proposition 8.
The structure of occurrences of the block x_m (respectively x_{m-1}) in the m-partition of $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the structure of occurrences of the letter a (respectively letter b) in $\mathrm{Sw}(\gamma_m, \ldots, \gamma_n)$.

Example 9.
Consider a standard word $\mathrm{Sw}(1,2,1,3,1)$. Its m-partitions (for $1 \leq m \leq 4$) together with its corresponding morphic reductions are depicted in the table below. See also Figure 2 for comparison.

m	m-partition	$\mathrm{Sw}(\gamma_m, \ldots, \gamma_n)$
1	$ab \cdot ab \cdot a \cdot ab \cdot ab \cdot ab \cdot a \cdot ab \cdot ab \cdot ab \cdot a \cdot ab \cdot ab \cdot ab \cdot a \cdot ab \cdot ab \cdot a \cdot ab$	$aabaaabaaabaaabaaba$
2	$ababa \cdot ab \cdot ababa \cdot ab \cdot ababa \cdot ab \cdot ababa \cdot ababa \cdot ab$	$abababaab$
3	$ababaab \cdot ababaab \cdot ababaab \cdot ababa \cdot ababaab$	$aaaba$
4	$ababaababababaababababa \cdot ababaab$	ab

3 The structure of maximal repetitions in standard words

The aim of this section is the presentation of some technical facts used further to prove the correctness of formulas for the sum of runs exponents. We start with recalling some technical facts presented in [8] and [9] related to the structure of factors in standard words.

Lemma 10 (see [9]).
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence. For every $0 \leq k \leq n$ and every $1 \leq i \leq \gamma_k$ the word $(x_k)^i x_{k-1}$ is primitive (i.e. is not of the form z^s, where z is nonempty and $s \geq 2$ is a natural number).

Lemma 11 (See [8]).
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and let $y \in \{a, b\}$ be a letter. For each occurrence of $y \cdot x_i$ in w, y is the last letter of the block x_{i-1} or x_i of the i-partition of w. Moreover, the type of this block is uniquely determined by y.

The following lemma is a key tool in the study of the runs structure in standard words. It is a version of Theorem 1 in [9] using a slightly different notation.

Lemma 12 (Structural Lemma).
The period of each maximal repetition in a standard word $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ is of the form x_i or $(x_i)^j x_{i-1}$, where $0 \leq i \leq n$, $0 < j < \gamma_i$ and x_i's are as in equation (1).

To prove the above lemma it is sufficient to show that no factor of a standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ that does not satisfy the condition given there could be the generator of some repetition, see the proof of Theorem 1 in [9] for more details.

Let us denote by $\widehat{w}$ the word w with two last letters removed and by $\widetilde{w}$ the word w with two last letters exchanged. The following fact can be proven by a simple induction, see for instance [15].

Lemma 13.
Let x_i be as in equation (1) and $i > 1$. Then:

1. *We have $x_{i-1} \cdot x_i \;=\; x_i \cdot \widetilde{x_{i-1}}$,*
2. *The longest prefix of $x_{i-1} \cdot x_i$ with the period of the length q_i is of the form $x_i \cdot \widehat{x_{i-1}}$.*

Example 14.
Recall the word $\mathrm{Sw}(1, 2, 1, 3, 1)$ from Example 2, where $x_3 = ababaab$, $x_2 = ababa$. Then we have $\widetilde{x_2} = abaab$, $\widehat{x_2} = aba$ and

$$x_2 \cdot x_3 \;=\; ababa \cdot ababaab \;=\; ababaab \cdot abaab \;=\; x_3 \cdot \widetilde{x_2}.$$

Moreover, the longest prefix of $x_2 \cdot x_3$ with the period of the length q_2 is of the form:

$$\overbrace{\underbrace{a\,b\,a\,b\,a}_{x_3 \cdot \widehat{x_2}}\overbrace{a\,b\,a\,b\,a}^{x_3}}^{x_2}\,b\,a$$

Observe that by equation (1) we have

$$\mathrm{Sw}(\gamma_0, \ldots, \gamma_n, 1) = (x_n)^{\gamma_n} \cdot x_{n-1} \cdot x_n \quad \text{and} \quad \mathrm{Sw}(\gamma_0, \ldots, \gamma_n + 1) = (x_n)^{\gamma_n} \cdot x_n \cdot x_{n-1}.$$

Therefore, as a straightforward corollary to the first point of Lemma 13 we get:

Corollary 15.
Standard words $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n, 1)$ and $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n + 1)$ differ only in the order of the last two letters.

