Warsaw University

Faculty of Mathematics, Informatics and Mechanics

Marcin Piątkowski

Faculty of Mathematics and Computer Science

Nicolaus Copernicus University

Efficient algorithms related to combinatorial structure of words

PhD dissertation

Supervisor
prof. dr hab. Wojciech Rytter
Institute of Informatics
Warsaw University

March 2011

Author's declaration:
Aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

....
date

....
Marcin Piątkowski

Supervisor's declaration:
This dissertation is ready to be reviewed.

....
date

....
prof. dr hab. Wojciech Rytter

Abstract [*]

Problems related to repetitions are central in the area of combinatorial algorithms on strings. The main types of repetitions are squares (strings of the form zz) and runs (also called maximal repetitions).

Denote by $sq(w)$ the number of distinct squares and by $\rho(w)$ the number of runs in a given string w; denote also by $sq(n)$ and $\rho(n)$ the maximal number of distinct squares and the maximal number of runs respectively in a string of the size n (we slightly abuse the notation by using the same names, but the meaning will be clear from the context). Despite a long research in this area the exact asymptotics of $sq(n)$ and $\rho(n)$ are still unknown. Also the algorithms for efficient calculation of $sq(w)$ and $\rho(w)$ are very sophisticated for general class of words.

In the thesis we investigate these problems for a very special class $\mathcal{S}$ of strings called the standard Sturmian words – one of the most investigated class of strings in combinatorics on words. They have very compact representations in terms of sequences of integers. For a sequence γ denote by $\mathrm{Sw}(\gamma)$ the standard word generated by γ. Usually the size of this word is exponential with respect to the size of γ, hence we are dealing here with repetition problems in compressed strings, which complicates the algorithms and proofs considerably. Standard words are the special kind of cyclic shifts of Chrisfoffel words, which constitute another important family of words with geometric applications.

Our main results are:

- the algorithm to compute $\rho(w)$ for $w \in \mathcal{S}$ in linear time with respect to the size of its (usually logarithmic) compact representation γ;
- the tight asymptotic bound: $\rho(w) \leq 0.8|w|$ for $w \in \mathcal{S}$;
- the algorithmic construction of an infinite sequence of standard words $\{w_k\}$ achieving asymptotically $\rho(w_k) = 0.8|w_k| - o(|w_k|)$;
- the tight asymptotic bound $sq(|w|) \leq 0.9|w|$ for $w \in \mathcal{S}$ ($\mathcal{S}$ is the largest class of words, for which we know the exact coefficient in asymptotic formula, for general words the best result is $n \leq sq(n) \leq 2n$ and any progress is very hard);
- the algorithmic construction of an infinite sequence of standard words $\{w_k\}$ achieving asymptotically $sq(w_k) = 0.8|w_k| - o(|w_k|)$;
- the investigation of the structure of compacted subword graphs of standard words – these graphs have linear size with respect to $|\gamma|$ and using them we derive several efficient algorithms for computing some important values related to $\mathrm{Sw}(\gamma)$;
- the relation of subword graphs of standard words to certain numerations systems and a special type of finite automata.

The results of the thesis were presented in [5], [6] and [61].

Keywords: combinatorics of words, Sturmian words, maximal repetitions, squares, data compression, subword graphs, numeration systems

AMS Mathematical Subject Classification: 68R15, 68P30, 68Q70.

[*] The research supported by Ministry of Science and Higher Education of Poland, grant N N206 258035

Streszczenie [*]

Badanie struktury powtórzeń jest jednym z podstawowych problemów spotykanych w kombinatoryce słów. Najważniejszymi typami badanych powtórzeń są kwadraty (powtórzenia postaci zz) oraz maksymalne powtórzenia.

Oznaczmy przez $sq(w)$ liczbę parami różnych kwadratów, zaś przez $\rho(w)$ liczbę maksymalnych powtórzeń w słowie w; dodatkowo niech $sq(n)$ oraz $\rho(n)$ oznaczają odpowiednio maksymalną liczbę kwadratów oraz maksymalnych powtórzeń w słowach długości n. Mimo wielu badań w tej dziedzinie dokładne asymptotyczne oszacowanie dla $sq(n)$ oraz $\rho(n)$ nadal nie jest znane, zaś algorytmy wyznaczania $sq(w)$ oraz $\rho(w)$ są skomplikowane.

W rozprawie zbadana została struktura wystąpień powtórzeń dla klasy standardowych słów Sturma $\mathcal{S}$ – jednej z intensywniej badanych klas w kombinatoryce słów. Posiadają one zwarte reprezentacje w postaci ciągów dodatnich liczb całkowitych. Słowo standardowe generowane przez ciąg γ oznaczamy $\mathrm{Sw}(\gamma)$. Długość $\mathrm{Sw}(\gamma)$ jest zwykle wykładniczo zależna od długości γ, mamy więc do czynienia z badaniem struktury powtórzeń w skompresowanej wersji słów, co dodatkowo komplikuje dowody i algorytmy. Słowa standardowe są szczególnym przypadkiem przesunięć cyklicznych słów Christoffela – stanowiących kolejną ważną klasę słów mającą wiele zastosowań geometrycznych.

Głównymi wynikami rozprawy są:

- algorytm znajdowania $\rho(w)$ dla $w \in \mathcal{S}$ w czasie liniowo zależnym od jego (zazwyczaj logarytmicznej) skompresowanej reprezentacji γ;
- asymptotyczna granica $\rho(w) \leq 0.8|w|$ dla $w \in \mathcal{S}$;
- algorytmiczna konstrukcja nieskończonego ciągu słów standardowych $\{w_k\}$ asymptotycznie osiągającego granicę $\rho(w_k) = 0.8|w_k| - o(|w_k|)$;
- asymptotyczna granica $sq(|w|) \leq 0.9|w|$ dla $w \in \mathcal{S}$ ($\mathcal{S}$ stanowi największą klasę słów, dla której znane jest dokładne asymptotyczne oszacowanie, dla ogólniejszych klas słów najlepszym oszacowaniem jest $n \leq sq(n) \leq 2n$, a uzyskanie dokładniejszego ograniczenia jest bardzo trudne);
- algorytmiczna konstrukcja nieskończonego ciągu słów standardowych $\{w_k\}$ asymptotycznie osiągającego granicę $sq(w_k) = 0.8|w_k| - o(|w_k|)$;
- zbadanie struktury skompresowanych grafów podsłów dla słów standardowych – rozmiar tych grafów jest liniowo zależny od γ, przy ich pomocy uzyskano kilka efektywnych algorytmów związanych z kombinatorycznymi własnościami słów standardowych;
- związek między strukturą grafów podsłów słów standardowych a pewnymi systemami liczbowymi oraz pewną szczególną klasą automatów skończonych.

Wyniki prezentowane w rozprawie zostały opublikowane w pracach [5], [6] oraz [61].

Słowa kluczowe: kombinatoryka słów, słowa Sturma, maksymalne powtórzenia, kwadraty, kompresja danych, grafy podsłów, systemy liczbowe

Klasyfikacja tematyczna AMS: 68R15, 68P30, 68Q70.

[*] Badania wspierane przez Ministerstwo Nauki i Szkolnictwa Wyższego, grant N N206 258035

Contents

Introduction

A sequence of symbols taken from a finite alphabet is one of the simplest and natural ways of information representation. Such a sequence is called a word or a string. Words are central objects of any standard model of computing. Even in cases when we compute on numbers, their representations of can be seen as strings, hence it is natural to study algorithmic properties of words.

The theory of combinatorics on words was started at the beginning of the XX-th century by A. Thue, see [73] (1906) and [74] (1912), but the notion of a word can be found in several older mathematical works. During the last three decades the research on combinatorial problems on words has grown enormously and culminated with the books [50] (1983), [51] (2002) and [52] (2005). Currently combinatorics on words has become a rich area, with many connections to algorithms, number theory, symbolic dynamics, and applications in biology and text processing.

Some examples of problems arising in combinatorics on words are:

String matching, also called pattern matching, is the problem of finding occurrences of one string called the *pattern* (or a finite set of patterns) in some larger text. It is the most thoroughly studied problem in combinatorics on words. The main theoretical tools in string-matching algorithms are related to the properties of regularities in strings, see [7], [23], [46].

Text compression is one of the basic problems in data storage, data transmission, etc. Compression means reducing the representation of a text in such a way that the original text can be easily recovered from its compressed form. To achieve better compression ratios it is necessary to analyse the data, hence the problem of detecting regularities in texts is very important. For more information see [36], [58], [65], [69].

Computational biology is a domain, where text algorithms appear in the study of molecular sequences. For example the reconstruction of a whole sequence from small segments can be seen as the shortest common superstring problem (construction of the shortest text containing several given smaller texts) and the alignment of two sequences of nucleotides – as the edit distance problem (computation of the minimal number of edit operations transforming one string into another). Because of the large amount of data involved the efficiency of the algorithms considered is necessary, see for example [34], [41].

Digital geometry can be seen as the geometry of a computer screen and is related to computer graphics. Its main problem is digitization of geometric objects, such as points, lines, polygons and so on, by finding their representation on a discrete plane $\mathbb{Z}$x$\mathbb{Z}$. It is done by selecting pixels that are close enough to the object they approximate. An interesting example of application is an approximation of a continuous curve by unit line segments (horizontal, vertical or diagonal) and encoding each line segment by a letter depending on its direction. These letters form a word and we can try to derive some properties of the encoded curve by analysing the word structure. More information and some recent results can be found in [13], [25], [45], [75].

For a detailed introduction to combinatorics on words, some historical background and more examples of applications we refer the reader to [10], [12], [23], [44], [50], [51], [52] and references therein.

Basic definitions

This section provides a brief introduction to the concepts used in the following chapters and fixes the general notations.

The symbols $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ denote the sets of nonnegative integer, integer, rational and real numbers. By $|X|$ we denote the cardinality of the set X.

Two positive integer numbers p and q are said to be relatively prime (denoted by $p \perp q$) if 1 is the only positive integer that divides both p and q.

Let Σ be a finite set called the *alphabet.* Elements of the alphabet are called *letters* (symbols, characters). A finite word over the alphabet Σ is a finite sequence of elements of Σ:

$$(a_1, a_2, \dots, a_n), \qquad a_i \in \Sigma.$$

The set of all finite words over Σ, denoted by Σ^*, is equipped with a binary associative operation $\cdot$ obtained by concatenating two sequences

$$(a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_n) \;=\; (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n).$$

We identify the letter $a \in \Sigma$ with the sequence (a) and for simplicity write a word as

$$a_1 a_2 \ldots a_n$$

and usually omit the symbol of the concatenation. The length of a word is equal to the number of its letters:

$$|a_1 a_2 \ldots a_n| = n.$$

For a letter $a \in \Sigma$ and a word w we denote by $|w|_a$ the number of occurrences of the letter a in w and the i-th letter of w is denoted by $w[i]$. An empty sequence of letters, called an *empty word* and denoted by ε, is the neutral element of concatenation, thus for any word w we have

$$\varepsilon \cdot w \;=\; w \cdot \varepsilon \;=\; w.$$

The set Σ^* with the operation of concatenation and the empty word has a monoid structure and is called the *free monoid.* The set $\Sigma^+ = \Sigma^* - \{\varepsilon\}$ of all nonempty words over Σ is called the *free semigroup.*

An infinite word is a sequence of symbols indexed by nonnegative integers. The set of all infinite words over the alphabet Σ is denoted by $\Sigma^{\mathbb{N}}$. Infinite words can be also defined as limits of infinite sequences of finite words. The set of all finite and infinite words over Σ is denoted by $\Sigma^\infty = \Sigma^{\mathbb{N}} \cup \Sigma^*$.

A word u is called a *factor* or a *subword* of a word w if there exist words x and y such that $w = xuy$. If $y = \varepsilon$ then u is called a *prefix* of w and if $x = \varepsilon$ then u is called a *suffix* of w. A factor is proper if $xy \neq \varepsilon$, a prefix is proper if $y \neq \varepsilon$ and a suffix is proper if $x \neq \varepsilon$. For a word $w = a_1 a_2 \ldots a_n$ and $1 \leq i, j \leq n$ we denote by $w[i..j]$ the factor of w of the form $a_i a_{i+1} \ldots a_j$. The set of all factors of w is denoted by $F(w)$.

A repetition is a word composed (as a concatenation) of several copies of another word: $w = v^k$. The exponent of a repetition is the number of such copies. We can extend the notion of a repetition and allow the exponent to be a rational number. In this case we have $w = v^k u$, where u is the proper prefix of v (possibly empty).

A word $w \in \Sigma^+$ is called *primitive* if the equality $w = u^n$ for some $u \in \Sigma^+$ implies $n = 1$. A word w is called *periodic* if $w = u^k$ for some nonempty word u. A word w is called *eventually periodic* if $w = vu^k$ for some words v

and u (u – nonempty). A word w is called *aperiodic* if it is neither periodic nor eventually periodic (no suffix of w is periodic).

Let $w = a_1a_2 \dots a_n$ ($a_i \in \Sigma$) be a word. A positive integer p is called a *period* of w if $a_i = a_{i+p}$ for $1 \le i \le n - p$. The smallest period of w is called *the period* of w and denoted $period(w)$.

A reverse of a word $w = a_1a_2 \dots a_n$ is the word $w^R = a_n \dots a_2a_1$. A word w is called a *palindrome* if $w = x \cdot x^R$ for a word x or $w = x \cdot a \cdot x^R$ for a word x and a letter a. Denote also by $\overline{w}$ the word w without the last two letters.

A word w is called *balanced* if for any letter a and any two factors x, y of w, such that $|x| = |y|$, we can state that

$$\Big| |x|_a - |y|_a \Big| \le 1.$$

The words x, y are called *conjugate* if there exist words u, v such that $x = uv$ and $y = vu$. Thus conjugate words are cyclic shifts of one another.

A function $f : A^* \longrightarrow B^*$ is a *morphism* if we have

$$f(x \cdot y) = f(x) \cdot f(y)$$

for all $x, y \in A^*$. A morphism is uniquely determined by its values on the alphabet and it is obvious that f maps the neutral element of A^* into the neutral element of B^*. A morphism is called *literal* if the image of a letter is a letter and *nonerasing* if the image of a letter is always a nonempty word.

The set of words over a finite alphabet Σ can be seen as a tree. Its vertices are elements of Σ^*. The root is the empty word ε. The sons of a node w are the words wa for $a \in \Sigma$. A word w can be also viewed as the path leading from the root to the node w. A word x is a prefix of a word y if it is its ancestor in the tree.

Sometimes we need an order relation for words to be defined. The *lexicographic order*, also called the *alphabetic order*, is defined as follows: given a strict order on the alphabet for any two words x and y we have $x < y$ if x is a proper prefix of y or there exist factorizations $x = uav_1$, $y = ubv_2$ where a, b are letters and $a < b$.

Outline of the thesis

In this thesis we investigate problems related to combinatorial structure and repetitions in standard Sturmian words – one of the most investigated class of strings in combinatorics on words. The thesis is organized as follows:

In the **chapter 1** we define the class $\mathcal{S}$ of standard Sturmian words – the binary aperiodic words with minimal combinatorial complexity. We start with a simple recurrent definition, which leads to an efficient compressed representation as a sequence of integer numbers with the size logarithmic with respect to the size of the word. Next we define them from the geometrical point of view as discretizations of straight lines on a discrete plane and a labeling of the edges of the Cayley graph of some finite group connected to the number of letters. We also present a simple morphic representation and describe standard words in arithmetic form using continued fractions.

The main purpose of the **chapter 2** is the investigation of the structure of subword graphs of standard words in more detail than in previous works. The special structure of those graphs (especially their compacted versions) leads to simple alternative graph-based proofs of several known facts and to special easy algorithms computing some properties of Sturmian words: the number of subwords, the critical factorization point, lexicographically maximal suffixes, occurrences of subwords of a fixed length and right special factors. The algorithms presented here work in linear time with respect to the size of the compressed representation of standard words.

The structure of subword graphs described in the chapter 2 implies a simple characterization of the periods of runs (maximal repetitions) in standard words. Using this characterization we derive in the **chapter 3** an explicit formula for the number $\rho(w)$ of runs in words $w \in \mathcal{S}$. This formula depends only on the compressed representation of standard words and leads to the algorithm computing $\rho(w)$ for $w \in \mathcal{S}$ in linear time with respect to the size of the compressed representation. We also show that

$$\frac{\rho(w)}{|w|} \leq \frac{4}{5} \quad \text{for each} \quad w \in \mathcal{S},$$

and there is an infinite sequence of strictly growing words $w_k \in \mathcal{S}$ such that

$$\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} = \frac{4}{5}.$$

The complete understanding of the structure of maximal repetitions for a large class $\mathcal{S}$ of complicated words is a step towards a better understanding of this problem in general class of words.

The **chapter 4** continues the investigation of the structure of repetitions started in the chapter 3. We use the results of [24], where exact (but not closed) complicated formulas were given for the number $sq(w)$ of squares in standard words. We slightly improve those formulas and show that:

$$sq(w) \leq \frac{9}{10}|w| \quad \text{for all} \quad w \in \mathcal{S}$$

and there is an infinite sequence of strictly growing words $w_k \in \mathcal{S}$ such that

$$\lim_{k\to\infty} \frac{sq(w_k)}{|w_k|} = \frac{9}{10}.$$

At present $\mathcal{S}$ is the largest class of words, for which the exact coefficient is known.

In this chapter we have also performed the asymptotical analysis of the maximal number of distinct squares and the maximal number of runs for Christoffel words.

In the **chapter 5** we use the structure of subword graphs of standard words investigated in the chapter 2 to describe the dual Ostrowski numeration system. It can be defined without any reference to those graphs, but representations of integer numbers in this numeration system have simple connections to lengths of paths in a subword graph of some $w \in \mathcal{S}$. We introduce also a new concept related to standard words: the Ostrowski automata.

The maximal repetition ratio 0.8 and the distinct square ratio 0.9 have been discovered by us doing experiments with very long standard Sturmian words. Similarly, we were tuning many intermediate formulas from the chapters 3 and 4 with the assistance of the computer. Some useful applets related to problems considered in this thesis can be found on the web site:

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

1 Standard Sturmian words

Sturmian words are infinite words over a binary alphabet that for each $k > 0$ have exactly $k+1$ distinct factors of the length k. They can be equivalently defined as balanced and aperiodic words with minimal combinatorial complexity. These words form an interesting class of words and have been studied by many researchers for their theoretical importance and their applications to various fields of science. They appear in many domains, such as: mathematics, computer science, digital geometry, computational biology, physics, astronomy and even music. Sturmian words can be found in literature under several different names: rotation sequences, cutting sequences, mechanical sequences, Christoffel words, Beatty sequences, characteristic sequences, balanced sequences and so on (see for example [1], [7], [9], [11], [51], [70], [71], [75] and references therein).

The theory of Sturmian sequences was started in the late XVIII-th century, see for instance [8] (1772), [15] (1875), [16] (1888), [72] (1786) and [56] (1882). The name *Sturmian*, first used in 1940 (see [57]), goes after the Swiss-born French mathematician Charles François Sturm (1803-1855), famous for his rule to compute the roots of an algebraic equation and the Sturm-Liouville problem – an eigenvalue problem in second order differential equations.

Suppose that $u(x)$ is the solution of the linear homogeneous differential equation

$$y'' + \phi(x)y \;=\; 0,$$

where $\phi(x)$ is continuous function of period 1, and k_n denotes the number of zeroes of $u(x)$ in the interval $[n, n+1)$, then the infinite word

$$w \;=\; ba^{k_0}ba^{k_1}ba^{k_2}\ldots$$

is either Sturmian or eventually periodic (see [57] for more details).

In this thesis we investigate Sturmian words from the algorithmic point of view and therefore we focus on their finite prefixes – the *standard words.* The number of applications of those words causes the existence of many equivalent definitions. The aim of this chapter is to present some of them.

We start with a recurrent definition of standard words based on a sequence of integer numbers and leading to an efficient compressed representation. Then we describe them from the geometrical point of view as discretizations of straight lines on a discrete plane and a labeling of the edges of Cayley graph of a finite group connected to the number of letters in the word generated. Next, we present a simple morphic representation of standard words and show a fast algorithm for checking if a given word is standard. Finally, we show the correspondence between the combinatorics on words and the number theory by describing standard words in arithmetic form using continued fractions.

1.1 Standard words

In this section we describe the class $\mathcal{S}$ of *standard words* – finite words that are prefixes of infinite characteristic Sturmian words. We present a definition based on recurrences, which leads to a grammar-based compression and very useful compact representation of standard words by sequences of positive integers. This will be the main definition used in the following chapters.

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:

$$\begin{aligned}
& x_{-1} = b, && x_0 = a, \\
& x_1 = x_0^{\gamma_0} x_{-1}, && x_2 = x_1^{\gamma_1} x_0, \\
& \vdots && \vdots \\
& x_n = x_{n-1}^{\gamma_{n-1}} x_{n-2}, && x_{n+1} = x_n^{\gamma_n} x_{n-1},
\end{aligned} \tag{1.1}$$

where $\mathrm{Sw}(\gamma) = x_{n+1}$.

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 $\mathrm{Sw}(0) = b$. The class of all standard words is denoted by $\mathcal{S}$.