See Figure 3 for an illustration of this fact. To properly count the exponents of runs in standard words we need also the following fact.

Proposition 16.
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $2 \leq i \leq n-2$. If x_{i-1} is the last block of the i-partition of w, then it is preceded by $(x_i)^{\gamma_i+1}$.

Proof.
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $2 \leq i \leq n-2$. By equation (1) we have $x_i = \mathrm{Sw}(\gamma_0, \ldots, \gamma_i)$ and $x_{i-1} = \mathrm{Sw}(\gamma_0, \ldots, \gamma_{i-1})$. Recall that x_i ends with ba for even $i > 0$ (i.e. for the odd length of a directive sequence) and with ab for odd $i > 0$ (i.e. for the even length of a directive sequence). Consider that w has the suffix $(x_i)^\alpha x_{i-1}$. Then n and i have the same parity and the number $n - m + 1$ is odd, hence the word $w^{(m)} = \mathrm{Sw}(\gamma_m, \ldots, \gamma_n)$ ends with ab. More precisely, due to Proposition 8, $w^{(m)}$ ends with $a^\alpha b$. By Lemma 4, the suffix $a^{\gamma_i} b$ of $w^{(m)}$ corresponds to the last letter a of $w^{(m+1)} = \mathrm{Sw}(\gamma_{m+1}, \ldots, \gamma_n)$. Since $n - m + 2$ is even and $w^{(m+1)}$ ends with ba, due to Lemma 4 the suffix $a^{\gamma_i} b$ of $w^{(m)}$ have to be preceded by a single occurrence of a. Therefore, we have $\alpha = \gamma_i + 1$ and this completes the proof. □

4 The sum of exponents of maximal repetitions

In this section we present and prove formulas for the the sum of exponents of maximal repetitions in any standard word, that depend only on its compressed representation – the directive sequence. The following zero-one functions for testing the parity of a nonnegative integer i will be useful to simplify those formulas:

$$even(i) \;=\; \begin{cases} 1 & \text{for even } i \\ 0 & \text{for odd } i \end{cases} \qquad \text{and} \qquad odd(i) \;=\; \begin{cases} 1 & \text{for odd } i \\ 0 & \text{for even } i \end{cases}.$$

Moreover, we define an auxilary function $\Delta_n : \mathbf{N} \to \mathbf{N}$:

$$\Delta_n(i) \;=\; |n - i + 1| \bmod 2.$$

In other words, $\Delta_n(i) = 1$ if and only if the numbers n and i have the same parity, and $\Delta_n(i) = 0$ otherwise. Recall also that for simplicity we denote $|x_i| = q_i$.

The main idea of the computation of the sum of runs exponents in a standard word w is the partition of the set of all maximal repetitions in w into separate categories depending on the length of their periods. Runs in w with the period of the form x_i and $(x_i)^k x_{i-1}$ (for $1 < k < \gamma_i$), where x_i are as in equation (1), are called the runs of *type* i. We study runs of each type separately.

Let $\sigma_i(\gamma)$ denotes the sum of exponents of type i runs. Then the sum of exponents of all runs in $\mathrm{Sw}(\gamma)$ can be computed using the following theorem.

Theorem 17.
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence. The sum of exponents of runs in $\mathrm{Sw}(\gamma)$ is given as:

$$\sigma(\gamma) \;=\; \sum_{i=1}^{n} \sigma_i(\gamma).$$

The detailed computation of $\sigma_i(\gamma)$ for each $0 \le i \le n$ is provided below.

4.1 The general case

We start with an investigation of a general case, i.e. maximal repetitions of the type i for $2 \le i \le n-1$. First, we consider runs with the period of the form x_i.

Lemma 18.
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. For $2 \le i \le n-1$ the sum of exponents of runs with the period x_i in w equals:

$$\sigma'_i(\gamma) \;=\; N_\gamma(i+1) \cdot \left(\gamma_i + 1 + \frac{q_{i-1} - 2}{q_i}\right) + \Big(N_\gamma(i+2) - 1\Big) + \Delta_n(i)\frac{2}{q_i}. \qquad (6)$$

Proof.
Let us denote

$$w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n), \quad w^{(i)} = \mathrm{Sw}(\gamma_i, \ldots, \gamma_n) \quad \text{and} \quad w^{(i+1)} = \mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n).$$

Due to Lemma 11, each maximal repetition with the period x_i in w is aligned to the i-partition of w, hence it corresponds to a block $(x_i)^\alpha x_{i-1}$, where $\alpha \in \{\gamma_i, \gamma_i + 1\}$.