Example 1.2
Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have:

$$\begin{array}{lclcl}
x_{-1} & = & b & & \\
x_0 & = & a & & \\
x_1 & = & (x_0)^1 \cdot x_{-1} & = & a \cdot b \\
x_2 & = & (x_1)^2 \cdot x_0 & = & ab \cdot ab \cdot a \\
x_3 & = & (x_2)^1 \cdot x_1 & = & ababa \cdot ab \\
x_4 & = & (x_3)^3 \cdot x_2 & = & ababaab \cdot ababaab \cdot ababaab \cdot ababa \\
x_5 & = & (x_4)^1 \cdot x_3 & = & ababaabababaabababaabababa \cdot ababaab
\end{array}$$

and finally

$$\mathrm{Sw}(1, 2, 1, 3, 1) \;=\; ababaabababaabababaabababaababaab.$$

□

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 the mapping:

$$E : \left\{ \begin{array}{l} a \longrightarrow b \\ b \longrightarrow a \end{array} \right. . \tag{1.3}$$

Example 1.4
Consider the directive sequence $\gamma = (0, 1, 2, 1, 3, 1)$. We have:

$$\begin{array}{lclcl}
x_{-1} & = & b & & \\
x_0 & = & a & & \\
x_1 & = & (x_0)^0 \cdot x_{-1} & = & \varepsilon \cdot b \\
x_2 & = & (x_1)^1 \cdot x_0 & = & b \cdot a \\
x_3 & = & (x_2)^2 \cdot x_1 & = & ba \cdot ba \cdot b \\
x_4 & = & (x_3)^1 \cdot x_2 & = & babab \cdot ba \\
x_5 & = & (x_4)^3 \cdot x_3 & = & bababba \cdot bababba \cdot bababba \cdot babab \\
x_6 & = & (x_5)^1 \cdot x_4 & = & bababbabababbabababbababab \cdot bababba
\end{array}$$

and finally

$$\mathrm{Sw}(0, 1, 2, 1, 3, 1) \;=\; bababbabababbabababbabababbababba.$$

Compare with the word $\mathrm{Sw}(1, 2, 1, 3, 1)$ from Example 1.2. □

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, every standard word consists either of repeated occurrences of the letter a separated by single occurrences of the letter b or repeated occurrences of the letter b separated by single occurrences of the letter a. Those letters are called the *repeating letter* and the *single letter*, respectively. If the repeating letter is a (letter b respectively), the word is called the Sturmian word of the type a (type b respectively), see the definition 6.1.4 in [63] for comparison.

Remark 1.5
Without loss of generality we consider in this thesis the standard Sturmian words of the type a, therefore we assume that $\gamma_0 > 0$. The words of the type b can be considered similarly and all the results hold.

Fact 1.6 (See [51])
Let p and q be relatively prime positive integers. There exist exactly two standard words with p letters b and q letters a, namely: $w \cdot ab$ and $w \cdot ba$.

Consider the directive sequences

$$\gamma^{I} = (\gamma_0, \gamma_1, \ldots, \gamma_n, 1) \qquad \text{and} \qquad \gamma^{II} = (\gamma_0, \gamma_1, \ldots, \gamma_n + 1).$$

By the equation (1.1) we have:

$$\mathrm{Sw}(\gamma^{I}) = x_n^{\gamma_n} \cdot x_{n-1} \cdot x_n \qquad \text{and} \qquad \mathrm{Sw}(\gamma^{II}) = x_n^{\gamma_n} \cdot x_n \cdot x_{n-1}.$$

Simple induction shows that the word $x_{n-1} \cdot x_n$ is the same as $x_n \cdot x_{n-1}$ up to the last two letters.

As a direct corollary from Fact 1.6 we know that the number of standard words of the length n is $2 \cdot \phi(n)$, where ϕ is the Euler's totient function defined for $n \geq 1$ as the number of positive integers less than n and relatively prime to n, see corollary 2.2.16 in [51].

Example 1.7
Consider the directive sequence $\gamma = (1, 2, 1, 4)$. By the equation (1.1) we have:

$$\begin{array}{lllll}
x_{-1} & = & b & & \\
x_0 & = & a & & \\
x_1 & = & (x_0)^1 \cdot x_{-1} & = & a \cdot b \\
x_2 & = & (x_1)^2 \cdot x_0 & = & ab \cdot ab \cdot a \\
x_3 & = & (x_2)^1 \cdot x_1 & = & ababa \cdot ab \\
x_4 & = & (x_3)^4 \cdot x_2 & = & ababaab \cdot ababaab \cdot ababaab \cdot ababaab \cdot ababa
\end{array}$$

and finally

$$\mathrm{Sw}(1, 2, 1, 4) \;=\; ababaabababaabababaabababaabababa.$$

Observe that Sw$(1, 2, 1, 4)$ and Sw$(1, 2, 1, 3, 1)$ from Example 1.2 differ only on the last two letters. In fact Sw$(1, 2, 1, 4)$ and Sw$(1, 2, 1, 3, 1)$ are the only standard words having 19 letters a and 14 letters b. □

The number $N = |\mathrm{Sw}(\gamma)|$ is the (real) size of the word, while $(n + 1) = |\gamma|$ can be thought as the compressed size. Observe that, by the definition of standard words, N is exponential with respect to n. 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 grammar is proportional to the length of the directive sequence.

Standard words can be also defined by the set of *standard pairs*. Every standard word, which is not a letter, is a product of two standard words, which are components of some standard pair, see section 2.2.1 in [51] for more details.

The following fact indicates the relation between finite standard words and infinite characteristic Sturmian words (see section 2.2.2 of [51] for the proof).

Fact 1.8 (See [51])
Let Sw(γ^I) *and* Sw(γ^{II}) *be standard words. If* γ^I *is a prefix of* γ^{II} *then* Sw(γ^I) *is a prefix of* Sw(γ^{II}). *Therefore, every infinite characteristic Sturmian word can be seen as the limit of a sequence of finite standard words.*

Fibonacci words

Fibonacci words are a well known family of strings. They are formed by repeated concatenation in the same way that the Fibonacci numbers are formed by repeated addition. The n-th Fibonacci word F_n is given by the recurrence:

$$F_{-1} \;=\; b, \qquad F_0 \;=\; a, \qquad \ldots, \qquad F_{n+1} \;=\; F_n \cdot F_{n-1}.$$

The lengths of Fibonacci words are given as the Fibonacci numbers:

$$\begin{array}{lcl@{\qquad\qquad}rcl}
F_0 & = & a & |F_0| & = & 1 \\
F_1 & = & ab & |F_1| & = & 2 \\
F_2 & = & aba & |F_2| & = & 3 \\
F_3 & = & abaab & |F_3| & = & 5 \\
F_4 & = & abaababa & |F_4| & = & 8 \\
F_5 & = & abaababaabaab & |F_5| & = & 13 \\
 & & & \vdots & &
\end{array}$$

By the 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).

These words satisfy a large number of interesting properties related to periods and repetitions. For example they have no factor of the form v^4, where v is some nonempty word. For more information on the Fibonacci words and their properties see for example [4], [28], [40], [51] and [67].

1.2 Christoffel words

In this section we define the class of Christoffel words, which are the special kind of cyclic shifts of standard words. We characterize them from the geometrical point of view as the discretization of a line segment in the plane by the path in the integer lattice $\mathbb{Z}$x$\mathbb{Z}$ and as a labelling the edges of the Cayley graph of some finite group (see [9], [11] and [15]).

Let p and q be two relatively prime integers. The *lower Christoffel path* of slope $\frac{p}{q}$ is the path in the discrete plane from the point $(0, 0)$ to (p, q) that consists of horizontal and vertical unit line segments and satisfies the following conditions:

- the path lies below the line segment beginning at the point $(0, 0)$ and ending at (p, q),
- the region in the plane enclosed by the path and the line segment contains no other points with integer coordinates besides those of the path.

The *upper Christoffel path* is defined in the same manner above the line segment. The *lower Christoffel word* and the *upper Christoffel word* are determined from the lower and upper Christoffel paths by encoding every horizontal line segment by the letter b and each vertical line segment by the letter a, see Figure 1.1 for an example. The unmodified term *Christoffel word* always means the lower Christoffel word.

Example 1.9
Let $p = 19$ and $q = 14$. The lower Christoffel word of slope $\frac{19}{14}$ is

$$c_l \;=\; bababaabababaabababaabababaababaa,$$

and the upper Christoffel word of slope $\frac{19}{14}$ is

$$c_u \;=\; aababaabababaabababaabababaababab.$$

The lower and upper Christoffel paths of slope $\frac{19}{14}$ and the labeling corresponding to c_l and c_u are depicted on Figure 1.1.

□

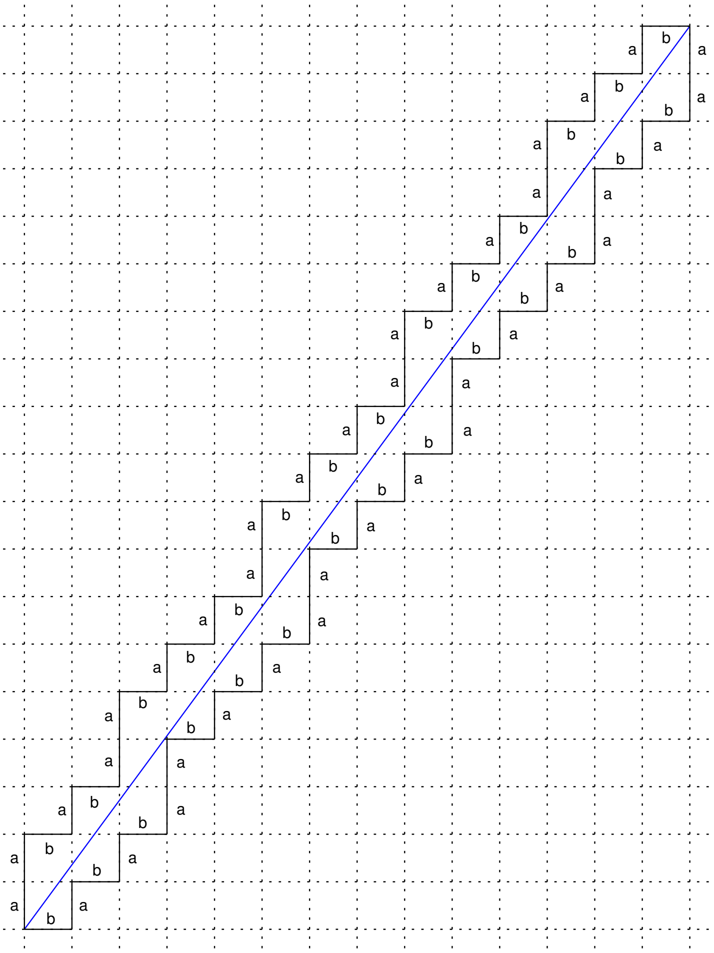

Figure 1.1: *The lower and upper Christoffel words of slope* $\frac{19}{14}$ *are* *bababaababababaababababaababababaababaa, aababaababababaababababaababababaababab.*

Recall the notion of the single and the repeating letter mentioned in the previous section. Observe also that the Christoffel word of slope $\frac{p}{q}$ consists of p letters a and q letters b. Therefore, if $p > q$ then a is the repeating letter and if $p < q$ then b is the repeating letter. Recall also that we consider the letter a to be the repeating letter (see Remark 1.5), hence assume $p > q$. Christoffel words of slope less than one can be considered similarly. In fact the morphism E from the equation (1.3) maps the lower Christoffel word of slope x to the upper Christoffel word of slope $\frac{1}{x}$, see lemma 2.6 in [11].

Since every positive rational number x can be uniquely expressed by the fraction $\frac{p}{q}$, where p and q are relatively primes, there are unique lower and upper Christoffel words of slope x, hence containing exactly q letters a and p letters b.

The relation between standard and Christoffel words is given by the following fact (see Proposition 9 in [9]):

FACT 1.10 (See [9])
Let a, $b \in \Sigma$ and $w \in \Sigma^*$. The following conditions are equivalent:

(1) *wab is a standard word,*

(2) *wba is a standard word,*

(3) *bwa is a lower Christoffel word,*

(4) *awb is an upper Christoffel word.*

Observe that the words from the points 1 an 3 (2 and 4 respectively) of Fact 1.10 are conjugate.

EXAMPLE 1.11
Let $w = ababaababababaabababaababababaababa$. We have:

- $w \cdot ab = ababaababababaabababaababababaababa \cdot ab$ – is the standard word given by the directive sequence $\gamma = (1, 2, 1, 3, 1)$, see Example 1.2,
- $w \cdot ba = ababaababababaabababaababababaababa \cdot ba$ – is the standard word given by the directive sequence $\gamma = (1, 2, 1, 4)$, see Example 1.7,
- $b{\cdot}w{\cdot}a = b{\cdot}ababaababababaabababaababababaababa{\cdot}a$ – is the lower Christoffel word of slope $\frac{19}{14}$, see Example 1.9,
- $a{\cdot}w{\cdot}b = a{\cdot}ababaababababaabababaababababaababa{\cdot}b$ – is the upper Christoffel word of slope $\frac{19}{14}$, see Example 1.9.

□

The following observation will be useful in the subsequent chapters:

REMARK 1.12
A standard word w without the last two letters ($\overline{w}$) is a palindrome. Hence every Christoffel word is of the form xuy, where x, y are letters and u is a palindrome word. Those palindrome words are called the *central words*. See proposition 4.2 in [11] and theorem 2.2.4 in [51] for the proof.

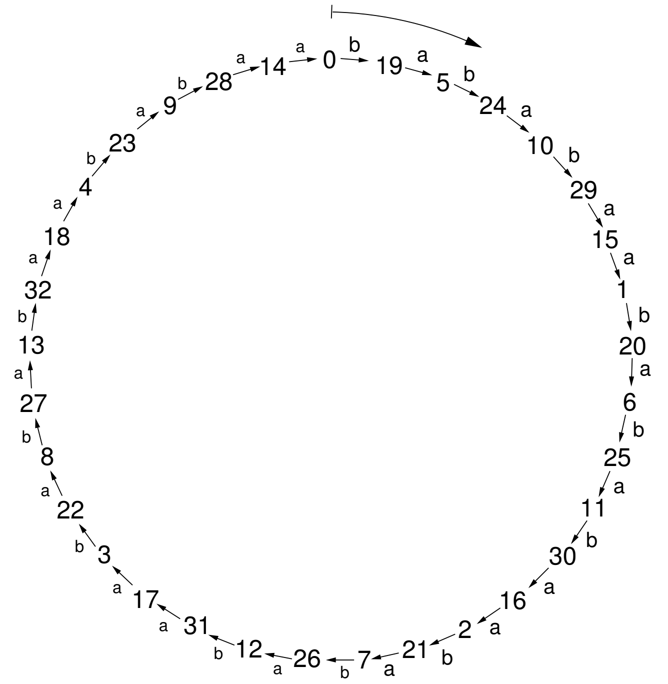

Figure 1.2: *The Cayley graph of the group* $\mathbb{Z}/(14+19)\mathbb{Z}$ *with the generator* 19*. Reading labels of its edges clockwise starting from* 0 *we obtain the lower Christoffel word of slope* $\frac{19}{14}$*:* $c_l = bababaabababaababababaabababaababaa$*.*

A Christoffel word of the slope $\frac{p}{q}$ can be equivalently defined by means of a Cayley graph (Cayley diagram, group graph) of the group $\mathbb{Z}/(p+q)\mathbb{Z}$. For a group G and a set of its generators S we define a directed graph $\mathcal{G}$, which consists of the vertices corresponding to all elements of G. An edge (u, v) is present in $\mathcal{G}$ if some generator from S transfers u into v. See [14], [53] and [59] for more details.

Let p, q be positive relatively prime integers. Consider the group $\mathbb{Z}/(p+q)\mathbb{Z}$ with the generator p. The Cayley graph of this group is the cycle $C_{(p+q)}$, with vertices 0, p, $2p$, $3p$,..., q, $0 \bmod (p+q)$. With every edge (s, t) we associate the label: "b" if $s < t$ and "a" if $s > t$. Those labels read consecutively starting from the vertex 0 form the lower Christoffel word of the slope $\frac{p}{q}$, see Figure 1.2.

If we consider the group $\mathbb{Z}/(p+q)\mathbb{Z}$ with the generator q instead of p and swap the roles of a and b in the above definition, we obtain the upper Christoffel word of the slope $\frac{p}{q}$. Observe also, that if we start reading the labels of edges from the edge labeled by the generator we obtain the standard word.

Example 1.13
Let $p = 19$, $q = 14$ and consider the Cayley graph of the group $\mathbb{Z}/33\mathbb{Z}$ with the generator 19. Consecutive vertices of this graph are: 0, 19, 5, 24, 10, 29, 15, 1, 20, 6, 25, 11, 30, 16, 2, 21, 7, 26, 12, 31, 17, 3, 22, 8, 27, 13, 32, 18, 4, 23, 9, 28, 14, 0. The labeling of its edges is depicted on Figure 1.2. By reading the edge labels clockwise (starting from 0) we obtain the lower Christoffel word of the slope $\frac{19}{14}$:

$$c_l = bababaabababaababababaabababaababaa.$$

Switching the generator and swapping the roles of a and b is the same as reading the labels of the edges of the Cayley graph counterclockwise (starting from 0). We then obtain the upper Christoffel word of the slope $\frac{19}{14}$:

$$c_u = aababaabababaababababaabababaababab.$$

Compare with Example 1.9. □

Christoffel words have a natural generalization to infinite sequences: we replace the defining line segment of slope $\frac{p}{q}$ with an infinite ray of irrational slope before building the lattice path. The resulting right infinite word is called a characteristic Sturmian word, see [1], [51] or [63] for more detailed information.

1.3 Morphic representation of standard words

The recurrent definition of standard words from the section 1.1 leads to the simple characterization by the 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

$$h_i : \begin{cases} a & \longrightarrow & a^{\gamma_i} b \\ b & \longrightarrow & a \end{cases} \qquad \text{for } 0 \leq i \leq n. \tag{1.14}$$

Lemma 1.15
For $0 \leq i \leq n$ the morphism h_i transforms a standard word into another standard word, and we have:

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

Proof
The induction on the length of the directive sequence.

Recall that the standard word given by the empty directive sequence is a. For $|\gamma| = 1$ we have, by definition of standard words and the morphism h_n,

$$\mathrm{Sw}(\gamma_n) \;=\; a^{\gamma_n} b \;=\; h_n(a).$$

Assume now that $|\gamma| = k \geq 2$ and for directive sequences shorter than k the thesis holds. We have then:

$$\begin{aligned}
\mathrm{Sw}(\gamma_i, \ldots, \gamma_n) &= \Big[\mathrm{Sw}(\gamma_i, \ldots, \gamma_{n-1})\Big]^{\gamma_n} \cdot \mathrm{Sw}(\gamma_i, \ldots, \gamma_{n-2}) \\
&\overset{ind.}{=} \Big[h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-1})\Big)\Big]^{\gamma_n} \cdot h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-2})\Big) \\
&= h_i\Big(\big[\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-1})\big]^{\gamma_n} \cdot \mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_{n-2})\Big) \\
&= h_i\Big(\mathrm{Sw}(\gamma_{i+1}, \ldots, \gamma_n)\Big),
\end{aligned}$$

which concludes the proof. □

As a direct conclusion from Lemma 1.15 we have that for the directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$

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

Example 1.16
Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$.
We have:

$$\begin{aligned}
\mathrm{Sw}(1) &= h_4(a) && = ab \\
\mathrm{Sw}(3,1) &= h_3\big(\mathrm{Sw}(1)\big) && = aaaba \\
\mathrm{Sw}(1,3,1) &= h_2\big(\mathrm{Sw}(3,1)\big) && = abababaab \\
\mathrm{Sw}(2,1,3,1) &= h_1\big(\mathrm{Sw}(1,3,1)\big) && = aabaaabaaabaaabaaba \\
\mathrm{Sw}(1,2,1,3,1) &= h_0\big(\mathrm{Sw}(2,1,3,1)\big) && = ababaababababaababababaabababaababaab.
\end{aligned}$$

Compare with Example 1.2. □

Recall the notion of the single and the repeating letter from the section 1.1. This concept can be similarly considered for larger factors. Observe that every standard word w can be divided into blocks of the form $a^k b$ $(k > 0)$, that can not be extended to the left in w, and sometimes an additional letter a at the end of w. The blocks in w can be of the two types: a^kb (the *short block*) and $a^{k+1}b$ (the *long block*). The word w can not include blocks of any other type without loosing the balance property. Hence w consists of repeated occurrences of the block of one type separated by single occurrences of the block of the second type, called the *repeating block* and the *single block* respectively (in the other case we have contradiction with aperiodicity and balance property).

Assume that the repeating block is a^kb and the single block $a^{k+1}b$ (the other case is similar) and consider the morphism g defined as:

$$g: \left\{ \begin{array}{ccc} a & \longrightarrow & a^kb \\ b & \longrightarrow & a^{k+1}b \end{array} \right. .$$

The word w is Sturmian if and only if the word $g^{-1}(w)$ is Sturmian (the single rightmost a can be omitted), see theorem 2.1 in [29]. This condition leads to a simple algorithmic method that allows us to check whether a given (finite) word w is standard. The algorithm presented here is a slight modification of the algorithm from [29].

For a given finite word w find out which block a^kb or $a^{k+1}b$ is the repeating block, and then reduce w to the word $g^{-1}(w)$. If w is Sturmian then we get the empty word after several steps, see [29] for more details.

Example 1.17

Consider the word from Example 1.2:

$$w_0 = ab \cdot ab \cdot aab \cdot ab \cdot ab \cdot aab \cdot ab \cdot ab \cdot aab \cdot ab \cdot ab \cdot aab \cdot ab \cdot aab.$$

Observe that the repeating block is ab and the single block is aab. We reduce w_0 to the word w_1 by encoding every repeating block by the letter a and every single block by the letter b:

$$w_1 = aab \cdot aab \cdot aab \cdot aab \cdot ab.$$

In this case, by encoding the repeating block ab by a and the single block aab by b, we reduce w_1 to:

$$w_2 = aaaab.$$

The word w_2 consists of one single block, which is encoded by a. If we omit the rightmost letter a we obtain $w_3 = \varepsilon$, hence w_0 is standard Sturmian word.

□

Some other interesting morphic representations of finite and infinite Sturmian words can be found for example in [1], [43] and [51].

1.4 Continued fractions

In this section we define the class $\mathcal{S}$ of standard words from the number theoretical point of view using the notion of continued fractions.

A *finite continued fraction* is an expression of the form:

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots \cfrac{1}{a_n}}}},$$

which is denoted for simplicity as $[a_0; a_1, a_2, a_3, \ldots, a_n]$, and called the continued fraction expansion (or CF-expansion in short). The integer number a_0 and the positive integer numbers $a_1, a_2, \ldots, a_n$ are called the *partial quotients.*

Every finite continued fraction represents some rational number and every rational number can be represented as a finite continued fraction in two ways:

$$\frac{p}{q} = [a_0; a_1, a_2, a_3, \ldots, a_n] \qquad \text{or} \qquad \frac{p}{q} = [a_0; a_1, a_2, a_3, \ldots, a_n - 1, 1].$$

The uniqueness of such a representation can be achieved by avoiding the number 1 to be the last element of a CF-expansion (see [64]).

Observe that if a rational number $\frac{p}{q}$ has the CF-expansion $[a_0; a_1, \ldots, a_n]$ and $a_0 > 0$, then its inversion has the CF-expansion $[0; a_0, a_1, \ldots, a_n]$.

Example 1.18
Consider a rational number $\frac{19}{14}$. Using simple arithmetic operations we can write it as

$$\frac{19}{14} = \mathbf{1} + \frac{5}{14} = \mathbf{1} + \cfrac{1}{\mathbf{2} + \cfrac{4}{5}} = \mathbf{1} + \cfrac{1}{\mathbf{2} + \cfrac{1}{\mathbf{1} + \cfrac{1}{\mathbf{4}}}} \tag{1.19}$$

and we achieve its CF-expansion:

$$\frac{19}{14} = [1; 2, 1, 4].$$

The last component of the equation (1.19) can be written as

$$\mathbf{1}+\cfrac{1}{\mathbf{2}+\cfrac{1}{\mathbf{1}+\cfrac{1}{\mathbf{4}}}} \;=\; \mathbf{1}+\cfrac{1}{\mathbf{2}+\cfrac{1}{\mathbf{1}+\cfrac{1}{\mathbf{3}+\cfrac{1}{\mathbf{1}}}}}$$

and we have an alternative CF-expansion

$$\frac{19}{14} \;=\; [1;2,1,3,1].$$

The inverse of $\frac{19}{14}$ can be written as

$$\frac{14}{19} \;=\; \mathbf{0}+\cfrac{1}{\mathbf{1}+\cfrac{5}{14}} \;=\; \mathbf{0}+\cfrac{1}{\mathbf{1}+\cfrac{1}{\mathbf{2}+\cfrac{4}{5}}} \;=\; \mathbf{0}+\cfrac{1}{\mathbf{1}+\cfrac{1}{\mathbf{2}+\cfrac{1}{\mathbf{1}+\cfrac{1}{\mathbf{4}}}}},$$

and finally

$$\frac{14}{19} \;=\; [0;1,2,1,4] \qquad \text{or} \qquad \frac{14}{19} \;=\; [0;1,2,1,3,1].$$

□

The notion of continued fractions can be extended to infinite CF-expansions, which correspond to irrational numbers. The infinite continued fraction is defined as the limit of a sequence of finite continued fractions:

$$[a_0;a_1,a_2,a_3,\ldots] \;=\; \lim_{n\to\infty}\,[a_0;a_1,a_2,a_3,\ldots,a_n].$$

Each rational number x_n given by $[a_0;a_1,a_2,a_3,\ldots,a_n]$ is a rational approximation of an irrational number x given by $[a_0;a_1,a_2,a_3,\ldots]$. The longer finite CF-expansions correspond to the better approximations of x.

To find a CF-expansion $[a_0;a_1,a_2,\ldots]$ of a given number x (rational or irrational) we can use a simple calculation:

$$y_0 \;=\; x, \quad a_n \;=\; \lfloor y_n \rfloor \quad \text{and} \quad y_{n+1} \;=\; \frac{1}{y_n - a_n} \quad \text{for} \quad (n \geq 0).$$

If x is rational then the computation stops for some n, when we have $y_n = a_n$, and we obtain the finite CF-expansion. For an irrational x the computation never stops and we achieve the infinite CF-expansion.

For a more comprehensive study of continued fractions and its applications see for example [3], [33], [37], [42] or [64].

The following fact indicates the relation between continued fractions, standard words and Christoffel words.

Fact 1.20 (See [11], [72])
Let $x,\ y \in \Sigma$ and $u \in \Sigma^*$. A word $w = xuy$ is a Christoffel word of the slope $\frac{p}{q}$ with CF-expansion $\frac{p}{q} = [\gamma_0; \gamma_1, \gamma_2, \dots, \gamma_n]$, if and only if uxy or uyx is a standard word given by a directive sequence $\gamma = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_n)$.

Example 1.21
Recall from Example 1.9 that

$$c_l \;=\; b \cdot ababaababababaababababaababababaababa \cdot a$$

is the lower Christoffel word of the slope $\frac{19}{14}$.

The two equivalent CF-expansion of $\frac{19}{14}$ (see Example 1.18) are

$$\frac{19}{14} = [1; 2, 1, 3, 1] \qquad \text{and} \qquad \frac{19}{14} = [1; 2, 1, 4].$$

Recall also from Example 1.2 and Example 1.7 that

$$\begin{aligned} \mathrm{Sw}(1,2,1,3,1) &= ababaababababaababababaababababaababa \cdot ab, \\ \mathrm{Sw}(1,2,1,4) &= ababaababababaababababaababababaababa \cdot ba. \end{aligned}$$

□

2

Algorithms related to subword graphs structure

This chapter is devoted to the investigation of the subword graphs structure of standard words in more detail than in previous works. The very special structure of those graphs (especially their compacted versions) leads to simple alternative graph-based proofs of several known facts and to special easy algorithms for computing some properties of Sturmian words: the number of subwords, the critical factorization point, lexicographically maximal suffixes, occurrences of subwords of a fixed length, and right special factors.

The *subword graph* is a classical data structure representing all subwords of a given word in a succinct manner. More precisely: the directed acyclic subword graph (the *dawg*, in short) of the word w is the minimal deterministic automaton (not necessarily complete) that accepts all suffixes of w. In this automaton we don't mark the accepting states and ignore the transitions leading to the dead-state (the rejecting state, in which the automaton loops). The most important property of the *dawg* is that its size is linear with respect to the length of the word w although the number of subwords of w can be quadratic. Subword graphs are designed to give a fast access to all factor of a string, and this is the reason why they have fairly large number of applications in text processing. See Figure 2.1 for an example of a subword graph and its compacted version.

For a more detailed study of this topic and online algorithms for constructing subword graph of a given word we refer the reader to [23] and [52].

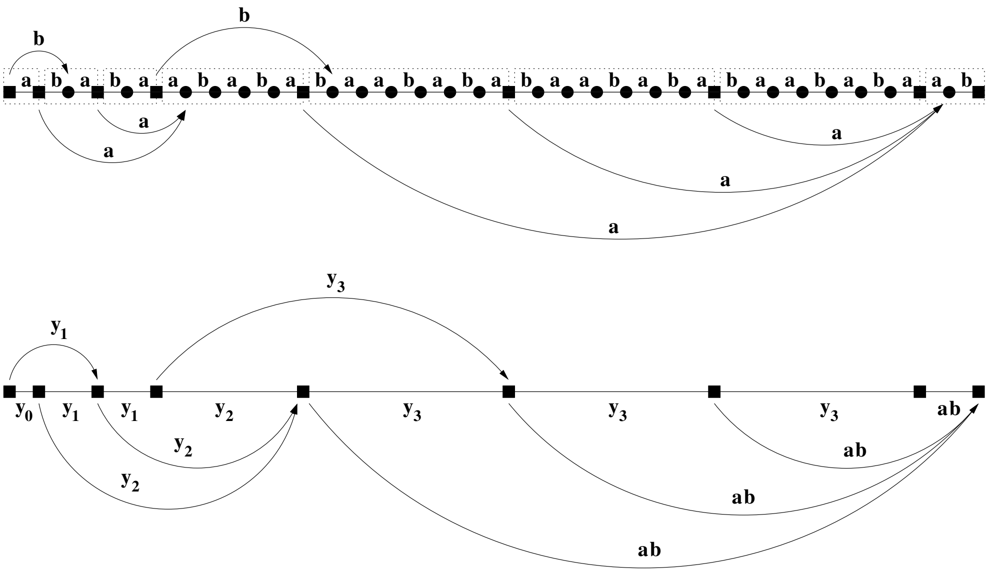
b
b
a b a b a a b a b a b a b a a b a b a b a b a a b a b a b a a b
a
a
a
a
a
y3
y1
y0 y1 y1 y2 y3 y3 y3 ab
y2
y2
ab
ab
ab

Figure 2.1: *The structure of the subword graph (*dawg*) and its compacted version (*cdawg*) of the word* Sw(1, 2, 1, 3, 1)*. The nodes of* dawg*, which are copied to* cdawg*, are marked with squares.*

2.1 Subword graphs of standard words

For the words w and u denote by $p_w(u)$ the shortest prefix of w having u as its suffix. The smallest number of states of the *dawg* of w is $|w| + 1$. We say that w is *simplistic* if the *dawg* of w has exactly $|w|+1$ nodes. The simplistic words have the simplest *dawgs*.

The following crucial fact describes the structure of the *dawg* of w.

Lemma 2.1 (See [70])
Let w be a standard Sturmian word. Then:

(1) *The word w is simplistic.*

(2) *The nodes of* dawg *of w can be identified with the prefixes of w.*

(3) *Each edge of the* dawg *of w is of the form $\alpha \xrightarrow{s} p_w(\alpha\, s)$, where $s \in \{a, b\}$ and α is a prefix of w.*

The *compacted subword graph* (the *cdawg*, in short) results from the subword graph by removing all nodes of out-degree one and replacing each chain by a single edge with the label representing the path label of this chain. We compact the chains of the *dawg* as much as possible but with the following restriction: for each node v all incoming edges of v have the same label (possibly long). This restriction implies that we can't fully compress the last chain going into the *sink* node. This chain is a concatenation of some basic subword y_k and the two-letter word u (ab or ba). We split this chain into two edges: the first labelled by y_k and the second labelled by u and going into the *sink* node. The internal nodes of *dawg* of out-degree two, which are copied to *cdawg*, are called the fork nodes. For the example of the *dawg* and *cdawg* see Figure 2.1.

Building blocks

We start by considering the relations between subwords, which are the *building blocks* of the subword graph of a standard word. Recall that for a word w the set of all nonempty factors of w is denoted by $F(w)$ and $\overline{w}$ denotes its reverse.

Let w be a standard word and x_i's are as in the equation (1.1). Subwords that are *building blocks* of the *dawg* and the *cdawg* of w are classified as:

a **special prefix** of w is a prefix z of w such that $za, zb \in F(w)$,

a **basic prefix** of w is a proper nonempty prefix of the type $x_k^j x_{k-1}$, where $0 \leq k \leq n$ and $0 \leq j \leq \gamma_k$.

a **basic subword** of w is a reverse of some x_k, denote $y_k = x_k^R$.

See Figure 2.2 for the *building blocks* structure of an example word.

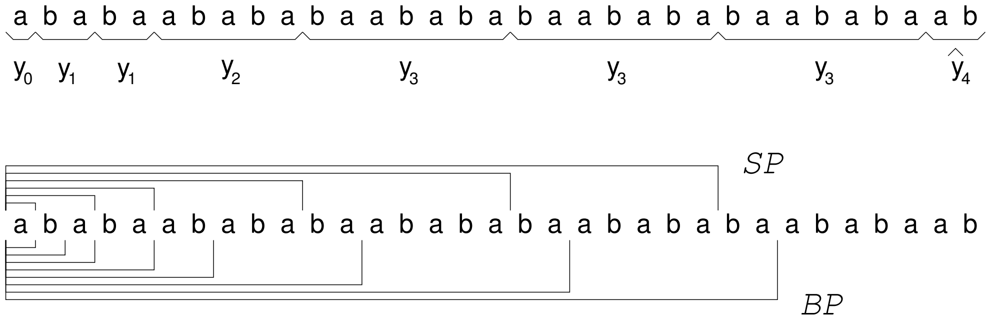
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
y_0 y_1 y_1 y_2 y_3 y_3 y_3 $\hat{y}_4$
SP
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
BP

Figure 2.2: *The structure of basic prefixes (BP), special prefixes (SP) and basic subwords of the word* Sw(1, 2, 1, 3, 1).

It follows directly from Lemma 2.1 that:

Fact 2.2
For a standard word w the nodes of the cdawg *of w with out-degree 2 (all except the last two nodes) correspond to the special prefixes of w.*

From the point of view of the compacted subword graphs of standard words special prefixes are the most important. On the other hand special prefixes are composed of basic subwords, and basic subwords are labels of the edges of the compacted subword graph. Hence special prefixes and basic subwords are the main *building blocks* of the standard words. The importance of the third type of factors – the basic prefixes – is an implication of the fact that special prefixes are *almost* the same as basic prefixes, and basic prefixes correspond more directly to the recurrences. They are the *link* between the directive sequence and special prefixes.

Denote by $BP(w)$ the set of basic prefixes of the word w and by $SP(w)$ the set of its special prefixes. Recall also that $\hat{w}$ is the prefix of w of the length 2 and exceptionally define $\hat{y}_0 = ab$. The following lemma expresses the relation between basic prefixes and special prefixes of standard words.

Lemma 2.3 (Building blocks)
Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a directive sequence and $x_{-1}, x_0, \ldots, x_{n+1}$ be the sequence of standard words given by the recurrence from the equation (1.1)*.*

(1) *For $i \geq 1$ we can represent the standard word x_i as*

$$x_i \ = \ y_0^{\gamma_0} y_1^{\gamma_1} \ldots y_{i-2}^{\gamma_{i-2}} y_{i-1}^{\gamma_{i-1}-1} \hat{y}_{i-1}.$$

(2) *Each special prefix z of the word x_n is of the the form $z = y_0^{\gamma_0} y_1^{\gamma_1} \ldots y_i^j$, where $0 \leq j \leq \gamma_i$ for $i < n-1$ and $0 \leq j \leq \gamma_i - 1$ for $i = n-1$.*

(3) *Each special prefix of x_n results by removing the last two letters from the corresponding basic prefix of x_n.*

Proof
We demonstrate each point separately.

Point (1).
Notice that for $i \geq 0$ we have $\hat{y}_i = \hat{y}_{i+2}$ and $y_{i+1} = y_{i-1}\, y_i^{\gamma_i}$.

First, we show by induction on i that

$$y_i = \hat{y}_i\, y_0^{\gamma_0}\, y_1^{\gamma_1} \ldots y_{i-1}^{\gamma_{i-1}-1}. \tag{2.4}$$

For $i = 1$ we have

$$y_1 = b\, a^{\gamma_0} = \hat{y}_1\, y_0^{\gamma_0 - 1}.$$

Assume that for $i \leq n$ the equation (2.4) is true. We have

$$\begin{aligned} y_{n+1} &= y_{n-1} \cdot y_n^{\gamma_n} \\ &= \left(\hat{y}_{n-1}\, y_0^{\gamma_0}\, y_1^{\gamma_1} \dots y_{n-3}^{\gamma_{n-3}}\, y_{n-2}^{\gamma_{n-2}-1}\right) \cdot \left(y_{n-2} \cdot y_{n-1}^{\gamma_{n-1}} \cdot y_n^{\gamma_n - 1}\right) \\ &= \hat{y}_{n+1}\, y_0^{\gamma_0}\, y_1^{\gamma_1} \dots y_{n-1}^{\gamma_{n-2}}\, y_{n-1}^{\gamma_{n-1}}\, y_n^{\gamma_n - 1}. \end{aligned}$$

Now we are ready to prove the equation from the point (1). We can do so by induction on i.

For $i = 1$ we have:

$$x_1 = x_0^{\gamma_0} x_{-1} = y_0^{\gamma_0 - 1} \hat{y}_0.$$

Assume now that for $i \leq n$ equation from the point (1) is true. We have:

$$\begin{aligned} x_{n+1} &= x_n^{\gamma_n}\, x_{n-1} \\ &= \left(y_0^{\gamma_0} \dots y_{n-2}^{\gamma_{n-2}}\, y_{n-1}^{\gamma_{n-1}-1}\, \hat{y}_{n-1}\right)^{\gamma_n} \cdot y_0^{\gamma_0} \dots y_{n-2}^{\gamma_{n-2}-1}\, \hat{y}_{n-2} \\ &= y_0^{\gamma_0} \dots y_{n-1}^{\gamma_{n-1}-1} \cdot \Big(\underbrace{\hat{y}_{n-1}\, y_0^{\gamma_0} \dots y_{n-2}^{\gamma_{n-2}-1}}_{y_{n-1}}\Big) \cdot \\ &\qquad \cdot \Bigg[\overbrace{\left(y_{n-2}\, y_{n-1}^{\gamma_{n-1}-1}\right) \cdot \Big(\underbrace{\hat{y}_{n-1}\, y_0^{\gamma_0} \dots y_{n-2}^{\gamma_{n-2}-1}}_{y_{n-1}}\Big)}^{y_n}\Bigg]^{\gamma_n - 1} \cdot \hat{y}_n \\ &= y_0^{\gamma_0} \dots y_{n-1}^{\gamma_{n-1}}\, y_n^{\gamma_n - 1}\, \hat{y}_n. \end{aligned}$$

Point (2).
First recall from Remark 1.12 that the standard word w without the last two letters (denoted by $\overline{w}$) is a palindrome.

Let $x_n = \mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_{n-1})$ and $z = y_0^{\gamma_0} y_1^{\gamma_1} \dots y_i^j$ be a prefix of x_n. Due to the point (1) we have $0 \leq j \leq \gamma_i$ for $i < n-1$ and $0 \leq j \leq \gamma_i - 1$ for $i = n-1$.

We can deduce that $z = \overline{v}$ for a word $v = \mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_{i-1}, j+1)$, hence it is a palindrome. Both $\overline{x_n}$ and z are palindromes, moreover z is a prefix of $\overline{x_n}$, therefore z is also a suffix of $\overline{x_n}$.

Assume that $i < n-1$ and i is odd. The case for even i is similar.

If $0 \leq j < \gamma_i$, then z is a prefix of x_{i+2} and zb is also a prefix of x_{i+2} (the first letter of y_i is b). We have $x_{i+2} = \overline{x_{i+2}} \cdot ab$ and z is a suffix of $\overline{x_{i+2}}$, hence za is also a subword of x_{i+2}.

If $j = \gamma_i$, then z is a prefix of x_{i+3} and za is also a prefix of x_{i+3} (the first letter of y_{i+1} is a). We have $x_{i+3} = \overline{x_{i+3}} \cdot ba$ and z is a suffix of $\overline{x_{i+3}}$, hence zb is also a subword of x_{i+3}.

Now assume that $i = n - 1$. For $0 \leq j < \gamma_{n-1}$ the proof is similar to the previous case and, due to the point (1), it is obvious that $j < \gamma_{n-1}$.

Point (3).
Notice that for $i \geq 0$ we have $\hat{y}_i = \hat{y}_{i+2}$ and $y_{i+1} = y_{i-1}\, y_i^{\gamma_i}$.

The point (1) implies that the basic prefix $x_k^j x_{k-1}$ equals:

$$\begin{aligned} x_k^j x_{k-1} &= \left(y_0^{\gamma_0} \ldots y_{k-2}^{\gamma_{k-2}}\, y_{k-1}^{\gamma_{k-1}-1}\ \hat{y}_{k-1}\right)^j \cdot y_0^{\gamma_0} \ldots y_{k-3}^{\gamma_{k-3}}\, y_{k-2}^{\gamma_{k-2}-1}\ \hat{y}_{k-2} \\ &= y_0^{\gamma_0} \ldots y_{k-1}^{\gamma_{k-1}}\, y_k^{j-1}\ \hat{y}_k. \end{aligned}$$

From the point (2) we conclude that the basic prefix $x_k^j x_{k-1}$ with the last two letters removed $(\hat{y}_k)$ is a special prefix.

□

Example 2.5
For $\mathrm{Sw}(1,2,1,3,1) = ababaababaababababaababaabababaab$ we have:

the set of the basic prefixes :

$BP = \{x_0,\ x_1,\ x_1\, x_0,\ x_2,\ x_3,\ x_3\, x_2,\ x_3^2\, x_2,\ x_4\}$,

where: $x_0 = a,\ x_1 = ab,\ x_2 = ababa,\ x_3 = ababaab,$

the set of the special prefixes :

$SP = \{y_0,\ y_0\, y_1,\ y_0\, y_1^2,\ y_0\, y_1^2\, y_2,\ y_0\, y_1^2\, y_2\, y_3,\ y_0\, y_1^2\, y_2\, y_3^2\}$,

where: $y_0 = a,\ y_1 = ba,\ y_2 = ababa,\ y_3 = baababa.$

and the decomposition of the word :

$$\begin{aligned} \mathrm{Sw}(1,2,1,3,1) &= a\ \ ba\ \ ba\ \ ababa\ \ baababa\ \ baababa\ \ baababa\ \ ab \\ &= y_0\, y_1^2\, y_2\, y_3^3\, \hat{y}_4. \end{aligned}$$

□

The structure of the compacted *dawg*

The regularity of the structure of compacted subword graphs has been discovered in [26]. The main point is that the *cdawg* is exponentially smaller than the *dawg* for the standard word w. The following fact is an implication of the results of [26], Lemma 2.1, Lemma 2.3 and our terminology.

Fact 2.6

Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word.

(1) *The labels of the edges in the* cdawg *of w are basic subwords of w.*

(2) *The compacted subword graph of w has the structure as follows:*

- *each node corresponding to special prefix $y_0^{\gamma_0}\, y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}}\, y_i^k$, for $0 \le k < \gamma_i$, has two outgoing edges:*
  - $y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}} y_i^k \xrightarrow{\quad y_i \quad} y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}} y_i^{k+1}$
  - $y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}} y_i^k \xrightarrow{\quad y_{i+1} \quad} y_0^{\gamma_0} y_1^{\gamma_1} \cdots y_{i-1}^{\gamma_{i-1}} y_i^k y_{i+1}$
- *each edge leading to the sink node has label $\hat{y}_n$ (ab or ba),*
- *the last but one node doesn't correspond to special prefix and has out-degree 1*

(see Figure 2.3).