Each internal block of this form is followed by a subsequent x_i. Due to Lemma 13, the longest prefix of $x_{i-1}x_i$ with the period x_i equals $x_i \cdot \widehat{x_{i-1}}$. Therefore, the period of the considered run repeats $\alpha+1$ times and its fractional part has the length $q_{i-1}-2$.

Consider the i-partition of w. By Proposition 8 occurrences of x_i correspond to occurrences of a in $w^{(i)}$ and occurrences of x_{i-1} correspond to occurrences of b in $w^{(i)}$. Therefore, a block $(x_i)^\alpha x_{i-1}$ correspond to the block $a^\alpha b$ in $w^{(i)}$. Moreover, due to Lemma 4, each block of the form $a^{\gamma_i+1}b$ in $w^{(i)}$ corresponds to the letter a preceded by the letter b in $w^{(i+1)}$ and each block of the form $a^{\gamma_i}b$ in $w^{(i)}$ corresponds to the letter a not preceded by the letter b in $w^{(i+1)}$.

The rightmost occurrence of $(x_i)^\alpha x_{i-1}$ have to be considered separately. Due to Proposition 7, if i and n have different parity the i-partition of w ends with $(x_i)^\alpha x_{i-1}x_i$. In this case the period of the considered repeats $\alpha+1$ times and its fractional part has the length $q_{i-1}-2$. On the other hand, if i and n have the same parity, the i-partition of w ends with $(x_i)^\alpha x_{i-1}$. Due to Proposition 16, $\alpha = \gamma_i + 1$. Moreover, since x_{i-1} is a prefix of x_i, the fractional part of considered run consists of the whole word x_i and has the length q_{i-1}.

Summing up, in the computation of the sum of runs exponents, we count $\gamma_i + 1 + \frac{q_{i-1}-2}{q_i}$ for each occurrence of a in $w^{(i+1)}$, namely $N_\gamma(i+1)$ times, and an additional 1 for each b in $w^{(i+1)}$ (except the rightmost one), namely $N_\gamma(i+2)-1$ times. Finally, we must take care of the remainder of the rightmost run with period x_i and we obtain the statement of the lemma. See Figure 3 for the illustration of type-2 runs structure in example words and two possible remainders of the rightmost run. □

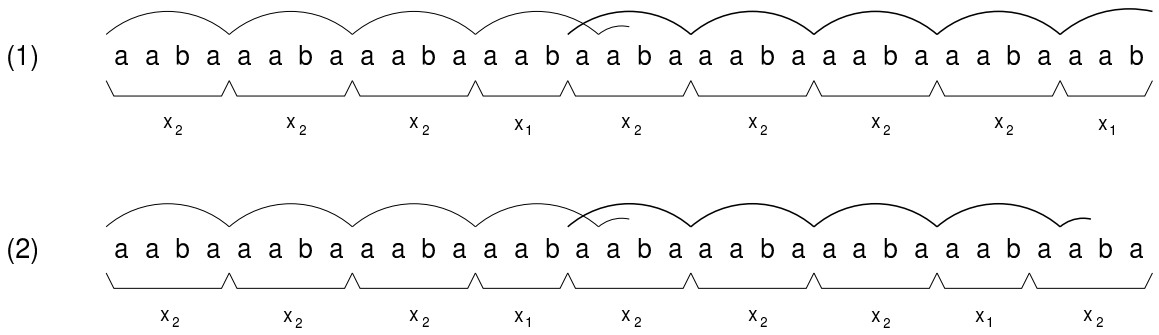

Figure 3. The structure of runs with the period x_2 in a standard word $\mathrm{Sw}(2,1,3,1,1)$ (1) compared to $\mathrm{Sw}(2,1,3,2)$ (2).

Observe that the maximal repetitions with the period of the form $(x_i)^k x_{i-1}$, where $1 \le k < \gamma_i$, appear only for $\gamma_i > 1$. The sum of exponents of such runs is given by the following fact.