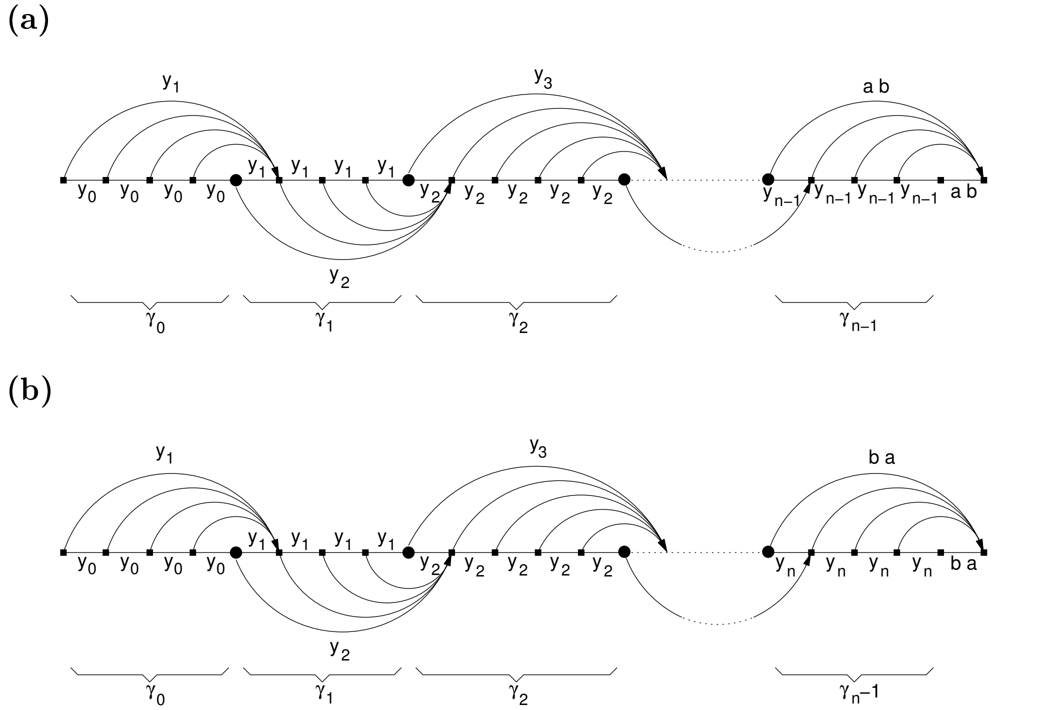

Figure 2.3: *The compacted subword graphs of the standard words (a):* $\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n)$ *and (b):* $\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_n - 1, 1)$ *are isomorphic (in the sense of the graph structure).*

Remark 2.7
Recall that standard words $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ and $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n - 1, 1)$ are the same up to the last two letters and observe that due to Lemma 2.3 we have

$$\begin{aligned} \mathrm{Sw}(\gamma_0, \ldots, \gamma_n) &= y_0^{\gamma_0}\, y_1^{\gamma_1} \ldots y_{n-q}^{\gamma_{n-1}}\, y_n^{\gamma_n - 1}\, \hat{y}_n, \\ \mathrm{Sw}(\gamma_0, \ldots, \gamma_n - 1, 1) &= y_0^{\gamma_0}\, y_1^{\gamma_1} \ldots y_{n-q}^{\gamma_{n-1}}\, y_n^{\gamma_n - 1}\, \hat{y}_{n+1}, \end{aligned}$$

since $y_{n+1}^0 = \varepsilon$. Therefore, Fact 2.6 implies that the compacted subword graphs of those words are isomorphic in the sense of graph structure, see Figure 2.3 for comparison.

2.2 The number of subwords

Recall that $F(w)$ denotes the set of factors of the word w and F_n the n-th Fibonacci word. The number of distinct subwords in F_n is (see [67]):

$$\big|F(F_{n+1})\big| = |F_n| \cdot |F_{n-1}| + 2 \cdot |F_n| - 1.$$

Surprisingly, we have similar result for the standard words.

Fact 2.8
Let $x_{n+1} = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard word and let $\gamma_n = 1$. The number of distinct factors of x_{n+1} is given by the formula

$$\big|F(x_{n+1})\big| = |x_n| \cdot |x_{n-1}| + 2 \cdot |x_n| - 1.$$

Proof
Let G be the compacted subword graph of the word x_{n+1}, v_0 the source node of G and $t_k = |x_k|$. In the graph G we define $\mathit{mult}(v)$ – the multiplicity of the vertex v – as the number of paths $v_0 \rightsquigarrow v$ and the weight of an edge as the length of the corresponding label-string of this edge. For a vertex v denote by $\mathit{edges}(v)$ the sum of the weights of all edges outgoing from v. See Figure 2.4 for edge-lengths and node-multiplicities structure in the *cdawg* of the example word.

Claim 2.9
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$. Then:

$$\big|F(w)\big| = \sum_{v \in G} \mathit{mult}(v) \cdot \mathit{edges}(v). \tag{2.10}$$

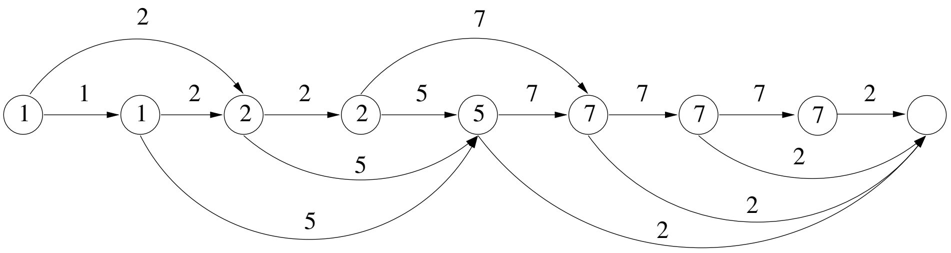

Figure 2.4: *The structure of the edge-lengths and the multiplicities of the nodes in the* cdawg *of* Sw(1, 2, 1, 3, 1)*. According to Fact 2.8 (and to the graph above) there are* $|x_4| \cdot |x_3| + 2 \cdot |x_4| - 1 = 26 \cdot 7 + 2 \cdot 26 - 1 = 233$ *subwords in our example word.*

We partition the set of edges of the graph G into chunks G_0, G_1,. . . ,G_{n-1}. The first chunk G_0 consists of the first γ_0 consecutive vertices (starting from the source node v_0), the second chunk G_1 contains the next γ_1 vertices, and so on. Only the last chunk G_{n-1} slightly differs.

The contribution of k-th internal chunk in the sum from the equation (2.10) is

$$\big(t_{k-1} + (\gamma_k - 1)t_k\big) \cdot (t_k + t_{k+1}) \;=\; t_{k+1}^2 - t_k^2,$$

where $t_{-1} = 1$, see Figure 2.5 for details.

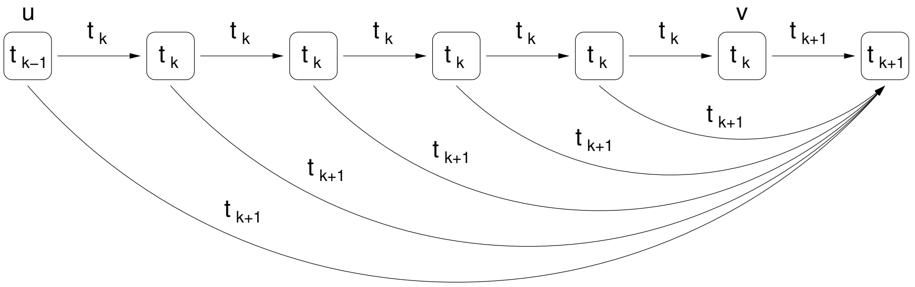

Figure 2.5: *The k-th internal chunk* G_k *of the subword graph consists of* γ_k *nodes from u to v (excluding u), and their outgoing edges. The multiplicity (the number of paths leading from* v_0*) of each node is written within the box corresponding to the node. The weight of the edges are the lengths of corresponding words in the* cdawg.

The contribution of the last chunk is in the sum from the equation (2.10) is

$$(t_{n-1} + 2)(t_n - t_{n-1}) + 2t_{n-1},$$

see Figure 2.6 for details.

Altogether, the number of subwords is

$$\sum_{k=0}^{n-2} \left(t_{k+1}^2 - t_k^2\right) + (t_{n-1}+2)(t_n - t_{n-1}) + 2 \cdot t_{n-1} \;=\; t_n \cdot t_{n-1} + 2 \cdot t_n - 1.$$

This completes the proof, since by the definition $t_k = |x_k|$.

□

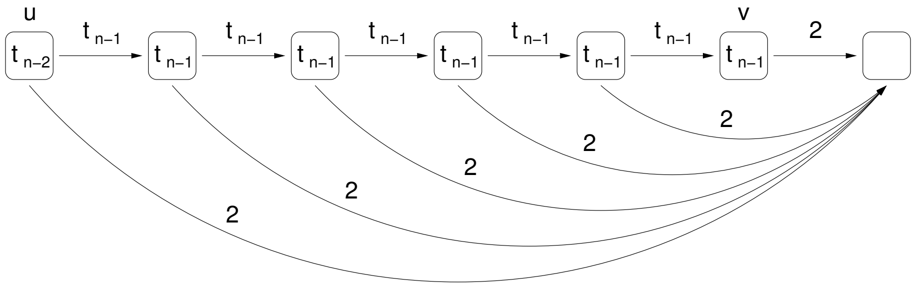

Figure 2.6: *The final chunk G_{n-1} of the subword graph consists of γ_{n-1} nodes from u to v, and their outgoing edges.*

The result from Fact 2.8 can be easily extended to the case $\gamma_n > 1$.

Fact 2.11

Let $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ *be a standard word and let* $\gamma_n > 1$. *Then*

$$\left|F\big(\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)\big)\right| \;=\; \left|F\big(\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n - 1, 1)\big)\right|.$$

Proof

Recall from Remark 2.7 that the compacted subword graphs of the words $\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)$ and $\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n - 1, 1)$ are isomorphic in the sense of the graph structure (see Figure 2.3).

For a word w the set $F(w)$ of its factors corresponds to its subword graph structure. Therefore, both sets

$$F\big(\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)\big) \qquad \text{and} \qquad F\big(\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n - 1, 1)\big)$$

have the same cardinality and the thesis holds.

□

2.3 The structure of occurrences of subwords

In this section we are interested in the structure of the first occurrences of the factors of a given length. One type of such subwords is particularly interesting – the right special factors.

Right special factors

A *right special factor* of the word $w \in \{a, b\}^*$ is any word x such that both xa and xb are subwords of w.

For each $k > 0$ there is at most one right special factor of the length k of a given standard word. Moreover, every right special factor of a standard word is either its special prefix or a suffix of some its special prefix.

Fact 2.12
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. Then:

(1) *For a given $k > 0$ the right special factor of w of the length k has the grammar-based representation of the size $O\big(|\gamma|\big)$.*

(2) *The compressed representation of the right special factor of w of the length k can be computed in time $O\big(|\gamma|\big)$.*

Proof
Let π be a path in the compacted subword graph of the word w. Define its value as the word created by concatenation of the labels of the edges in π.

Let v be a fork node in the *cdawg* of w (whichever except the last two nodes), π be a path leading to v from some other node v_1, and z_π be the value of π. It is clear that z_π is a subword of w.

The node v has two outgoing edges: one with the label starting with the letter a and the second with the label starting with the letter b. Consequently $z_\pi a$ and $z_\pi b$ are also subwords of w and therefore z_π is a right special factor of the word w.

Observe that the value of every path in the *cdawg* of w, which ends in some fork node v, is the suffix of the value of the longest path from the root to v. Moreover, the value of this longest path from the root to v is the prefix of w, hence the special prefix of w. This implies that every right special factor of w is a suffix of some its special prefix.

Every right special factor of the word w is the concatenation of some its basic subwords. It follows easily from Lemma 2.3 that every right special factor of w has the grammar-based representation of the size $O(|\gamma|)$, which can be computed in linear time with respect to the length of the directive sequence γ.

□

Example 2.13

Let $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ be a standard word. We have then

$$w = ababaabababaababababaababababaababaab$$

and

$$y_0 = a, \qquad y_1 = ba, \qquad y_2 = ababa, \qquad y_3 = baababa.$$

The right special factors of the word w with their lengths are (the special prefixes are marked in bold):

1	$\boldsymbol{y_0}$			11	$y_1^2y_3$	18	$y_1^2y_3^2$
				12	y_2y_3	19	$y_2y_3^2$
2	y_1	6	y_0y_2	13	$y_0y_2y_3$	20	$y_0y_2y_3^2$
3	$\boldsymbol{y_0y_1}$	7	y_1y_2	14	$y_1y_2y_3$	21	$y_1y_2y_3^2$
		8	$y_0y_1y_2$	15	$y_0y_1y_2y_3$	22	$y_0y_1y_2y_3^2$
4	y_1^2	9	$y_1^2y_2$	16	$y_1^2y_2y_3$	23	$y_1^2y_2y_3^2$
5	$\boldsymbol{y_0y_1^2}$	10	$\boldsymbol{y_0y_1^2y_2}$	17	$\boldsymbol{y_0y_1^2y_2y_3}$	24	$\boldsymbol{y_0y_1^2y_2y_3^2}$

Compare with the graph on Figure 2.1.

□

The structure of the *dawg* of $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ implies the following fact.

Fact 2.14

Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. Every factor v of w has the unique decomposition into subwords

$$v = y_{i_1}\, y_{i_2} \cdots y_{i_k}\, \tilde{y}_{i_{k+1}},$$

where $i_1 \in \{0, 1\}$, $i_{k+1} \in \{i_k, i_k + 1, i_k + 2\}$ and $\tilde{y}_{i_{k+1}}$ is a prefix (possibly the whole word) of $y_{i_{k+1}}$.

Remark 2.15

As a direct consequence of Fact 2.14 we obtain the easy linear algorithm for checking if v is a subword of $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$, since the next factor of the above decomposition is determined by the next scanning letter.

Final positions of the first occurrences of subwords

For the words w and u we define *first-fin*(u, w) as the position of the last symbol of u in its first occurrence in the word w.

For $k \geq 1$ we define also the set

$$FIN(k, w) \;=\; \{\textit{first-fin}(u, w) \;:\; u \text{ is a subword of } w \text{ of the length } k\;\}.$$

See Figure 2.7 for an example.

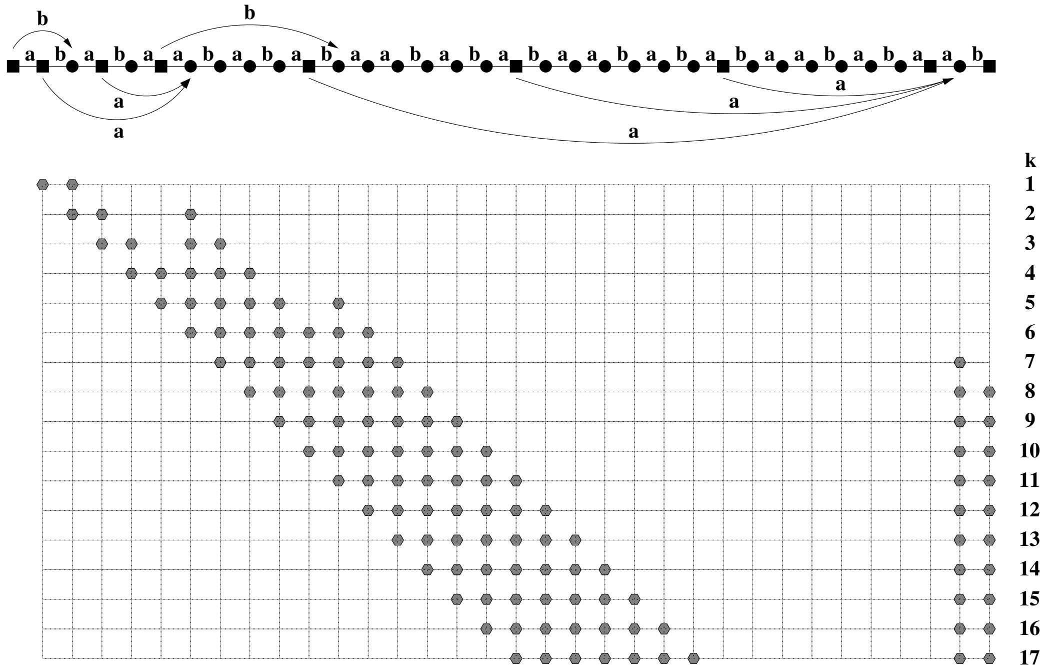

Figure 2.7: *The subword graph of the word* $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ *and the structure of the set* $FIN(k, w)$.

The following fact describes the structure of the set $FIN(k, w)$ for the standard word w.

Fact 2.16
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *be a standard Sturmian word. Then:*

(1) *The set* $FIN(k, w)$ *consists of a single interval or two disjoint intervals.*

(2) *For a given* $k \geq 1$ *we can compute the intervals representing* $FIN(k, w)$ *in linear time with respect to the size of the directive sequence.*

Proof
The structure of the set $FIN(k,w)$ easily follows from the way the paths of the length $k-1$ in the subword graph of the word w are extended into the paths of the length k. Only the fork nodes $i \in FIN(k-1,w)$ generate two elements of the set $FIN(k,w)$. Each other node $i \in FIN(k-1,w)$ generates a single element $i+1$ in $FIN(k,w)$ (see Figure 2.7).

It is clear that the set $FIN(k+1,w)$ results from $FIN(k,w)$ by shifting each position by one to the right and adding an extra position for each fork node. Hence the thesis follows from the structure of subword graphs of a standard Sturmian words.

□

2.4 Critical factorization and maximal suffixes

The *minimal local period* in a word w at the position k is the positive integer p (minimal having this property) such that $w[i-p] = w[i]$ for every $k < i \leq k+p$, whenever $w[i]$ and $w[i-p]$ are defined. If either $w[i]$ or $w[i-p]$ is not defined, we consider that the equality needed is true. In other words the *minimal local period* in a word w at the position k is the length of the shortest subword of w that ends at the position k and repeats directly after this position.

The *critical factorization point* in a word w is the position k in w, for which the minimal local period at k equals the (global) minimal period of w. We refer the reader to [23] for the more detailed definition.

Example 2.17
Let w = Sw(1, 2, 1, 3, 1) = *ababaababab aababab aabababaababaab*.
Minimal local periods of w are as follows:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	······
	a	b	a	b	a	a	b	a	b	a	b	a	a	b	a	b	a	······
$p(i)$	1	2	2	2	5	1	7	2	2	2	2	7	1	7	2	2	2	2······

i	······	18	19	20	21	22	23	24	25		26	27	28	29	30	31	32	33
	······	b	a	a	b	a	b	a	b		a	a	b	a	b	a	a	b
$p(i)$	······	2	7	1	7	2	2	2	4	**33**	1	5	2	2	5	1	3	1

where i denotes the position in the word w and $p(i)$ – the minimal local period at the position i.

The critical factorization point in w is at the position $i = 25$, where the minimal local period equals the length of w.

□

The following nontrivial fact has been shown by Crochemore and Perrin in [21].

Fact 2.18 (See [21])
The critical factorization point of the word w is given as the starting position of a lexicographically maximal suffix of w, maximized over two reversed orders of the alphabet.

The above fact has an interesting interpretation in terms of the subword graphs structure of standard words.

Let w be a standard word. In the subword graph of w we define the path $\pi_a(w)$ leading from the root to the sink, in which we use the letter a whenever we have a choice (in every fork node). The path $\pi_b(w)$ is defined in a similar way. The only difference is that the letter a is replaced by the letter b. Both $\pi_a(w)$ and $\pi_b(w)$ can be also defined in the compacted subword graph of w. In this case, in every fork node we choose the edge with the label starting with the letter a or b, respectively. The length of the path π, denoted by $|\pi|$, is defined as the length of the word given by π.

It is easily seen that the lexicographically maximal suffix of the word w with respect to the letter ordering $''a < b''$ is given by the path $\pi_b(w)$ and the lexicographically maximal suffix of w with respect to the letter ordering $''a > b''$ is given by the path $\pi_a(w)$.

The structure of the paths $\pi_a(w)$ and $\pi_b(w)$ is given by the following lemma.

Lemma 2.19
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word and $\pi_a(w)$, $\pi_b(w)$ be defined as above. Then:

$$\begin{aligned}\pi_a(w) &= y_0^{\gamma_0}\, y_2^{\gamma_2} \cdots y_k^{\alpha}\, \hat{y}_n, \\ \pi_b(w) &= y_1^{\gamma_1} y_3^{\gamma_3} \cdots y_l^{\beta}\, \hat{y}_n,\end{aligned}$$

where

$$\begin{cases} k = n, \quad \alpha = \gamma_n - 1, \quad l = n-1, \quad \beta = \gamma_{n-1} & \textit{for even } n \\ k = n-1, \quad \alpha = \gamma_{n-1}, \quad l = n, \quad \beta = \gamma_n - 1 & \textit{for odd } n \end{cases}.$$

Proof
It follows by the definition of basic subwords that y_i starts with the letter a for even i and with the letter b for odd i.

We are constructing the path $\pi_a(w)$ in the *cdawg* of w by choosing the edge with the label starting with the letter a whenever it is possible. The structure of the *cdawg* of a standard Sturmian word (see Figure 2.3) implies that every fork node has two outgoing edges: one with the label y_{2i} (starting with the letter a) and the second with the label y_{2i+1} (starting with the letter b).

In order to construct $\pi_a(w)$ we have to choose γ_0 times the edge with the label y_0, then γ_2 times the edge with the label y_2, and so on. Depending on the parity of n we end either with y_n taken $\gamma_n - 1$ times (for even n) or with y_{n-1} taken γ_{n-1} times (for odd n). Finally, by Lemma 2.3, it suffices to add $\hat{y}_n$ – the last two letters of w (ab or ba respectively).

The structure of the path $\pi_b(w)$ could be proved by a similar reasoning.

□

The following fact is a conclusion from Fact 2.18 and the construction of the paths $\pi_a(w)$ and $\pi_b(w)$.

Fact 2.20
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)$ be a standard Sturmian word. Then:

(1) *The critical factorization point of w is at the position*

$$k = |w| - \min\{\ |\pi_a(w)|,\ |\pi_b(w)|\ \}.$$

(2) *The critical factorization point of w can be computed in linear time with respect to the size of the directive sequence.*

Proof
Recall that $\pi_a(w)$ and $\pi_b(w)$ correspond to lexicographically maximal suffixes of w with respect to the letter orderings $''a > b''$ and $''a < b''$ respectively. The thesis is a direct consequence of Fact 2.18 and Fact 2.19.

□

Example 2.21
Let $w = \mathrm{Sw}(1, 2, 1, 3, 1) = ababaababababaababababaababababaababaab$.
See Figure 2.1 for its subword graph structure. We have:

$$\begin{aligned} \pi_a(w) &= y_0 y_2 ab = a\ ababa\ ab, \\ \pi_b(w) &= y_1^2 y_3^3 ab = ba\ ba\ baababa\ baababa\ baababa\ ab. \end{aligned}$$

Hence the position

$$i = |w| - |y_0\, y_2\, ab| \;=\; 33 - 8 \;=\; 25$$

is the critical factorization point of w (compare with Example 2.17).

□

Similar computations have been given in [35] and [67] for Fibonacci words.

The paths $\pi_a(w)$ and $\pi_b(w)$ defined above have a very regular structure, consequently the words represented by them are well compressible. This results with the following fact.

Fact 2.22
Let $w = \mathrm{Sw}(\gamma)$ be a standard Sturmian word. Then:

(1) *The lexicographically maximal suffix of w has a grammar-based representation of the size $O\big(|\gamma|\big)$.*

(2) *The compressed representation of the lexicographically maximal suffix of w can be computed in time $O\big(|\gamma|\big)$.*

Proof
Recall that the lexicographically maximal suffix of a standard Sturmian word w is given either by the path $\pi_a(w)$ (for the letter ordering $''b < a''$) or by the path $\pi_b(w)$ (for the letter ordering $''a < b''$). The thesis follows directly from the construction of the paths $\pi_a(w)$ and $\pi_b(w)$ (see Lemma 2.19) and the structure of the subword graph of w (see Fact 2.6).

□

3
Maximal repetitions in standard words

A *run* (a *maximal repetition*) is a non-extendable (with the same period) periodic segment in a string, in which the period repeats at least twice. Runs are important in combinatorics on words and many practical applications: data compression, computational biology, pattern-matching and so on. The structure of repetitions is almost completely understood for the class of Fibonacci words, see [48], [67], [40]. In this chapter we investigate the structure of runs in class $\mathcal{S}$ of standard Sturmian words and give the exact formula and the tight asymptotic bound for the number of maximal repetitions.

We continue here the work of [29], where it was shown how to compute the number of runs for block-complete Sturmian words (not all standard Sturmian words have this property) in linear time with respect to the size of the whole word. We show the algorithm, which computes the number of runs in any standard word in linear time with respect to the size of its compressed representation – the directive sequence – hence in logarithmic time with respect to the length of the word.

Recall that a number i is a period of the word w if $w[j] = w[i+j]$ for all i with $i + j \le |w|$. The minimal period of w will be denoted by $period(w)$. We say that a word w is periodic if $period(w) \le \frac{|w|}{2}$. A word w is said to be *primitive* if w is not of the form z^k, where z is a finite word and $k \ge 2$ is a natural number.

A *maximal repetition* (a *run*, in short) in a word w is an interval $\alpha = [i..j]$ such that $w[i..j] = u^k v$ $(k \ge 2)$ 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 $|u|$).

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 creates no ambiguity we identify sometimes 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 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.

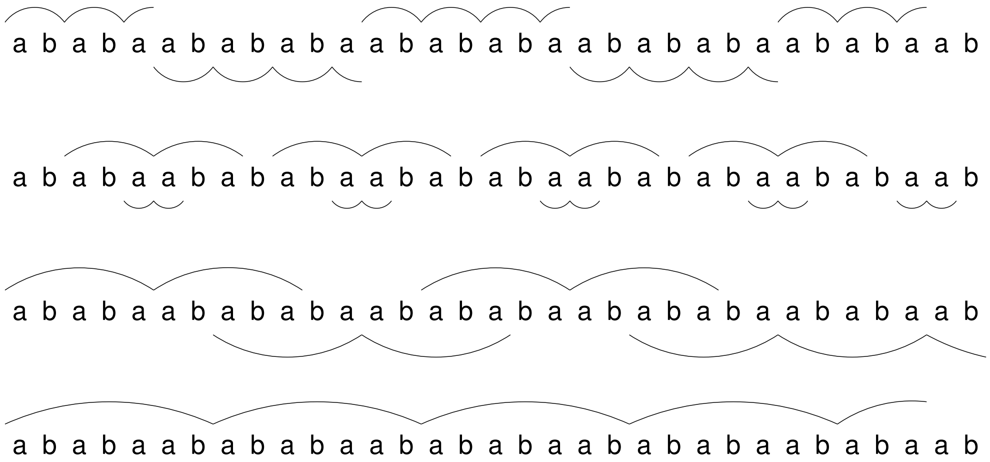
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b

Figure 3.1: *The structure of runs in the word* Sw$(1,2,1,3,1)$. *There are 19 runs including:* 10 short *runs (periods* 1 *and* 2*),* 8 medium *runs (periods* 3 *and* 5*) and* 1 large *run (period* 7*).*

Example 3.1

Let $w = ababaabababaababababaabababaababaab$.

There are 5 runs with the period $|a|$:

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

5 runs with the period $|ab|$:

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

4 runs with the period $|aba|$:

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

4 runs with the period $|ababa|$:

$$\begin{aligned} w[1..10] &= (ababa)^2, & w[8..17] &= (ababa)^2, \\ w[15..24] &= (ababa)^2, & w[22..33] &= (ababa)^2ab, \end{aligned}$$

and 1 run with the period $|ababaab|$:

$$w[1..31] \;=\; (ababaab)^4aba.$$

All together we have 19 runs, see Figure 3.1 for comparison. □

Denote by $\rho(w)$ the number of runs in the word w and by $\rho(n)$ the maximal number of runs in the words of length n. The most interesting and open conjecture about maximal repetitions is:

$$\rho(n) \;<\; n.$$

In 1999 Kolpakov and Kucherov (see [47]) showed that the number $\rho(w)$ of runs in a string w is $O\big(|w|\big)$, but the exact multiplicative constant coefficient is still unknown. The best known results related to the value of $\rho(n)$ are

$$0.944542\,n \;\leq\; \rho(n) \leq\; 1.048\,n.$$

The upper bound is by [18], [20] and the lower bound is by [30], [31], [49]. See Table 3.1 for the maximal number of runs and the repetition ratio in the words over a binary alphabet for the small values of n.

3.1 Morphic reduction of standard words

In this section we introduce a *reduction sequence* that allows us to reduce the computation of runs in the word $\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)$ to the computation in $\mathrm{Sw}(\gamma_1, \gamma_2, \dots, \gamma_n)$. The relation between the words $\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)$ and $\mathrm{Sw}(\gamma_1, \gamma_2, \dots, \gamma_n)$ is described in terms of the morphism transforming one of them to the other. Our concept is similar to the one shown in [29], but is closely related to the combinatorial structure of standard words.

Recall form the section 1.3 that for a directive sequence $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_n)$ we define the sequence of morphisms $\{h_i\}_{i=0}^{n}$, where:

$$h_i : \begin{cases} a & \longrightarrow \quad a^{\gamma_i}b \\ b & \longrightarrow \quad a \end{cases} \qquad \text{for } 0 \leq i \leq n. \tag{3.2}$$

n	$\rho(n)$	$\rho(n)/n$	Example word
10	6	0.6	aabaabbaab
11	7	0.636364	aabaabbaabb
12	8	0.666667	aabaabbaabaa
13	8	0.615385	aaabaabbaabaa
14	10	0,714286	aabaabbaabaabb
15	10	0.666667	aaabaabbaabaabb
16	11	0.6875	aabaabbaabaabbaa
17	12	0.705882	aabaababbabaababb
18	13	0.722222	aabaabbaabaabbaabb
19	14	0.736842	aabaabbaabaabbaabaa
20	15	0.75	aababaababbabaababaa
21	15	0.714286	aaababaababbabaababaa
22	16	0.727273	aabaababaababbabaababb
23	17	0.73913	aabaababaababbabaababaa
24	18	0.75	aabaabbaabaabbabbaabbabb
25	19	0.76	aabaabbaabaaabaabbaabaabb
26	20	0.769231	aababaababbabaababaababbab
27	21	0.777778	aabaababaababbabaababaababb
28	22	0.785714	aababaababbabaababaababbabaa
29	23	0.793103	aababaababbabaababaababbababb
30	24	0.8	aababbabaababbababbabaababbaba
31	25	0.806457	aababaababbabaababaababbabaabab

Table 3.1: *The maximal number of runs and the repetition ratio for the binary words of the given length.*

Due to Lemma 1.15, we have

$$\mathrm{Sw}(\gamma_i, \gamma_{i+1}, \ldots, \gamma_n) \;=\; h_i\big(\mathrm{Sw}(\gamma_{i+1}, \gamma_{i+2}, \ldots, \gamma_n)\big).$$

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

Recall that $|w|_a$ denotes the number of occurrences of the letters a in the word w. In the rest of this chapter we use the following notation:

$$\begin{aligned} N_\gamma(k) &= \big|S(\gamma_k, \gamma_{k+1}, \ldots, \gamma_n)\big|_a, \\ M_\gamma(k) &= \big|S(\gamma_k, \gamma_{k+1}, \ldots, \gamma_n)\big|_b, \end{aligned} \tag{3.3}$$

which enables us to simplify the formulas for the number of runs.

Remark 3.4

As a direct conclusion from the above definition, the equation (1.1) and the equation (3.2) we have that the numbers $N_\gamma(k)$ and $M_\gamma(k)$ satisfy:

$$\begin{aligned} N_\gamma(k) &= \gamma_k \, N_\gamma(k+1) + N_\gamma(k+2), \\ M_\gamma(k) &= N_\gamma(k+1). \end{aligned} \tag{3.5}$$

Remark 3.6

Recall form the section 1.1 that the n-th Fibonacci word is defined as

$$F_n \;=\; \mathrm{Sw}(1, 1, \ldots, 1).$$

Observe that the number of the letters a in the word F_n equals the length of the word F_{n-1}, and therefore

$$N_\gamma(k) = |F_{n-k-1}| \qquad \text{and} \qquad M_\gamma(k) = |F_{n-k-2}|.$$

Example 3.7

Let $\gamma \;=\; (1, 2, 1, 3, 1)$ be a directive sequence. We have then

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

□

The following lemma enables us to express the length of any standard word in terms of the numbers $N_\gamma(k)$ and $M_\gamma(k)$.

Lemma 3.8

Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$, $A = N_\gamma(2)$ *and* $B = N_\gamma(3)$. *Then*

$$|w| \;=\; \big((\gamma_0 + 1)\, \gamma_1 + 1\big)\, A \;+\; (\gamma_0 + 1)\, B.$$

Proof

By the definition of $N_\gamma(k)$ and $M_\gamma(k)$ we have

$$|w| \;=\; N_\gamma(0) + M_\gamma(0) \qquad \text{and} \qquad M_\gamma(0) \;=\; N_\gamma(1).$$

By repeated application of the formulas from the equation (3.3) we obtain:

$$\begin{aligned} |w| &= N_\gamma(0) + N_\gamma(1) \\ &= (\gamma_0 + 1)\, N_\gamma(1) + N_\gamma(2) \\ &= \big((\gamma_0 + 1)\, \gamma_1 + 1\big)\, N_\gamma(2) + (\gamma_0 + 1)\, N_\gamma(3) \\ &= \big((\gamma_0 + 1)\, \gamma_1 + 1\big)\, A + (\gamma_0 + 1)\, B \end{aligned}$$

and the proof is complete. □

3.2 Counting runs and repetition ratios

In this section we present the formula for the number of runs in standard Sturmian words and investigate its asymptotic behaviour. The proof of its correctness is the aim of the section 3.3.

We begin with the definition of some useful zero-one functions for testing the parity of a nonnegative integer i:

$$\mathit{even}(i) = \begin{cases} 1 & \text{for even } i \\ 0 & \text{for odd } i \end{cases} \quad \text{and} \quad \mathit{odd}(i) = \begin{cases} 1 & \text{for odd } i \\ 0 & \text{for even } i \end{cases}$$

and for testing if a positive integer i equals 1:

$$\mathit{unary}(i) = \begin{cases} 1 & \text{for } i = 1 \\ 0 & \text{for } i > 1 \end{cases} .$$

These functions will be used to simplify the formula for the number of runs in standard words.

Theorem 3.9 (Formula for the number of runs)
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $n \geq 3$. The number of runs in a standard word $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ is given by the following formula:

$$\rho(w) = \begin{cases} 2A + 2B + \Delta(\gamma) - 1 & \textit{for } \gamma_0 = \gamma_1 = 1 \\ (\gamma_1 + 2)A + B + \Delta(\gamma) - \mathit{odd}(n) & \textit{for } \gamma_0 = 1;\ \gamma_1 > 1 \\ 2A + 3B + \Delta(\gamma) - \mathit{even}(n) & \textit{for } \gamma_0 > 1;\ \gamma_1 = 1 \\ (2\gamma_1 + 1)A + 2B + \Delta(\gamma) & \textit{for } \gamma_0 > 1;\ \gamma_1 > 1 \end{cases} , \quad (3.10)$$

where:

$$\begin{aligned} A &= N_\gamma(2) = |S(\gamma_2, \gamma_3, \ldots, \gamma_n)|_a, \\ B &= N_\gamma(3) = |S(\gamma_3, \gamma_4, \ldots, \gamma_n)|_a, \\ \Delta(\gamma) &= n - 1 - (\gamma_1 + \ldots + \gamma_n) - \mathit{unary}(\gamma_n). \end{aligned}$$

The detailed proof of the above theorem is shown in the section 3.3. Now we can use the formula from the equation (3.10) to compute the number of runs in some example standard words.

Example 3.11
Let $\gamma = (1, 2, 1, 3, 1)$ be a directive sequence. We have $n = 4$ and

$$\mathrm{Sw}(\gamma) \ = \ ababaabababaabababaabababaababaab.$$

In this case

$$A = N_\gamma(2) = 5, \quad B = N_\gamma(3) = 4, \quad \Delta = (4-1) - 7 - 1 = -5, \quad odd(4) = 0.$$

Theorem 3.9 implies:

$$\begin{aligned} \rho(w) &= (\gamma_1 + 2)\ A + B + \Delta - odd(4) \\ &= 4\ A + B - 5 \\ &= 4 \cdot 5 + 4 - 5 \\ &= 19, \end{aligned}$$

see Figure 3.1 and Example 3.1 for comparison. □

The number of runs in the n-th Fibonacci word is given by the formula

$$\rho(F_n) \ = \ 2\ F_{n-2} + 3,$$

see [48] for the proof. As the next example we derive the same formula using the results from Theorem 3.9.

Example 3.12
Recall that $F_n = \mathrm{Sw}(1, 1, \ldots 1)$ (n ones) and in this case $N_\gamma(k) = F_{n-k-1}$. According to the formula from the equation (3.10) we have:

$$\begin{aligned} \rho(F_n) &= 2\ N_\gamma(2) + 2\ N_\gamma(3) + n - 1 - n - 1 - 1 \\ &= 2\ F_{n-3} + 2\ F_{n-4} - 3 \\ &= 2\ F_{n-2} - 3. \end{aligned}$$

□

The following lemma gives us the bound for the number of runs in standard words described by the directive sequences of the length at most 2.

Lemma 3.13
Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence, $w = \mathrm{Sw}(\gamma)$ and $n \leq 2$. Then

$$\rho(w) < \frac{4}{5}\ |w|.$$

Proof

Recall that the standard word given by the empty directive sequence is a and does not include any repetition. Therefore, we have to consider two cases: $|\gamma| = 1$ and $|\gamma| = 2$.

First assume that $|\gamma| = 1$. Then

$$w = \mathrm{Sw}(\gamma_0) = a^{\gamma_0} b \qquad \text{and} \qquad |w| = \gamma_0 + 1.$$

There is one run for $\gamma_0 > 1$, no run for $\gamma_0 = 1$ and obviously $\rho(w) < \frac{4}{5}|w|$.

Assume now that $|\gamma| = 2$. We have

$$w \;=\; \mathrm{Sw}(\gamma_0, \gamma_1) \;=\; \big(a^{\gamma_0} b\big)^{\gamma_1} a \qquad \text{and} \qquad |w| \;=\; (\gamma_0 + 1)\gamma_1 + 1.$$

The number of runs in w depends on the values of γ_0 and γ_1 as follows:

$$\rho(w) = \begin{cases} 0 & \text{for} \quad \gamma_0 = 1, \; \gamma_1 = 1 \\ 1 & \text{for} \quad \gamma_0 > 1, \; \gamma_1 = 1 \\ 1 & \text{for} \quad \gamma_0 = 1, \; \gamma_1 > 1 \\ \gamma_1 + 1 & \text{for} \quad \gamma_0 > 1, \; \gamma_1 > 1 \end{cases} .$$

In each case we have

$$\rho(w) \;<\; \frac{4}{5}\Big((\gamma_0 + 1)\gamma_1 + 1\Big) \;=\; \frac{4}{5}|w|$$

and the proof is complete. □

Now we are ready to estimate the asymptotic bound for the number of runs in all standard Sturmian words.

Theorem 3.14

For each standard Sturmian word w we have $\rho(w) \leq \frac{4}{5}\,|w|$.

Proof

Let $\gamma = (\gamma_0, \ldots, \gamma_n)$ be a directive sequence and $w = \mathrm{Sw}(\gamma_0, ..., \gamma_n)$ be a standard word. Recall the formula (3.10) from Theorem 3.9 and observe that $\Delta(\gamma) \leq 0$.

The case when $n \leq 2$ follows from the lemma 3.13. It is sufficient to prove the thesis for $n \geq 3$. We consider the four cases depending on the values of γ_0 and γ_1.

Case 1: $\gamma_0 = \gamma_1 = 1$.

We have, due to Lemma 3.8 and the equation (3.10):

$$|w| \;=\; 3A + 2B \qquad \text{and} \qquad \rho(w) \;\leq\; 2\,A + 2\,B.$$

Hence

$$\frac{\rho(w)}{|w|} \;\leq \frac{2A + 2B}{3A + 2B} \;\leq \frac{4}{5},$$

due to inequalities $A \geq B \geq 1$. This completes the proof of this case.

Case 2: $\gamma_0 = 1;\; \gamma_1 > 1$.

We have, due to Lemma 3.8 and the equation (3.10):

$$|w| \;=\; (2\,\gamma_1 + 1)\;A + 2B \qquad \text{and} \qquad \rho(w) \leq (\gamma_1 + 2)\;A + B.$$

Consequently:

$$\frac{\rho(w)}{|w|} \leq \frac{(\gamma_1 + 2)\;A + B}{(2\,\gamma_1 + 1)\;A + 2B} \leq \frac{4}{5},$$

because $\gamma_1 \geq 2$ and $\frac{\gamma_1+2}{2\,\gamma_1+1} \leq \frac{4}{5}$.

Case 3: $\gamma_0 > 1;\; \gamma_1 = 1$.

Due to the equation (3.10) and Lemma 3.8, we have:

$$\rho(w) \;\leq\; 2A + 3B,$$

$$|w| \;=\; \Big((\gamma_0 + 2)\;A + (\gamma_0 + 1)\;B\Big) \;\geq\; 4A + 3B,$$

and consequently:

$$\frac{\rho(w)}{|w|} \;\leq\; \frac{2A + 3B}{4A + 3B} \;\leq\; \frac{3A + 2B}{4A + 3B} \;\leq\; \frac{3}{4} \;<\; \frac{4}{5}.$$

Case 4: $\gamma_0 > 1;\; \gamma_1 > 1$.