Lemma 19.
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. For $1 \le i \le n-1$ the sum of exponents of runs in w with the period $(x_i)^k x_{i-1}$, where $1 \le k < \gamma_i$, equals:

$$\sigma_i''(\gamma) \;=\; \Big(N_\gamma(i+1)-1\Big) \cdot \sum_{k=1}^{\gamma_i - 1} \left(2 + \frac{q_i - 2}{k \cdot q_i + q_{i-1}}\right). \tag{7}$$

Proof.
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ and $u = (x_i)^k x_{i-1}$, where $1 \leq k < \gamma_i$. Due to Lemma 11, each occurrence of u is aligned to the i-partition of w. Consider a repetition of the form u^m in w and denote it as $u^{(1)}u^{(2)}\cdots u^{(m)}$. Observe that each $u^{(2)}$, …, $u^{(m)}$ have to be preceded by the suffix of u, namely x_{i-1}. Since each two consecutive occurrences of x_{i-1} in the i-partition of w are separated by at least γ_i occurrences of x_i and $k < \gamma_i$, the factor u cannot have more than two consecutive occurrences. Therefore, the considered run with the period u has the form $u^{(1)}u^{(2)} \cdot v$, where v is a prefix of u.

The suffix x_{i-1} of $u^{(2)}$ starts at the beginning of an x_i block followed by x_{i-1}, which appears either as block of the i-partition of w or as a prefix of a subsequent block x_i. Due to Lemma 13, the considered factor has the form $x_i \cdot x_{i-1} = x_{i-1} \cdot \widetilde{x_i}$. Therefore, the fractional part of the considered run has the length $q_i - 2$.

Observe, that occurrences of $u^{(1)}$ in w are aligned with occurrences of x_{i-1} in the i-partition of w. Therefore, each such occurrence of x_{i-1} (except the rightmost one) corresponds to $\gamma_i - 1$ runs with a period $(x_i)^k x_{i-1}$, for $1 \leq k < \gamma_i$. Due to Proposition 8, each occurrence of x_{i-1} in the i-partition of w corresponds to an occurrence of b in $\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$. Summing up exponents of all $\gamma_i - 1$ runs for each b in $\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)$ (except the rightmost one), namely $N_\gamma(i+1) - 1$ occurrences, we obtain the statement of the lemma. See Figure 4 for an illustration of the structure of runs of this type. □

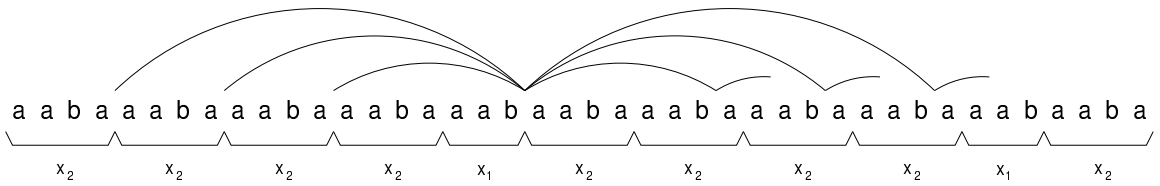

Figure 4. The structure of runs with the period $(x_2)^k x_1$ $(1 \leq k \leq 3)$ in $\mathrm{Sw}(2, 1, 4, 2)$.

The complete formula for the sum of exponents of all type-i runs can be obtained by combining the formulas from Lemma 18 and Lemma 19.

Lemma 20.
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. For $2 \leq i \leq n-1$ the sum of exponents of type i runs in w equals:

$$\begin{aligned}\sigma_i(\gamma) = N_\gamma(i+1) \cdot \left(\gamma_i + 1 + \frac{q_{i-1} - 2}{q_i}\right) + \Big(N_\gamma(i+2) - 1\Big) + \Delta_n(i)\frac{2}{q_i} \\ + \Big(N_\gamma(i+1) - 1\Big) \cdot \sum_{k=1}^{\gamma_i - 1} \left(2 + \frac{q_i}{k \cdot q_i + q_{i-1}}\right).\end{aligned} \tag{8}$$

4.2 Boundary cases

For a standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ runs of types 0, 1 and n have to be investigated differently. We start with the analyze of runs of type 0, i.e. the runs with the period of the form a.