In this case, due to the equation (3.10) and Lemma 3.8, we have:

$$\rho(w) \;\leq\; (2\,\gamma_1 + 1)\;A \;+\; 2\;B,$$

$$|w| \;=\; \big((\gamma_0 + 1)\;\gamma_1 + 1\big)\;A \;+\; (\gamma_0 + 1)\;B,$$

and consequently

$$\frac{\rho(w)}{|w|} \;\leq\; \frac{(2\,\gamma_1 + 1)\;A \;+\; 2\;B}{\big((\gamma_0 + 1)\;\gamma_1 + 1\big)\;A \;+\; (\gamma_0 + 1)\;B} \;\leq\; \frac{(2\,\gamma_1 + 1)\;A \;+\; 2\;B}{(3\,\gamma_1 + 1)\;A \;+\; 3\;B} \;\leq\; \frac{4}{5},$$

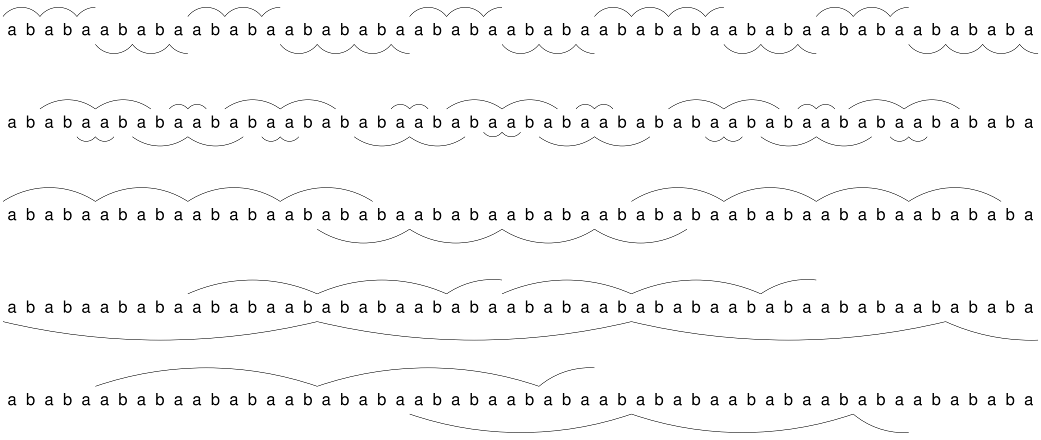
a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a
a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a
a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a
a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a
a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a a b a b a a b a b a a b a b a b a

Figure 3.2: *The structure of runs in the word* $\mathrm{Sw}(1,2,k,k)$ *for* $k = 3$. *There are* $4k^2 - k + 3 = 36$ *runs.*

because

$$\frac{2\,\gamma_1 + 1}{3\,\gamma_1 + 1} \leq \frac{4}{5}.$$

This completes the proof of the theorem. □

The above results give us the asymptotic bound for the number of runs in standard words. Below we construct a strictly growing sequence of standard words to show that this estimation is tight.

Theorem 3.15
For the class $\mathcal{S}$ *of standard words we have:*

$$\sup\left\{ \frac{\rho(w)}{|w|} \;:\; w \in \mathcal{S} \right\} \;=\; 0.8.$$

Proof
Let $\gamma = (1,2,k,k)$ and $w_k = \mathrm{Sw}(1,2,k,k)$. By the definition we have

$$w_k \;=\; \Big((ababa)^k\; ab\Big)^k ababa \qquad \text{and} \qquad |w_k| \;=\; 5k^2 + 2k + 5,$$

and due to the equation (3.10):

$$\big|\rho(w_k)\big| \;=\; 4k^2 - k + 3,$$

see Figure 3.2 for the case $k = 3$. Consequently

$$\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} \;=\; \lim_{k\to\infty} \frac{4k^2 - k + 3}{5k^2 + 2k + 5} \;=\; 0.8,$$

which completes the proof. □

The formula from the equation (3.10) leads to a simple and fast algorithm to compute the number of runs in standard words.

Theorem 3.16
We can count the number of runs in any standard word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *in linear time with respect to the length of the directive sequence* $|\gamma|$.

Proof
The formula for the number of runs in standard words from Theorem 3.9 depends directly on the components of the directive sequence γ and the numbers $N_\gamma(2)$ and $N_\gamma(3)$. It is sufficient to prove that we can compute the numbers $N_\gamma(k)$ for $k = 1, 2, 3$ in time $O(n)$. For this purpose we can iterate the equation (1.1) from the section 1.1:

Algorithm to compute $N_\gamma(k)$

```
Input: $\gamma = (\gamma_0, \ldots, \gamma_n)$
Output: $N_\gamma(k)$

$x \leftarrow 1$
$y \leftarrow 0$
for $i = n$ downto $k$ do
    $x \leftarrow \gamma_i \cdot x + y$
    $y \leftarrow x$
end for
return $x$
```

Using the above algorithm and applying the formula from the equation (3.10) we can count the number of runs in any standard word in linear time with respect to the size of the directive sequence (logarithmic with respect to the length of the whole word). □

3.3 The proof of the main theorem

This section is devoted to the proof of Theorem 3.9. We begin with the characterization of the structure of the possible periods of the maximal repetitions in standard words. Recall their recurrent definition given by the equation (1.1) and the words x_i defined there.

The following lemma is the consequence of the structure of subword graphs of standard words, see Lemma 2.3 comparison.

Lemma 3.17 (Structural Lemma)
The period of each run in the word $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *is of the form* $x_i^j x_{i-1}$*, where* $0 \leq j < \gamma_i$*.*

To prove the above lemma it is sufficient to show that no factor of the 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 [24] for details.

The main idea of the proof of the correctness of the formula given in the equation (3.10) is the partition of the set of the maximal repetitions in the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ into three separate categories depending on the length of their periods. We say that a run is:

short – if the length of its period does not exceed $|x_1|$,

large – if the length of its period exceeds $|x_2|$,

medium – otherwise.

Denote by $\rho_S(w)$, $\rho_M(w)$ and $\rho_L(w)$ the number of short, medium and large runs in the word w, respectively.

Example 3.18
Recall the word $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ from Example 3.1 and the set of its maximal repetitions. In this case we have:

- 10 short runs (periods *a* and *ab*),
- 8 medium runs (periods *aba* and *ababa*),
- 1 large run (the period *ababaab*),

see Figure 3.1 for comparison. □

Counting short runs

We start with the computation of the number of the *short* runs. These are the runs with the periods of the form a or a^+b. Their number depends on the value of γ_0 and γ_1.

Lemma 3.19 (Short Runs)
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *be a standard word. The number of* short *runs in* w *is given by the formula:*

$$\rho_S(w) = \begin{cases} N_\gamma(2) + N_\gamma(3) - 1 & \text{for } \gamma_0 = 1,\ \gamma_1 = 1 \\ 2\, N_\gamma(2) - \mathit{odd}(n) & \text{for } \gamma_0 = 1,\ \gamma_1 > 1 \\ N_\gamma(1) + N_\gamma(3) - \mathit{even}(n) & \text{for } \gamma_0 > 1,\ \gamma_1 = 1 \\ N_\gamma(1) + N_\gamma(2) & \text{for } \gamma_0 > 1,\ \gamma_1 > 1 \end{cases}.$$

Proof

Short runs are the runs with the periods of the form a or $a^k b$. We estimate the number of runs with the periods of each type separately.

Case 1: runs with the periods of the form a.

First assume that $\gamma_0 > 0$. 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 1.15, 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.

Assume now that $\gamma_0 = 1$. In this case the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ consists of the blocks of the two types: ab or aab and only the blocks of the second type include the runs with the period a. Due to Lemma 1.15, 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 in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$ equals the number of the letters b in $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$, hence $N_\gamma(2)$. 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 $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$, hence $N_\gamma(2) - 1$. Finally the whole case can be summarized as:

$$\begin{cases} N_\gamma(2) - \mathit{odd}(n) & \text{for} \quad \gamma_0 = 1 \\ N_\gamma(1) & \text{for} \quad \gamma_0 > 1 \end{cases}.$$

Case 2: runs with the periods of the form $a^k b$.

Notice that, due to the equation (1.14) and Lemma 1.15, the runs with the periods $a^{\gamma_0}b$ and $a^{\gamma_0+1}b$ in in the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ correspond to the runs with the periods a in the word $\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)$. Similar reasoning as above shows that the number of such in the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ equals;

$$\begin{cases} N_\gamma(3) - \mathit{even}(n) & \text{for} \quad \gamma_1 = 1 \\ N_\gamma(2) & \text{for} \quad \gamma_1 > 1 \end{cases}.$$

Combining the results from the two above cases we conclude the thesis of the lemma. □

Counting medium runs

Recall that *medium* runs are the maximal repetitions with the periods $x_1^k x_0$ for $0 < k < \gamma_1$ and x_2, where x_i are as in the equation (1.1). Observe that the medium runs appear in the standard words generated by the directive sequences of the length at least 3. We have to consider two cases: the directive sequences of the length 3 and the longer directive sequences. The value of γ_0 does not affect the number of the medium runs, hence to simplify the calculations we assume in further proofs that $\gamma_0 = 1$. We start with counting the medium runs in the standard words generated by the directive sequences of the length greater than 3.

Lemma 3.20 (Medium runs, n $\geq$ 3)
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word and $n \geq 3$. The number of medium *runs in w is given by the formula:*

$$\rho_M(w) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1.$$

The thesis of the lemma is the consequence of the following stronger claim:

Claim 3.21
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. There are:

(1) $N_\gamma(2) - 1$ runs with the period $x_1^i x_0$ for each $0 < i < \gamma_1$.

(2) $N_\gamma(3)$ runs with the period x_2.

Proof
Point (1)
The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ has the form:

$$(ab)^{\alpha_1} a (ab)^{\alpha_2} a \ldots (ab)^{\alpha_s} a\; ab \qquad \text{or} \qquad (ab)^{\alpha_1} a (ab)^{\alpha_2} a \ldots (ab)^{\alpha_s} a,$$

where $0 < \alpha_i \leq (\gamma_1 + 1)$ and $s = N_\gamma(2)$, because, due to Lemma 1.15, every factor $(ab)^{\alpha_i} a$ in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ corresponds to the letter a in $\mathrm{Sw}(\gamma_2, \ldots, \gamma_n)$. For example, see Figure 3.3, the word $\mathrm{Sw}(1, 4, 2, 2)$ has the form

$$\mathrm{Sw}(1,4,2,2) \;=\; (ab)^4 a (ab)^4 a (ab)^5 a (ab)^4 a (ab)^5 a.$$

Each pair of neighboring factors: $(ab)^{\alpha_i} a \cdot (ab)^{\alpha_{i+1}} a$ produces $\gamma_1 - 1$ runs with the period $(ab)^k a$ for each $0 < k < \gamma_1$. In the word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ we have $N_\gamma(2) - 1$ such pairs and therefore $(N_\gamma(2) - 1)(\gamma_1 - 1)$ medium runs with the periods $x_1^k x_0$ for $0 < k < \gamma_1$.

Point (2)
The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ can be represented as a sequence of concatenated words x_1 and x_2 and has the form:

$$(a): \quad x_2^{\alpha_1} x_1 x_2^{\alpha_2} x_1 \ldots x_2^{\alpha_s} x_1 x_2 \qquad \text{or} \qquad (b): \quad x_2^{\beta_1} x_1 x_2^{\beta_2} x_1 \ldots x_2^{\beta_s} x_1.$$

For example the word $\mathrm{Sw}(1,4,2,2)$ has the decomposition $x_2^2 x_1 x_2^2 x_1 x_2$, see Figure 3.3.

First assume the case (a). Each run with the period x_2 has the form $x_2^k x_1$. By the definition of standard words the factor $x_1 x_2$ has x_2 as a prefix. Therefore, the number of such runs in $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ equals the number of factors x_1 in the decomposition mentioned above, which, due to Lemma 1.15, corresponds to the number of the letters b in $\mathrm{Sw}(\gamma_2, \ldots, \gamma_n)$, namely $N_\gamma(3)$.

Assume now the case (b). The word $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ has the suffix x_1 but in this case we have $\beta_s \geq 2$. Hence the number of runs with the period x_2 is the same as in the case (a). □

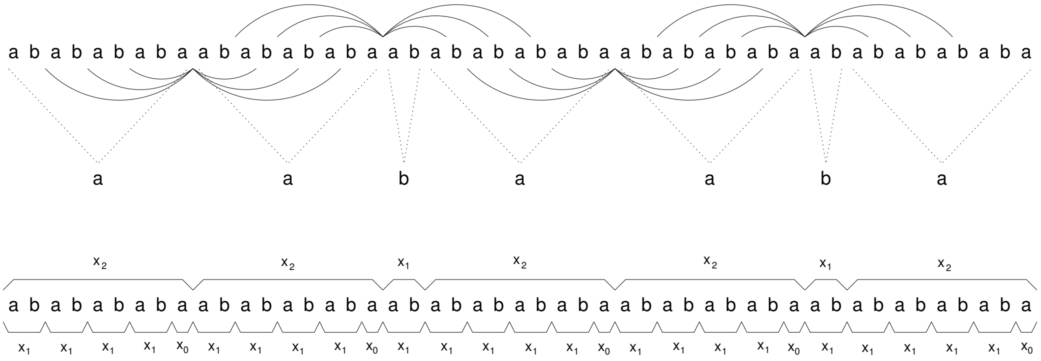
a b a b a b a b a a b a b a b a b a a b a b a b a b a b a a b a b a b a b a a b a b a b a b a
a a b a a b a
x2 x2 x1 x2 x2 x1 x2
a b a b a b a b a a b a b a b a b a a b a b a b a b a b a a b a b a b a b a a b a b a b a b a
x1 x1 x1 x1 x0 x1 x1 x1 x1 x0 x1 x1 x1 x1 x1 x0 x1 x1 x1 x1 x0 x1 x1 x1 x1 x1 x0

Figure 3.3: *The structure of runs with the periods $|x_1| < p < |x_2|$ for the word* $\mathrm{Sw}(1,4,2,2)$ *and its decomposition into words x_1, x_2 and x_0, x_1.*

Proof of Lemma 3.20
Summing up the formulas from the points (1) and (2) of Claim 3.21 we obtain:

$$\begin{aligned}
\rho_M &= \big(N_\gamma(2) - 1\big)\,(\gamma_1 - 1) + N_\gamma(3) \\
&= \big(\gamma_1 N_\gamma(2) + N_\gamma(3)\big) \;-\; N_\gamma(2) - \gamma_1 + 1 \\
&= N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1
\end{aligned}$$

and this completes the proof of the lemma. □

The structure of the medium runs in standard words defined by the directive sequences of the length 3 is slightly different.

Lemma 3.22 (Medium runs, n=2)
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2)$ be a standard word. The number of medium *runs in w is given by the formula:*

$$\rho_M(w) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1 - unary(\gamma_2)$$

Proof
The proof for the case $\gamma_2 > 1$ follows the same argumentation as the one for Lemma 3.20.

In the case $\gamma_2 = 1$ the word $\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2)$ has the decomposition

$$\mathrm{Sw}(\gamma_0, \gamma_1, \gamma_2) \;=\; \big(a^{\gamma_0}b\big)a \cdot a^{\gamma_0}b \;=\; x_2 \cdot x_1.$$

There is no run with the period x_2, and we have to subtract 1 from the number of the factors x_1 in this case. □

The recurrence for large runs

Recall that the run is called *large* if it has the period of the length greater than $|x_2|$, where x_2 is as in the equation (1.1). We reduce the problem of counting the large runs to the one for counting the medium runs, using the morphic representation of the standard words introduced in the section 1.3.

Let h be a morphism and let $v = a_1a_2\dots a_k$ be the word of the length k. The morphism h defines the partition of the word $w = h(v)$ into segments $h(a_1)$, $h(a_2)$,..., $h(a_t)$. These segments are called the *h-blocks*. We say that a factor x of the word w is *synchronized* with the morphism h in w if and only if each occurrence of x in w starts at the beginning of some h-block and ends at the end of some h-block. Observe that every factor in w that is synchronized with h corresponds to some factor in v, hence the morphism h preserves the structure of the factors that are synchronized with it.

Example 3.23
Let $w = \mathrm{Sw}(1, 2, 1, 3, 1)$ and $v = \mathrm{Sw}(2, 1, 3, 1)$ be standard words and h_0 be the morphism defined as:

$$h_0 : \begin{cases} a \longrightarrow ab \\ b \longrightarrow a \end{cases}.$$

Recall that

$$\begin{aligned}\mathrm{Sw}(1,2,1,3,1) &= h_0\bigl(\mathrm{Sw}(2,1,3,1)\bigr),\\ \mathrm{Sw}(1,2,1,3,1) &= ababaababababaababababaababababaababaab,\\ \mathrm{Sw}(2,1,3,1) &= aabaaabaaabaaabaaba.\end{aligned}$$

The factors $w[6..8] = aba$ and $w[13..17] = abaab$ are not synchronized with the morphism h_0, because both of them ends in the middle some h_0-block. The factor $w[22..28]$ starts at the beginning of some h_0-block and ends at the end of some h_0-block, hence is synchronized with the morphism h_0. Moreover it corresponds with the factor $v[13..16] = aaba$, see Figure 3.4 for comparison. □

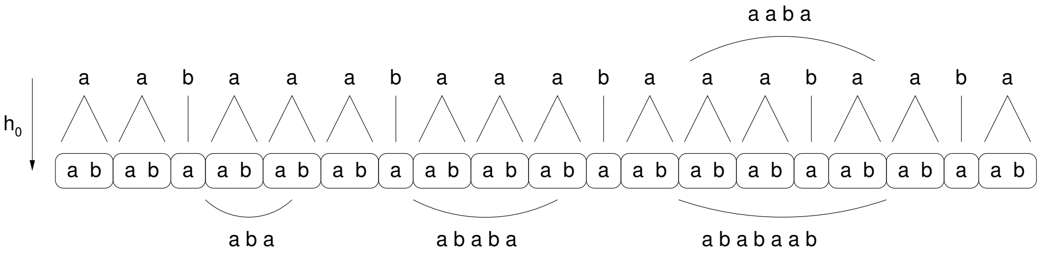
aaba
a a b a a a b a a a b a a a b a a b a
h0
ab ab a ab ab ab a ab ab ab a ab ab ab a ab ab a ab
aba
ababa
abababaab

Figure 3.4: *The periods of the medium runs $x_1x_0 = aba$ and $x_2 = ababa$ do not synchronize with the morphism h_0 in the word* Sw(1, 2, 1, 3, 1)*, while the period of the large run $x_3 = ababaab$ is synchronized with h_0 and corresponds to the medium run with the period $x_1x_0 = aaba$ in the word* Sw(2, 1, 3, 1).

Lemma 3.24 (Synchronization Lemma)
The periods of the large *runs in the word* $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ *are synchronized with the morphism h_0.*

Proof
Let h_0 be the morphism defined as

$$h_0 : \begin{cases} a \longrightarrow a^{\gamma_0}b \\ b \longrightarrow a \end{cases}.$$

Due to Lemma 1.15, we have

$$\mathrm{Sw}(\gamma_0, \ldots, \gamma_n) = h_0\bigl(\mathrm{Sw}(\gamma_1, \ldots, \gamma_n)\bigr).$$

Moreover, h_0 determines the partition of $\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ into h_0-blocks of the form $a^{\gamma_0}b$ and a, see Figure 3.4 for the partition of Sw(1, 2, 1, 3, 1).

Recall that the period of each large run in the word $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ is of the form $x_i^k x_{i-1}$, where $0 \leq k < \gamma_i$ and $i \geq 2$. By the definition of standard words the factor $x_i^k x_{i-1}$ starts with $a^{\gamma_0}b$, hence at the beginning of some h_0-block.

For even $i \geq 2$ the subword $x_i^k x_{i-1}$ ends with $x_1 = a^{\gamma_0}b$, hence at the end of some h_0-block, and is obviously synchronized with h_0.

For odd $i \geq 2$ the factor $x_i^k x_{i-1}$ ends with

$$x_3 \cdot x_2 \;=\; x_2^{\gamma_2}x_1 \cdot x_1^{\gamma_1}x_0 \;=\; x_2^{\gamma_2} \cdot (a^{\gamma_0}b)^{\gamma_1+1}a.$$

First assume that $x_i^k x_{i-1}$ is followed by the block $a^{\gamma_0}b$. The single letter a at the end of $x_i^k x_{i-1}$ is then the whole h_0-block and $x_i^k x_{i-1}$ is synchronized with the morphism h_0.

Assume now that $x_i^k x_{i-1}$ ends with $(a^{\gamma_0}b)^{\gamma_1+1}a$ and is followed by $(a^{\gamma_0-1}b)$, namely it ends in the middle of some h_0-block. In this case we have the occurrence of the factor $(a^{\gamma_0}b)^{\gamma_1+2}$ in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$, which is reduced by the morphism h_0^{-1} to the factor $a^{\gamma_1+2}b$ in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. By the definition the standard word $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$ can include only the blocks of the two types: the short block – $a^{\gamma_1}b$ and the long block – $a^{\gamma_1+1}b$ (see the section 1.1), hence we have the contradiction and the proof is complete. □

The following lemma, which is a direct conclusion from Synchronization lemma, allows us to reduce the problem of counting the large runs in the word $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ to those in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$.

Lemma 3.25 (Recurrence Lemma)
Let $w = \mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ and $v = \mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$ be standard words. The number of large *runs in w is given by the recurrence*

$$\rho_L(w) \;=\; \rho_L(v) \;+\; \rho_M(v).$$

Proof
The synchronization lemma implies that the morphism defined as in the equation (3.2) preserve the structure of the long runs in standard words. Recall that $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ is reduced by h_0^{-1} to $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. Therefore, every large run α in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ corresponds to some run β in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$.

Due to Lemma 3.17, the period of the run α is of the form $x_i^k x_{-1}$, where $0 < k \leq \gamma_i$ and $i \geq 2$. The corresponding run β is either large (for $i = 2$) or medium (for $i = 2$). Hence to compute all large runs in $\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)$ it is sufficient to compute all large and medium runs in $\mathrm{Sw}(\gamma_1,\ldots,\gamma_n)$. □

The thesis of the next lemma gives us the compact formula for the number of the medium and the large runs in standard words.

Lemma 3.26 (Large Runs)
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ be a standard word. We have

$$\rho_L(w) \ + \ \rho_M(w) \ = \ N_\gamma(1) \ + \ n-1 \ - \ (\gamma_1 + \ldots + \gamma_n) \ - \ unary(\gamma_n).$$

Proof
Due to the formulas from Lemma 3.20 and Lemma 3.22 and the recurrence from Lemma 3.25 we have

$$\begin{aligned}\rho_L(w) \ + \ \rho_M(w) \ &= \ \sum_{i=0}^{n-2} \rho_M\big(\mathrm{Sw}(\gamma_i, \ldots, \gamma_n)\big) \\ &= \ \Big(N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1\Big) \ + \\ & \qquad\qquad \vdots \\ & \quad \Big(N_\gamma(n-2) - N_\gamma(n-1) - \gamma_{n-2} + 1\Big) \ + \\ & \quad \Big(N_\gamma(n-1) - N_\gamma(n) - \gamma_{n-1} + 1 - unary(\gamma_n)\Big).\end{aligned}$$

Taking into account that $N_\gamma(n) = \gamma_n$ the above formula can be written as

$$\rho_L(w) \ + \ \rho_M(w) \ = \ N_\gamma(1) \ + \ (n-1) \ - \ (\gamma_1 + \ldots + \gamma_n) \ - \ unary(\gamma_n),$$

which concludes the thesis. □

Now we are ready to prove the formula for the number of runs in standard words given by the equation (3.10).

Proof of Theorem 3.9
To obtain the formula from the equation (3.10) it is sufficient to combine the formulas from Lemma 3.19 and Lemma 3.26. □

4 Squares in Sturmian words

A square in a string is a subword of the form ww, where w (called the *period* or the *generator*) is nonempty. Squares are the simplest forms of repetitions, but despite the simple formulation many combinatorial problems related to squares are not well understood. The subject of computing the number of distinct squares and other types of repetitions in words is one of the fundamental topics in combinatorics on words as well as it is important in other areas: lossless compression, word representation, computational biology, etc. See for instance [10], [44], [52], [73].

Denote by $sq(w)$ the number of distinct squares in the word w and by $sq(n)$ the maximal number of distinct squares in words of length n. The behaviour of the function $sq(n)$ is not well understood, although the subject has been studied by many authors, see for example [22], [23] and [40]. The best known results related to the value of $sq(n)$ are:

$$n - O(n) \;\leq\; sq(n) \;\leq\; 2n - O(\log n),$$

compare with the results in [27], [38] and [39]. See Table 4.1 for the maximal number of squares and the square ratio in words over a binary alphabet for small values of n. In this chapter we concentrate on the asymptotic behaviour of the maximal number of distinct squares in standard Sturmian words and give the thigh asymptotic bound for $sq(n)$ for this class of strings.

There are known efficient algorithms for the computation of integer powers in words, see [2], [17], [24], [54], [55]. The powers in words are related to maximal repetitions, also called *runs*, see the chapter 3 for more information. It is surprising that the known bounds for the number of runs are much tighter than for squares, which is due to the work of many people [5], [19], [20], [32], [47], [48], [62], [66], [68].

One of the interesting questions related to squares is the relation of their number to the number of runs. In case of Fibonacci words the numbers of squares and runs differ only by 1 and have the same asymptotic behaviour, see [28], [48]. The analysis of such relation for standard word is done in the section 4.4.

n	$sq(n)$	$sq(n)/n$	Example word
10	6	0.6	abbabbabaa
11	7	0.636364	bababbabbaa
12	7	0.583333	abbabbabaaaa
13	8	0.615385	bababbabbaaaa
14	9	0.642857	ababaababaabaa
15	10	0.666667	aababaababaabaa
16	11	0.6875	baabaababaababaa
17	12	0.705882	bbaabaababaababaa
18	12	0.666667	baabaababaababaaaa
19	13	0.684211	bbaabaababaababaaaa
20	13	0.65	baabaababaababaaaaaa
21	14	0.666667	bbaabaababaababaaaaaa
22	15	0.681818	aaabaabaaabaabaaabaaaa
23	16	0.695652	baaabaaabaabaaabaabaaaa
24	17	0.708333	bbaaabaaabaabaaabaabaaaa
25	18	0.72	abaabaababaabaababaababaa
26	19	0.730769	babaabaababaabaababaababaa
27	20	0.740741	bbabaabaababaabaababaababaa
28	20	0.714286	aabaabaaabaaaabaaabaaaabaaaa
29	21	0.724138	ababaabaaabaaaabaaabaaaabaaaa
30	22	0.733333	bbaaaabaaabaaaabaaabaaaabaabaa
31	23	0.741935	ababaaabaaaabaaabaaaabaaabaabaa

Table 4.1: *The maximal number of distinct squares and the square ratio for the binary words of the given length.*

4.1 Formulas for the number of squares

The exact formulas for the number of distinct squares in standard Sturmian words were given by Damanik and Lenz in [24]. In this section we reformulate their equations to have compact version more suitable for the asymptotic analysis. The formulas are rather complicated and such an analysis is nontrivial. It is the matter of the section 4.3.

Denote $q_i = |x_i|$, where x_i are as in the equation (1.1). We have then

$$q_{-1} = q_0 = 1 \qquad \text{and} \qquad q_{i+1} = \gamma_i q_i + q_{i-1}. \tag{4.1}$$

The following lemma characterize the possible lengths of the periods of squares in Sturmian words.

Lemma 4.2 (See [24])
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word. Each primitive period of a square in w has the length kq_i for $1 \leq k \leq \gamma_i$ or $kq_i + q_{i-1}$ for $1 \leq k < \gamma_i$.

The squares in the standard Sturmian word w with the period of the length kq_i, for $1 \leq k \leq \gamma_i$, or $kq_i + q_{i-1}$, for $1 \leq k < \gamma_i$, are said to be of the *type i*. The squares with the period of the form a^+ are said to be of the *type* 0.

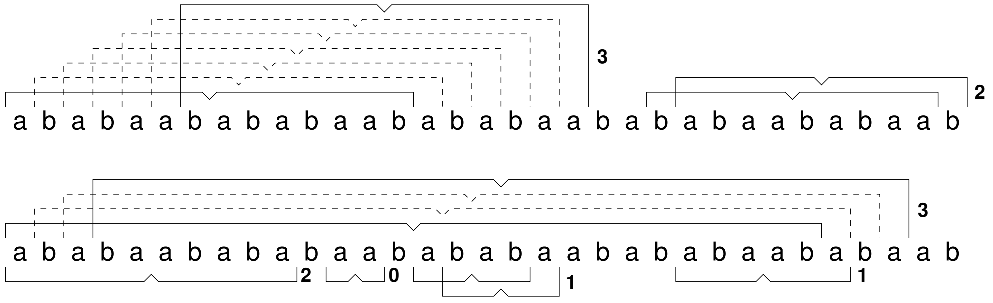
3
2
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
3
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
2
0
1
1

Figure 4.1: *The squares in the word* $\mathrm{Sw}(1,2,1,3,1)$ *with their types.*

Example 4.3
Consider the word: $\mathrm{Sw}(1,2,1,3,1) = ababaabababaabababaababababaababaab$.
We have: one square of type 0: $a{\cdot}a$,
three squares of type 1 (periods 2, 3): $ab{\cdot}ab$, $ba{\cdot}ba$, $aba{\cdot}aba$,
three squares of type 2 (period 5): $ababa{\cdot}ababa$, $babaab{\cdot}babaab$, $abaab{\cdot}abaab$,
and eleven squares of type 3 (periods 7, 14):

$ababaab{\cdot}ababaab$, $babaaba{\cdot}babaaba$, $abaabab{\cdot}abaabab$,
$baababa{\cdot}baababa$, $aababab{\cdot}aababab$, $abababa{\cdot}abababa$,
$bababaa{\cdot}bababaa$,
$ababaababababaab{\cdot}ababaababababaab$, $babaababababaaba{\cdot}babaababababaaba$,
$abaabababaabab{\cdot}abaabababaabab$, $baababababaababa{\cdot}baababababaababa$,

see Figure 4.1 for comparison. □

For a standard word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ denote by $sq(\gamma_0, \gamma_1, ..., \gamma_n)$ the number of distinct squares in w and by $sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n)$ the number of distinct squares of the type i in w (for $0 \leq i \leq n$).

We count all squares in w by counting separately the squares of each type:

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \; = \; \sum_{i=0}^{n} \; sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n).$$

For a directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ and $1 \leq i \leq n$ denote:

$$\begin{aligned}
d(0) &= \left\lfloor \tfrac{\gamma_0+1}{2} \right\rfloor, \\
d_1(i) &= \begin{cases} \frac{\gamma_i}{2} q_i + q_{i-1} - 1 & \text{for even } \gamma_i \\ \frac{\gamma_i}{2} q_i + \frac{1}{2}\; q_i & \text{for odd } \gamma_i \end{cases}, \qquad\qquad (4.4) \\
d(i) &= d_1(i) + \gamma_i \, q_i - q_i - \gamma_i + 1,
\end{aligned}$$

where q_i are as in the equation (4.1).

The number of distinct squares $sq(\gamma_0, \ldots, \gamma_n)$ is determined as follows, see [24]:

Summation formulas:

(**1**) $\quad sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \; = \; \sum_{i=0}^{n} \; sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n).$

(**2**) $\quad (0 \leq i \leq n-3)$ or $\; (i = n-2 \;\&\; \gamma_n \geq 2) \;\Rightarrow\; sq_i(\gamma) = d(i).$

(**3**) $$\gamma_n = 1 \;\Rightarrow\; sq_{n-2}(\gamma) \; = \; \begin{cases} d(n-2) - q_{n-3} + 1 & \text{for even } \gamma_{n-2} \\ d(n-2) - q_{n-2} + q_{n-3} + 1 & \text{otherwise} \end{cases}$$

(**4**) $$\gamma_n = 1 \;\Rightarrow\; sq_{n-1}(\gamma) \; = \; \begin{cases} d_1(n-1) - q_{n-2} + 1 & \text{for even } \gamma_{n-1} \\ d_1(n-1) - q_{n-1} + q_{n-2} - 1 & \text{otherwise} \end{cases}$$

(**5**) $$\gamma_n > 1 \;\Rightarrow\; sq_{n-1}(\gamma) \; = \; \begin{cases} d(n-1) - q_{n-2} + 1 & \text{for even } \gamma_{n-1} \\ d(n-1) - q_{n-1} + q_{n-2} - 1 & \text{otherwise} \end{cases}$$

(**6**) $$sq_n(\gamma) \; = \; \begin{cases} d_1(n) - q_n + 2 & \text{for even } \gamma_n \\ d_1(n) - q_n & \text{otherwise} \end{cases}$$

(4.5)

4.2 Standard words with many squares

In this section we present and analyze the sequence $\{w_k\}$ of strictly growing standard words achieving asymptotically maximal ratio of the number of distinct squares to the length of the word:

$$\lim_{k\to\infty} |w_k| = \infty \qquad \text{and} \qquad \lim_{k\to\infty} \frac{sq(w_k)}{|w_k|} = \frac{9}{10}.$$

Recall that the squares with the periods of the length kq_i, for $1 \le k \le \gamma_i$, or $kq_i + q_{i-1}$, for $1 \le k < \gamma_i$, where q_i are as in the equation (4.1), are said to be of the type i and the squares with the periods of the form a^+ – the type 0.

Consider a directive sequence $\gamma_k = (k, k, 2, 1, 1)$ and a sequence of words $w_k = \mathrm{Sw}(k, k, 2, 1, 1)$, where $k > 0$.

Example 4.6
We have:

$$\begin{aligned}
w_1 &= \mathrm{Sw}(1,1,2,1,1) &&= (aba)^2ab(aba)^3ab,\\
w_2 &= \mathrm{Sw}(2,2,2,1,1) &&= \Big((aab)^2a\Big)^2 aab\Big((aab)^2a\Big)^3 aab,\\
w_3 &= \mathrm{Sw}(3,3,2,1,1) &&= \Big((aaab)^3a\Big)^2 aaab\Big((aaab)^3a\Big)^3 aaab,\\
& && \vdots\\
w_k &= \mathrm{Sw}(k,k,2,1,1) &&= \Big(\big(a^kb\big)^ka\Big)^2 a^kb\Big(\big(a^kb\big)^ka\Big)^3 a^kb.
\end{aligned}$$

See Figure 4.2 for the structure of the distinct squares in $\mathrm{Sw}(3,3,2,1,1)$. □

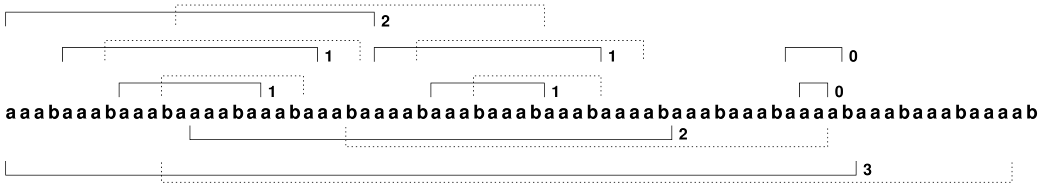
2
1
1
0
1
1
0
aaabaaabaaabaaaabaaabaaabaaaabaaabaaabaaabaaaabaaabaaabaaaabaaabaaabaaaab
2
3

Figure 4.2: The squares in the standard word $\mathrm{Sw}(3,3,2,1,1)$ *with their shifts and types.*

Theorem 4.7
For the standard word $\mathrm{Sw}(k,k,2,1,1)$ *we have*

$$sq(k,k,2,1,1) \;\longrightarrow\; \frac{9}{10}\cdot\big|\mathrm{Sw}(k,k,2,1,1)\big|,$$

for $k \;\longrightarrow\; \infty$.

Proof
Let $\gamma_k = (k,k,2,1,1)$. We have:

$$\mathrm{Sw}(\gamma_k) \;=\; \Big((a^kb)^ka\Big)^2 a^kb\Big((a^kb)^ka\Big)^3 a^kb$$

and

$$|\mathrm{Sw}(\gamma_k)| \;=\; 5k^2+7k+7.$$

We compute separately the number of distinct squares of each type $sq_i(\gamma_k)$ for $0 \le i \le 4$ in the word w_k.

There are two cases depending on the parity of the parameter k. We are interested in the asymptotic behaviour of the number of distinct squares, hence there is no loss of generality in assuming that $k > 1$.

Case 1: k is odd.

We have (according to formulas (1-6) from the equation (4.5)):

$$\begin{aligned}
sq_0(\gamma_k) &= \frac{1}{2}\Big(k+1\Big),\\
sq_1(\gamma_k) &= \frac{1}{2}\Big(3k^2+1\Big),\\
sq_2(\gamma_k) &= 2k^2+2k+1,\\
sq_3(\gamma_k) &= k^2+k,\\
sq_4(\gamma_k) &= 0.
\end{aligned}$$

Summing altogether we obtain:

$$sq(\gamma_k) \;=\; \frac{1}{2}\Big(9k^2+7k+4\Big),$$

and finally

$$\lim_{k\to\infty}\frac{sq(\gamma_k)}{|\mathrm{Sw}(\gamma_k)|} \;=\; \lim_{k\to\infty}\frac{9k^2+7k+4}{10k^2+14k+14} \;=\; \frac{9}{10}.$$

Case 2: k is even.

We have (according to formulas (1-6) from the equation (4.5)):

$$\begin{aligned} sq_0(\gamma_k) &= \frac{1}{2}k, \\ sq_1(\gamma_k) &= \frac{1}{2}\Big(3k^2 - k\Big), \\ sq_2(\gamma_k) &= 2k^2 + 2k + 1, \\ sq_3(\gamma_k) &= k^2 + k, \\ sq_4(\gamma_k) &= 0. \end{aligned}$$

Summing altogether we obtain:

$$sq(\gamma_k) = \frac{1}{2}\Big(9k^2 + 6k + 2\Big),$$

and finally

$$\lim_{k\to\infty} \frac{sq(\gamma_k)}{|\mathrm{Sw}(\gamma_k)|} = \lim_{k\to\infty} \frac{9k^2 + 6k + 2}{10k^2 + 14k + 14} = \frac{9}{10}.$$

□

4.3 Asymptotics of the number of squares

The formulas (1–6) from the equation (4.5) give us the value of $sq(\gamma_0, \dots, \gamma_n)$, however there is no close simple formula. Therefore, the tight asymptotic estimations are nontrivial.

We start with the two lemmas that allow us to restrict the value of the last two elements of the directive sequence $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_n)$ in the asymptotic estimation of the maximal number of distinct squares in $\mathrm{Sw}(\gamma_0, \dots, \gamma_n)$.

Lemma 4.8 (Reduction of γ_n)
Let $\mathrm{Sw}(\gamma_0, \gamma_1, \dots, \gamma_n)$ *be a standard Sturmian word. If* $\gamma_n > 1$ *then*

$$sq(\gamma_0, \dots, \gamma_{n-1}, \gamma_n) \leq sq(\gamma_0, \dots, \gamma_{n-1}, \gamma_n - 1, 1) + 2.$$

Proof
Recall from the section 1.1 that the standard words $w_1 = \mathrm{Sw}(\gamma_0, \dots, \gamma_n)$ and $w_2 = \mathrm{Sw}(\gamma_0, \dots, \gamma_n - 1, 1)$ differ only in the last two letters – *ab* or *ba*.

The squares of the types 0, 1, ..., $n-1$ and short squares of the type n are the same for w_1 and w_2 (see Figure 4.3). The difference is possible only for the longest squares of the type n. Exchanging the last two letters enables (or disables respectively) the shift of the longest square of the type n by one and two positions (see the squares marked in bold on Figure 4.3). In w_2 we have $\gamma_{n+1} = 1$ and, due to formulas (1-6) from the equation (4.5), there is no square of the type $n+1$. Therefore, the difference between the numbers of squares in w_1 and w_2 is not greater than 2, what follows the thesis.

□

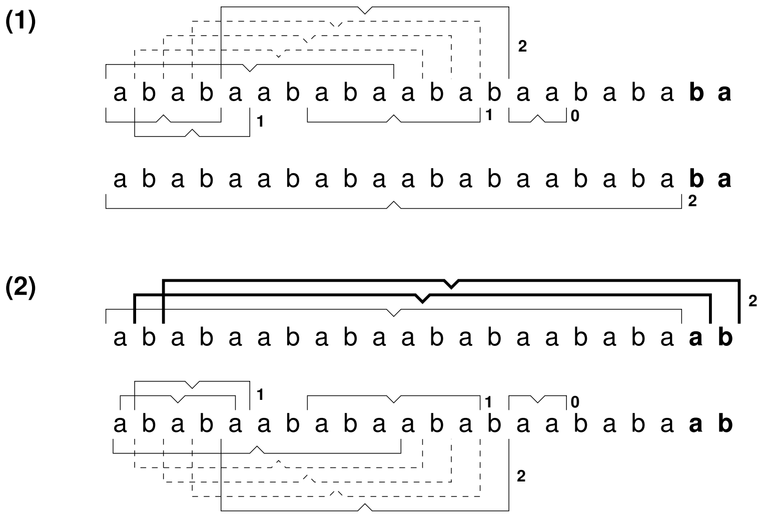
(1)
2
a b a b a a b a b a a b a b a a b a b a b a
1
1
0
a b a b a a b a b a a b a b a a b a b a b a
2
(2)
2
a b a b a a b a b a a b a b a a b a b a a b
1
1
0
a b a b a a b a b a a b a b a a b a b a a b
2

Figure 4.3: The squares in the standard words (1): Sw(1, 2, 3, 1) *and (2):* Sw(1, 2, 4)*. Two additional squares of the type 2 are marked in bold.*

Lemma 4.9 (Reduction of γ_{n-1})
Let $w = \mathrm{Sw}(\gamma_0, \dots, \gamma_{n-2}, \gamma_{n-1}, 1)$, $w_1 = \mathrm{Sw}(\gamma_0, \dots, \gamma_{n-2}, 1, 1)$ *and* $w_2 = \mathrm{Sw}(\gamma_0, \dots, \gamma_{n-2}, 2, 1)$ *be standard words. If the inequalities*

$$sq(w_1) \leq 0.9\,|w_1| - 2 \qquad \text{and} \qquad sq(w_2) \leq 0.9\,|w_2| - 2$$

are satisfied then

$$sq(w) \leq 0.9\,|w| - 2.$$

Proof
If γ_{n-1} is odd then let $\Delta = \gamma_{n-1} - 1$ otherwise let $\Delta = \gamma_{n-1} - 2$.

Consider what happens when we change γ_{n-1} by the quantity Δ (see the formula(4) from the equation (4.5)). The increase of the number of distinct squares is $\frac{1}{2}\,\Delta\, q_{n-1}$, while the increase of the length of the word is $\Delta\, q_{n-1}$. The increase of the number of squares is amortized by half of the increase of the length of the word. Therefore, we can subtract Δ from γ_{n-1} and the thesis follows.

□

Now we are ready to estimate the number of distinct squares in the standard words defined by short directive sequences.

Lemma 4.10 (Short γ)
If $n < 3$ then

$$sq(\gamma_0,\ldots,\gamma_n) \;\le\; \frac{9}{10}\big|\mathrm{Sw}(\gamma_0,\ldots,\gamma_n)\big|.$$

Proof
There are three types of short directive sequences: $\gamma^{I} = (\gamma_0)$, $\gamma^{II} = (\gamma_0,\gamma_1)$ and $\gamma^{III} = (\gamma_0,\gamma_1,\gamma_2)$. We consider each of them separately.

Case 1: $\gamma^{I} = (\gamma_0)$.
We have (due to formulas (1-6) from the equation (4.5)):

$$sq\big(\gamma^{I}\big) \le \frac{1}{2}\big(\gamma_0+1\big) \qquad \text{and} \qquad \big|\mathrm{Sw}\big(\gamma^{I}\big)\big| = \gamma_0+1.$$

Therefore,

$$sq\big(\gamma^{I}\big) \;\le\; \frac{1}{2}\Big|\mathrm{Sw}\big(\gamma^{I}\big)\Big| \;<\; \frac{9}{10}\Big|\mathrm{Sw}\big(\gamma^{I}\big)\Big|.$$

Case 2: $\gamma^{II} = (\gamma_0,\gamma_1)$.
We have (due to formulas (1-6) from the equation (4.5)):

$$\begin{aligned}
\big|\mathrm{Sw}\big(\gamma^{II}\big)\big| \;&=\; \big(\gamma_0+1\big)\gamma_1+1,\\
sq\big(\gamma^{II}\big) \;&=\; sq_0\big(\gamma^{II}\big) \;+\; sq_1\big(\gamma^{II}\big),\\
sq_0\big(\gamma^{II}\big) \;&\le\; \tfrac{1}{2}\big(\gamma_0+1\big).
\end{aligned}$$

There are two cases depending on the value of γ_1.