Lemma 21 (Type 0).
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard

word. The sum of exponents of type 0 runs in w equals:

$$\sigma_0(\gamma) = \begin{cases} 2\Big(N_\gamma(2) - odd(n)\Big) & \text{for } \gamma_0 = 1 \\ \gamma_0 N_\gamma(1) + N_\gamma(2) - odd(n) & \text{for } \gamma_0 > 1 \end{cases}. \tag{9}$$

Proof.
Each standard word consists of blocks of repeated occurrences of the letter a separated by single occurrences of the letter b. The length of the blocks of the form $a \cdots a$ depends on the value of γ_0.

First assume that $\gamma_0 = 1$. In this case the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ consists of the blocks of two types: ab or aab and only the blocks of the second type include the runs with the period a and exponent 2. Due to Lemma 4, every such run in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the letter b followed by the letter a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Hence, the number of such runs equals the number of blocks ba in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$.

Recall that for an even length of the directive sequence $|(\gamma_1, \ldots, \gamma_n)|$ (n is even) the word $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ ends with ba and in this case the number of runs with the period a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ equals the number of the letters b in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$, namely $N_\gamma(2)$. On the other hand, for an odd length of the directive sequence $|(\gamma_1, \ldots, \gamma_n)|$ (n is odd) the word $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ ends with ab and the last letter b does not correspond to a run in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$. In this case, the number of runs with the period a in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is one less than the number of the letters b in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$, namely $N_\gamma(2) - 1$. Hence, in this case the sum of type-0 runs exponents equals

$$\sigma_0(\gamma) = 2\Big(N_\gamma(2) - odd(n)\Big).$$

Assume now that $\gamma_0 > 1$. Every run with the period a in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ equals a^{γ_0} or a^{γ_0+1} and is followed by the single letter b. Due to Lemma 4, every such run in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the letter a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Hence in this case we have $N_\gamma(1)$ runs with the period a.

By Lemma 4 each occurrence of a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ preceded by b produces a run a^{γ_0+1} in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$, and each occurrence of a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ not preceded by b produces a run a^{γ_0} in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$. Therefore, in computation of the sum of runs exponents, we count γ_0 for each a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ and an additional 1 for each b. As in the previous case, for odd n, the rightmost b does not correspond to a run in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Therefore, in this case the sum of type-0 runs exponents equals

$$\sigma_0(\gamma) = \gamma_0 N_\gamma(1) + N_\gamma(2) - odd(n).$$

□

The next boundary case, strongly related to the case considered above, is the sum of exponents of runs with the period of the form x_1.

Lemma 22 (Type 1).
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. The sum of exponents of runs with the period x_1 in w equals:

$$\sigma_1'(\gamma) = \begin{cases} \Big(N_\gamma(3) - 1\Big) \cdot \left(2 + \dfrac{\gamma_0}{\gamma_0 + 1}\right) + odd(n) \cdot \left(2 + \dfrac{1}{\gamma_0 + 1}\right) & \text{for } \gamma_1 = 1 \\ N_\gamma(2) \cdot \left(\gamma_1 + \dfrac{\gamma_0}{\gamma_0 + 1}\right) + \Big(N_\gamma(3) - 1\Big) + odd(n) \cdot \dfrac{\gamma_0 - 1}{\gamma_0 + 1} & \text{for } \gamma_1 > 1 \end{cases}. \tag{10}$$

Proof.
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$. By definition we have $x_1 = a^{\gamma_0}b$. Therefore, the remainder of each internal run with the period x_1 has the length γ_0.

Consider the 1-partition of w. By Lemma 4 occurrences of blocks of the form $a^{\gamma_0}b$ correspond to occurrences of letters a in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ and occurrences of blocks of the form a to occurrences of letters b in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Therefore, following similar argumentation as in proof of Lemma 21, we obtain the formula for the sum of exponents of internal runs with the period x_1 in w.

Let us now consider the rightmost run with the period x_1 in w. If n is even, w ends with $a \cdot a^{\gamma_0}b$ and this occurrence of x_1 does not correspond to a run in w. On the other hand, if n is odd, due to Proposition 16 w ends with $(a^{\gamma_0}b)^{\gamma_1+1}a$. Such a suffix corresponds to a run with the total part of exponent equal γ_1+1 and the remainder a, and we should include it in our formula. □

The sum of exponents of runs with the period $(x_1)^k x_0$ for $1 \leq k < \gamma_1$ follows from Lemma 19. As a final step of investigation we count the sum of exponents of type-n runs.

Lemma 23 (Type n).
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. The sum of exponents of runs of type n in w is given by the formula:

$$\sigma_n(\gamma) \;=\; \begin{cases} 0 & \text{for} \quad \gamma_n = 1 \\ \gamma_n + \dfrac{q_{n-1}}{q_n} & \text{for} \quad \gamma_n > 1 \end{cases} \tag{11}$$

Proof.
We have $w = (x_n)^{\gamma_n} x_{n-1}$. Therefore, for $\gamma_n = 1$ there is no run of type n in w. On the other hand, for $\gamma_n > 1$, w contains only one run of type n. Its generator – x_n – repeats undivided γ_n times. Moreover, since x_{n-1} is a prefix of x_n, the total exponent of α equals $\gamma_n + \frac{q_{n-1}}{q_n}$. □

Now we can combine the formulas from equations (6), (7), (9), (10) and (11) and obtain the formula from Theorem 17.

4.3 Algorithm

The formulas from equations (6), (7), (9), (10) and (11) lead to simple and efficient algorithm for computation of the sum of runs exponents in any standard word. Its time complexity depends only on the coefficients of the directive sequence, which is the compressed representation of a considered word.

Theorem 24.
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma)$ be a standard word. The sum of exponents of maximal repetitions in w can be computed in time $O(\|\gamma\|)$, where $\|\gamma\| = \gamma_0 + \gamma_1 + \cdots + \gamma_n$.

Proof.
Observe that, by equations (6), (7), (9), (10) and (11), the value of each formula $\sigma_i(\gamma)$ depends only on coefficients of γ and the values of $N_\gamma(i+1)$, $N_\gamma(i+2)$, q_i and q_{i-1}.

Therefore, we can iterate through all types of runs from 0 to n computing the value of $\sigma_i(\gamma)$ and simultaneously updating the values of $N_\gamma(i+1)$, $N_\gamma(i+2)$, q_i and q_{i-1}. See Algorithm 1 for details.

The main loop of presented algorithm (lines 8-13) performs $n+1$ iterations. The most time consuming part of each iteration is the computation of the sum of exponents of maximal repetitions with the period $(x_i)^k x_{i-1}$ (line 10), namely the component

$$\sum_{k=1}^{\gamma_i - 1} \left(2 + \frac{q_i - 2}{k \cdot q_i + q_{i-1}}\right).$$

It can be done in $O(\gamma_i)$ time. Hence, the time complexity of the whole algorithm is $O(\|\gamma\|)$, where $\|\gamma\| = \gamma_0 + \gamma_1 + \cdots + \gamma_n$. □

Algorithm 1: *Sum-Of-Exponents*$\big(\mathrm{Sw}(\gamma)\big)$

Input: $\gamma = (\gamma_0, \ldots, \gamma_n)$
Output: $\sigma(\gamma)$

```
1  result ⟵ 0;
2  N_γ(n+1) ⟵ 1;
3  N_γ(n+2) ⟵ 0;
4  q_0 ⟵ 1;
5  q_{-1} ⟵ 1;
6  for i := 1 to n do
7      (q_{i+1}, q_i) ⟵ (γ_i · q_i, q_i);
8  for i := n downto 0 do
9      compute σ'_i(γ);    // runs with period x_i;
10     compute σ''_i(γ);   // runs with period (x_i)^k x_{i-1};
11     result ⟵ result+σ'_i(γ)+σ''_i(γ);
12     (q_i, q_{i-1}) ⟵ (q_{i-1}, q_i − γ_{i-1} · q_{i-1});
13     (N_γ(i), N_γ(i+1)) ⟵ (γ_i · N_γ(i) + N_γ(i+1), N_γ(i));
14 return result;
```

Final remarks

The aim of this paper was to study problems related to repetitions in standard Sturmian words – one of the most thoroughly investigated class of strings in combinatorics of words. We presented the formulas for the sum of exponents of maximal repetitions in any standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ that depend only on its compressed representation (the directive sequence). We proposed also an algorithm based on those formulas that computes the sum of runs exponents in any standard word in linear time with respect to the (total) size of the directive sequence, i.e. in time $O(\|\gamma\|)$, where $\|\gamma\| = \gamma_0 + \gamma_1 + \cdots + \gamma_n$.

The notion of total run length (TRL) proposed in [12] can be considered similarly. To obtain the formulas for the total run length of a standard word we can use modified formulas for the sum of runs exponents. We only needed to multiply the total part of each exponent by the length of related period (either q_i or $k \cdot q_i + q_{i-1}$) and remove

the denominator from its fractional part. The described change could be also taken into account in the presented algorithm.

In the case of the total run length computation, the component

$$\sum_{k=1}^{\gamma_i-1}\left(2+\frac{q_i-2}{k\cdot q_i+q_{i-1}}\right)$$

of equation 7 has the form

$$\sum_{k=1}^{\gamma_i-1}\Big((k+1)q_i+q_{i-1}-2\Big).$$

The above formula is a sum of an arithmetic progression, hence it can be simplified as

$$(\gamma_i-1)\frac{(\gamma_i+2)q_i+2q_i-4}{2}.$$

Therefore, in each iteration of the main loop of the modified algorithm, we have to compute the value of a single arithmetic formula and update the values of $N_\gamma(i+1)$, $N_\gamma(i+2)$, q_i and q_{i-1}. This way we obtain the algorithm computing the total run length of any standard word $\mathrm{Sw}(\gamma)$ in time $O(|\gamma|)$, where $|\gamma|$ denotes the length of the directive sequence.

References

1. J. Allouche and J. Shallit: *Automatic Sequences. Theory, Applications, Generalizations.*, Cambridge University Press, 2003.
2. P. Baturo, M. Piątkowski, and W. Rytter: *The Number of Runs in Standard Sturmian Words.* Electronic Journal of Combinatorics, 20(1) 2013.
3. P. Baturo and W. Rytter: *Compressed string-matching in standard Sturmian words.* Theoretical Computer Science, 410(30–32) 2009, pp. 2804–2810.
4. J. Berstel: *Sturmian and Episturmian words: a survey of some recent results*, in Proceedings of the 2nd international conference on Algebraic informatics, vol. 4728 of Lecture Notes in Computer Science, Springer, 2007, pp. 23–47.
5. J. Berstel, A. Lauve, C. Reutenauer, and F. Saliola: *Combinatorics on Words: Christoffel Words and Repetitions in Words*, CRM monograph series, Providence, R.I: American Mathematical Society, 2009.
6. M. Crochemore and L. Ilie: *Analysis of maximal repetitions in strings*, in Proceedings of the 32nd International Conference on Mathematical Foundations of Computer Science, vol. 4708 of Lecture Notes in Computer Science, Springer, 2007, pp. 465–476.
7. M. Crochemore, L. Ilie, and L. Tinta: *Towards a solution of the "runs" conjecture*, in Proceedings of the 19th annual symposium on Combinatorial Pattern Matching, vol. 5029 of Lecture Notes in Computer Science, Springer, 2008, pp. 290–302.
8. D. Damanik and D. Lenz: *The index of sturmian sequences.* European Journal of Combinatorics, 23(1) 2002, pp. 23–29.
9. D. Damanik and D. Lenz: *Powers in Sturmian sequences.* European Journal of Combinatorics, 24(4) 2003, pp. 377–390.
10. F. Franek, R. J. Simpson, and W. F. Smyth: *The maximum number of runs in a string*, in Proceedings of 14th Australian Workshop on Combinatiorial Algorithms, 2003, pp. 26–35.
11. F. Franek and Q. Yang: *An asymptotic lower bound for the maximal number of runs in a string.* Jnternational Journal of Foundations of Computer Science, 19(1) 2008, pp. 195–203.
12. A. Glen and J. Simpson: *The total run length of a word.* arXiv:1301.6568, 2013.
13. R. Kolpakov and G. Kucherov: *On the sum of exponents of maximal repetitions in a word*, Tech. Rep. 99-R-034, LORIA, 1999.

14. K. Kusano, W. Matsubara, A. Ishino, H. Bannai, and A. Shinobara: *New lower bound for the maximum number of runs in a string.* Computing Research Repository, abs/0804.1214 2008.
15. M. Lothaire: *Algebraic Combinatorics on Words*, vol. 90 of Encyclopedia of mathematics and its application, Cambridge University Press, 2002.
16. M. Piątkowski: *Stringology applets* . `http://www.mat.umk.pl/~martinp/stringology/applets`.
17. M. Sciortino and L. Zamboni: *Suffix automata and standard Sturmian words*, in Proceedings of the 11th International Conference on Developments in Language Theory, vol. 4588 of Lecture Notes in Computer Science, Springer, 2007, pp. 382–398.
18. J. Shallit: *Characteristic words as fixed points of homomorphisms*, Tech. Rep. CS-91-72, University of Waterloo, Department of Computer Science, 1991.