I: If $\gamma_1 = 1$ then $sq_1\big(\gamma^{II}\big) \;=\; 0$ and we have:

$$sq\big(\gamma^{I}\big) \le \frac{1}{2}\big(\gamma_0+1\big),$$

and consequently

$$\frac{sq\big(\gamma^{II}\big)}{\big|\mathrm{Sw}\big(\gamma^{II}\big)\big|} \;\le\; \frac{1}{2} \;-\; \frac{3}{2\gamma_0+4} \;\le\; \frac{9}{10}.$$

II: If $\gamma_1 > 1$ then

$$sq_1\big(\gamma^{II}\big) \;\le\; \frac{1}{2}\big(\gamma_0+1\big)\big(\gamma_1-1\big)+2.$$

We have:

$$sq\big(\gamma^{II}\big) \;\le\; \frac{1}{2}\big(\gamma_0+1\big)\gamma_1+2,$$

and finally

$$\frac{sq\big(\gamma^{II}\big)}{\big|\mathrm{Sw}\big(\gamma^{II}\big)\big|} \;\le\; \frac{1}{2} \;+\; \frac{3}{2\big(\gamma_0+1\big)\gamma_1+2} \;\le\; \frac{9}{10}.$$

Case 3: $\gamma^{III} = (\gamma_0, \gamma_1, \gamma_2)$.
There are two cases depending on the value of γ_2.

I: If $\gamma_2 = 1$ then we have (due to formulas (1-6) from the equation (4.5)):

$$\begin{aligned}
\big|\mathrm{Sw}\big(\gamma^{III}\big)\big| \;&=\; \big(\gamma_0+1\big)\big(\gamma_1+1\big)+1,\\
sq\big(\gamma^{III}\big) \;&=\; sq_0\big(\gamma^{III}\big) \;+\; sq_1\big(\gamma^{III}\big) \;+\; sq_2\big(\gamma^{III}\big),\\
sq_0\big(\gamma^{III}\big) \;&\le\; \tfrac{1}{2}\big(\gamma_0+1\big),\\
sq_1\big(\gamma^{III}\big) \;&\le\; \tfrac{1}{2}\big(\gamma_0+1\big)\big(\gamma_1+1\big),\\
sq_2\big(\gamma^{III}\big) \;&=\; 0.
\end{aligned}$$

Therefore,

$$sq(\gamma_0, \gamma_1, 1) \;\le\; \frac{1}{2}\big(\gamma_0+1\big)\big(\gamma_1+2\big),$$

and finally

$$\frac{sq\big(\gamma^{III}\big)}{\big|\mathrm{Sw}\big(\gamma^{III}\big)\big|} \;\le\; \frac{1}{2} \;+\; \frac{\gamma_0}{2\big(\gamma_0+1\big)\big(\gamma_1+1\big)+2} \;\le\; \frac{9}{10}.$$

II: If $\gamma_2 > 1$ then we have (due to Lemma 4.8)

$$sq(\gamma_0, \gamma_1, \gamma_2) \;\le\; sq(\gamma_0, \gamma_1, \gamma_2-1, 1)+2,$$

and the proof is similar to the proof of Theorem 4.14.

□

The next two facts will be useful in the estimation of the number of distinct squares in standard words defined by longer directive sequences.

Remark 4.11
Let $d(i)$ be as in the equation (4.4). We have

$$d(i) \quad \leq \begin{cases} \left(\frac{3}{2}\,\gamma_i - 1\right)\, q_i + q_{i-1} - 1 & \text{for even } \gamma_i \\ \left(\frac{3}{2}\,\gamma_i - \frac{1}{2}\right)\, q_i & \text{for odd } \gamma_i \end{cases} .$$

Lemma 4.12
For $0 \leq r \leq n-3$ *we have*

$$\sum_{i=0}^{r} d(i) \;<\; \frac{3}{2}\, q_{r+1} + q_r - 2.$$

Proof
Recall that $q_i = |x_i|$ (see the equation (1.1) and (4.1)). According to the observation above and the implication

$$\gamma_i \geq 2 \quad \Rightarrow \quad q_i - q_{i+1} < -\frac{1}{2}\, q_{i+1},$$

we have:

$$d(i) \leq \frac{3}{2}\,\gamma_i\, q_i - \frac{1}{2}\, q_i.$$

Observe now that $\gamma_i\, q_i \;=\; q_{i+1} - q_{i-1}$. Therefore,

$$\begin{aligned} \sum_{i=0}^{r} \gamma_i\, q_i \quad &= \quad q_{r+1} + q_r - q_0 - q_{-1} \\ &= \quad q_{r+1} + q_r - 2. \end{aligned}$$

Consequently

$$\begin{aligned} \sum_{i=0}^{r} d(i) \quad &< \quad \frac{3}{2} \sum_{i=0}^{r} \gamma_i\, q_i \; - \frac{1}{2}\, q_r \\ &\leq \quad \frac{3}{2}\left(q_{r+1} + q_r - 2\right) - \frac{1}{2}\, q_r \qquad (4.13) \\ &\leq \quad \frac{3}{2}\, q_{r+1} + q_r - 2. \end{aligned}$$

This completes the proof. □

Now we are ready to prove the tight bound for the number of distinct squares in the class $\mathcal{S}$ of standard Sturmian words.

Theorem 4.14
Let $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *be a standard Sturmian word. Then*

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \;\leq\; \frac{9}{10} \cdot \big|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)\big|.$$

Proof
If $n < 3$ then the thesis follows from Lemma 4.10. Therefore, we can assume that $n \geq 3$.

We start with the assumption that:

$$\gamma_n = 1 \qquad \text{and} \qquad \gamma_{n-1} \in \{1, 2\}.$$

Let us shorten the notation and denote:

$$A = q_{n-2}, \qquad B = q_{n-3}, \qquad \alpha = \gamma_{n-2}.$$

Claim 4.15
Recall that $q_i = |x_i|$. Due the equation (1.1) we have $A > B$ and the values of A and B increase exponentially. The smallest growth of A and B we have for the Fibonacci words, where A and B are consecutive Fibonacci numbers. For other standard Sturmian words the growth of A and B is significantly larger and the difference between A and B is bigger.

Lemma 4.12 can be rewritten in terms of A and B as follows:

Claim 4.16

$$\sum_{i=0}^{n-3} sq_i(\gamma) \;=\; \sum_{i=0}^{n-3} d(i) \;\leq\; \frac{3}{2}\, A + B - 2.$$

This, together with the fact that $sq_n(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}, 1) = 0$, implies:

Claim 4.17

$$sq(\gamma) \;\leq\; \Phi(\gamma) \;\overset{def}{=}\; \frac{3}{2}\, A + B - 2 + sq_{n-1}(\gamma) + sq_{n-2}(\gamma).$$

Our goal is to prove the inequality

$$\Phi(\gamma) \leq \frac{9}{10}\, |w| - 2. \tag{4.18}$$

Using our terminology we can write:

(a) $$\big|\mathrm{Sw}(\gamma)\big| \;=\; \begin{cases} 2\,\alpha\,A + A + 2B & \text{for } \gamma_{n-1} = 1 \\ 3\,\alpha\,A + A + 3B & \text{for } \gamma_{n-1} = 2 \end{cases},$$

(b) $$sq_{n-2}(\gamma) \;\leq\; \begin{cases} \frac{3}{2}\,\alpha\,A - A & \text{for even } \gamma_{n-2} \\ \frac{3}{2}\,\alpha\,A - \frac{3}{2}A + B + 1 & \text{for odd } \gamma_{n-2} \end{cases},$$

(c) $$sq_{n-1}(\gamma) \;\leq\; \begin{cases} \alpha\,A + B & \text{for } \gamma_{n-1} = 2 \\ A - 1 & \text{for } \gamma_{n-1} = 1 \end{cases}.$$

We have to consider 4 cases depending on the value of $\gamma_{n-1} \in \{1, 2\}$ and the parity of α.

Case 1: ($\gamma_{n-1} = 1$, α is even)

In this case the inequality $\Phi(\gamma) \leq \frac{9}{10}\,|w|$ reduces to:

$$\frac{3}{2}\Big(\alpha + 1\Big)\;A + B \;\leq\; \frac{9}{10}\,\Big((2\,\alpha + 1)\;A + 2\;B\Big).$$

This reduces to:

$$(3\alpha - 6)A + 8B \geq 0,$$

which obviously holds for $\alpha \geq 2$.

This completes the proof of this case.

Case 2: ($\gamma_{n-1} = 1$, α is odd)

In this case the inequality $\Phi(\gamma) \leq \frac{9}{10}\,|w| - 2$ reduces to:

$$\left(\frac{3}{2}\,\alpha + 1\right)\;A + 2\;B \leq\; \frac{9}{10}\,\Big((2\;\alpha + 1)\;A + 2\;B\Big).$$

This reduces to

$$(3\alpha - 1)A \geq 2B,$$

which holds since $\alpha \geq 1$ and due to Claim 4.15.

This completes the proof of this case.

Case 3: ($\gamma_{n-1} = 2$, α is even)

In this case

$$\Phi(\gamma) \leq \left(\frac{5}{2}\,\alpha + \frac{1}{2}\right) A + 2\,B - 2.$$

Consequently, the inequality $\Phi(\gamma) \leq \frac{9}{10}\,|w| - 2$ reduces to:

$$\left(\frac{5}{2}\,\alpha + \frac{1}{2}\right) A + 2\,B \;\leq\; \frac{9}{10}\,\Big(3\,\alpha\,A + A + 3B\Big).$$

This reduces to

$$(2\alpha + 4)A + 7B \geq 0$$

and holds since $\alpha \geq 2,\;\, A > B > 0$.

Case 4: ($\gamma_{n-1} = 2$, α is odd)

In this case

$$\Phi(\gamma) \;\leq\; \frac{5}{2}\,\alpha\,A + 3\,B - 1.$$

Now the inequality $\Phi(\gamma) \leq \frac{9}{10}\,|w| - 2$ reduces to:

$$\frac{5}{2}\,\alpha\,A + 3\,B + 1 \;\leq\; \frac{9}{10}\,\Big(3\,\alpha\,A + A + 3B\Big).$$

This reduces to

$$3B + 10 \leq (2\alpha + 9)A,$$

which holds since $\alpha \geq 1$ and due to Claim 4.15.

We have proved that

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 1, 1) \;\leq\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 1, 1)| - 2$$

and

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 2, 1) \;\leq\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 2, 1)| - 2.$$

Due to Lemma 4.8 and Lemma 4.9 we have

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \;\leq\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)|,$$

which completes the proof of the theorem.

□

4.4 Squares vs. maximal repetitions

Recall form the chapter 3 that the *maximal repetition* (the *run* in short) in a word w is a nonempty subword $w[i..j] = u^k v$ ($k \geq 2$), 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 $|u|$).

Let $\rho(w)$ be the number of runs in the word w. For the n-th Fibonacci word F_n we have:

$$\begin{aligned} sq(F_n) &= 2|F_{n-2}| - 2, \\ \rho(F_n) &= 2|F_{n-2}| - 3, \end{aligned}$$

hence $sq(F_n) = \rho(f_n) + 1$ and consequently $\frac{sq(F_n)}{|F_n|}$ and $\frac{\rho(F_n)}{|F_n|}$ have the same asymptotic behaviour, see [28], [48].

For standard Sturmian words the situation is quite different. We have:

$$\frac{\rho(w)}{|w|} \longrightarrow 0.8 \qquad \text{and} \qquad \frac{sq(w)}{|w|} \longrightarrow 0.9,$$

see the chapter 3 for details. Below we will investigate 3 different sequences of standard Sturmian words to see that the number of squares and the number of runs are not so closely related as in case of Fibonacci words.

Case 1: $w_k = \mathrm{Sw}(k,k,2,1,1)$.
The word w_k has the form:

$$w_k = \Big((a^k b)^k a\Big)^2 a^k b \Big((a^k b)^k a\Big)^3 a^k b$$

and length

$$|w_k| = 5k^2 + 7k + 7.$$

In the section 4 we have computed the number of squares for the word w_k, and we have seen that

$$\frac{sq(w_k)}{|w_k|} \longrightarrow \frac{9}{10}.$$

Now we compute the number of runs in w_k using the formula from the equation (3.10). We have:

$$\rho(w_k) = 9k + 7$$

and therefore

$$\frac{\rho(w_k)}{|w_k|} \longrightarrow 0.$$

We can see that w_k is an example of a word that is rich in squares and at the same time has a very small number of runs.

Case 2: $v_k = \mathrm{Sw}(1,2,k,k)$.
The word v_k has the form

$$v_k \;=\; \Big((ababa)^k ab\Big)^k ababa,$$

and length

$$|v_k| \;=\; 5k^2+2k+5.$$

Using the formula from the equation (3.10) we obtain

$$\rho(v_k) \;=\; 4k^2-k+3,$$

hence

$$\frac{\rho(v_k)}{|v_k|} \;\longrightarrow\; \frac{8}{10}.$$

Using the formulas (1-6) from the equation (4.5) we have:

$$sq(v_k) \;=\; \begin{cases} \frac{5}{2}k^2+\frac{5}{2}k+4 & \text{for even } \; k \\ \frac{5}{2}k^2+5k-\frac{5}{2} & \text{for odd } \; k \end{cases}$$

and consequently

$$\frac{sq(v_k)}{|v_k|} \;\longrightarrow\; \frac{1}{2}.$$

We can see that v_k is an example of a word, for which the number of squares is significantly smaller than the number of runs.

Case 3: $z_k = \mathrm{Sw}(1,2,k,k,2,1,1)$.
The word z_k has the form

$$z_k \;=\; \Big[\Big((ababa)^k ab\Big)^k ababa\Big]^2 (ababa)^k ab \Big[\Big((ababa)^k ab\Big)^k ababa\Big]^3 (ababa)^k ab$$

and length

$$|z_k| \;=\; 25k^2+20k+29.$$

Using the formulas (1-6) from the equation (4.5) we compute the number of squares:

$$sq(z_k) \;=\; \begin{cases} \frac{45}{2}k^2+\frac{19}{2}k+17 & \text{for even } \; k \\ \frac{45}{2}k^2+12k+\frac{31}{2} & \text{for odd } \; k \end{cases}$$

and consequently

$$\frac{sq(z_k)}{|z_k|} \;\longrightarrow\; \frac{9}{10}.$$

Using the formula from the equation (3.10) we compute the number of runs

$$\rho(z_k) \;=\; 20k^2 + 11k + 20,$$

hence

$$\frac{\rho(z_k)}{|z_k|} \;\longrightarrow\; \frac{8}{10}.$$

We can see that z_k is an example of a word, for which both the number of distinct squares and the number of runs are high.

The results above show that the maximal number of squares and the maximal number of runs for standard Sturmian words are not closely related. The asymptotic limits are close, but for different types of words the number of squares and the number of runs could have different asymptotic behaviour.

4.5 Repetitions in Christoffel words

Recall the geometric definition of the class of Christoffel words presented in the section 1.2. These words are strongly related to standard words. Due to Fact 1.10, every Christoffel word is the cyclic shift of some standard word of the one position. Therefore, the results related to the number of runs from the chapter 3 and to the number of distinct squares from the chapter 4 could be simply extended to the class of Christoffel words.

Lemma 4.19
Let w be a Christoffel word of the length n.

(1) *We have $\rho(w) \;\leq\; 0.8\, n + \log(n)$.*

(2) *There is an infinite sequence $\{w_k\}$ of strictly growing Christoffel words, such that*

$$\lim_{k\to\infty} |w_k| \;=\; \infty \qquad \text{and} \qquad \lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} \;=\; 0.8.$$

(3) *We have $sq(w) \;\leq\; 0.9\, n + \log(n)$.*

(4) *There is an infinite sequence $\{v_k\}$ of strictly growing Christoffel words, such that*

$$\lim_{k\to\infty} |v_k| \;=\; \infty \qquad \text{and} \qquad \lim_{k\to\infty} \frac{sq(v_k)}{|v_k|} \;=\; 0.9.$$

Proof
For each position i in a word w there is a logarithmic number of runs that begins (respectively ends) at the position i in w.

Let w_S be a standard word. The corresponding Christoffel word w_C is formed from w_S by moving its last letter from the end to the begin of the word.

The removing of the last letter of w_S causes the possible disappearance of the logarithmic number of runs (squares respectively) in w_C ending at this position. On the other hand, the new letter at the beginning of the word w_C can be the starting position of the logarithmic number of new runs (squares respectively).

The construction of the strictly growing sequences of Christoffel words that achieve asymptotically the maximal ratio of the number of runs (squares respectively) to the length of the word could be done in similar way as for the standard words.

□

5

Numeration systems related to Sturmian words

In this chapter we will consider the numeration systems connected to the structure of compacted subword graphs of standard words (see the chapter 2).

A numeration system is a way of expressing numbers as a sequences of digits. Therefore, numbers can be seen as finite words over an alphabet of digits and are field of interest of combinatorics on words.

More formally, let $B \geq 2$ be an integer number called the *base*. The B-ary representation of the integer number $N \geq 0$ is the finite word $d_k \dots d_0$ over the alphabet $A = \{0, 1, \dots, B-1\}$, such that

$$N = \sum_{i=0}^{k} d_i B^i.$$

If we omit the leading zeroes such a representation is unique. The set of all representations of the positive integers equals A^*. An example of such a representation is the decimal numerations system – the most natural way of representing integers.

The notion of a numeration system can be generalized by replacing the base by the infinite strictly increasing sequence of integers $Q = (q_n)_{n \geq 0}$, with $q_0 = 1$, called the *base sequence*. A representation of an integer number N in this system is the finite sequence of integers $(d_0, \dots, d_k)$, where

$$N = \sum_{i=0}^{k} d_i q^i.$$

The representation of the number 0 is the empty word ε.

Example 5.1
Let $Q = \{2^n : n \geq 0\}$ be the base sequence. The representation of a number N in the system defined by Q is simply its binary representation.

□

For more information on numeration system see for instance the chapter 7 of [51] or the chapter 3 of [1].

5.1 The Ostrowski numeration systems

The dual Fibonacci numeration system has been introduced in [67], where its relation to the subword structure of Fibonacci words has been investigated. We extend these results to standard Sturmian words.

For an infinite directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$ we introduce $[*]_\gamma$-numeration system: a version of the Ostrowski's numeration system described in [1], which is a generalization of the Fibonacci numeration system. Let us define the *base* sequence as:

$$Q \;=\; (q_0, q_1, \ldots) \;=\; \big(|x_0|, |x_1|, ...\big),$$

where x_i's are standard words given by the equation (1.1).

The base sequence Q can be defined without any reference to standard words as follows:

$$q_{-1} \;=\; q_0 \;=\; 1, \qquad \text{and} \qquad q_{i+1} \;=\; q_i \cdot \gamma_i + q_{i-1} \quad \text{for } i > 0.$$

Example 5.2
For $\gamma = (1, 2, 1, 2, \ldots)$ the base sequence is $Q = (1, 2, 5, 7, 19, \ldots)$.
For $\gamma = (1, 2, 1, 1, 1, \ldots)$ the base sequence is $Q = (1, 2, 5, 7, 12, 19, \ldots)$.

□

The representation of an integer N in the *Ostrowski* numeration system is defined as:

$$[\, N \,]_\gamma \;=\; (d_0, d_1, \ldots, d_n),$$

where we require:

(1) $N \;=\; d_0 \cdot q_0 + d_1 \cdot q_1 + \ldots + d_n \cdot q_n,$

(2) $\forall_{0 \leq j \leq n}\; d_j \leq \gamma_j,$

(3) $d_{j+1} = \gamma_{j+1} \implies d_j = 0.$

In other words, in the representation of a number N, for each k we take at most γ_k numbers $|x_k|$, and if we take exactly γ_k numbers $|x_k|$, then we take zero numbers $|x_{k-1}|$.

The uniqueness of the representation in the Ostrowski numeration system has been proved in [1].

Example 5.3

Let $\gamma = (1, 2, 1, 3, 1, \ldots)$. In this case the base sequence is

$$q = \big(|x_0|, |x_1|, ...\big) = (1, 2, 5, 7, 26, 33, \ldots).$$

We have $[16]_\gamma = (0, 1, 0, 2)$, because $16 \; = \; 0 \cdot 1 + 1 \cdot 2 + 0 \cdot 5 + 2 \cdot 7$.
We have $[29]_\gamma = (1, 1, 0, 0, 1)$, because $29 = 1 \cdot 1 + 1 \cdot 2 + 0 \cdot 5 + 0 \cdot 7 + 1 \cdot 26$.

□

The representation of the number N in the *dual Ostrowski* numeration system is defined as:

$$[\,\hat{N}\,]_\gamma = (d_0, d_1, \ldots, d_n),$$

where we require:

(1) $N \; = \; d_0 \cdot q_0 + d_1 \cdot q_1 + \ldots + d_n \cdot q_n$,

(2) $\forall_{0 \le j \le n} \; d_j \le \gamma_j$,

(3) $\big(d_j < \gamma_j$ and $\exists_{(i>j)} \; d_i > 0\;\big) \implies d_{j+1} > 0$.

In other words, in the representation of a number N in the numeration system defined above, for each k, we take at most γ_k numbers $|x_k|$ and if we take $d_k < \gamma_k$ numbers $|x_k|$ and d_k is not the last component of this representation, then we must take at least one number $|x_{k+1}|$.

For the proof of the uniqueness of the representation in the dual Ostrowski numeration system we refer the reader to [26].

Example 5.4

Let $\gamma = (1, 2, 1, 3, 1, \ldots)$. In this case the base sequence is

$$q = \big(|x_0|, |x_1|, ...\big) = (1, 2, 5, 7, 26, 33, \ldots).$$

We have $[\hat{16}]_\gamma = (0, 2, 1, 1)$, because $16 \; = \; 0 \cdot 1 + 2 \cdot 2 + 1 \cdot 5 + 1 \cdot 7$.
We have $[\hat{29}]_\gamma = (1, 1, 1, 3)$, because $29 = 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 5 + 3 \cdot 7$.

□

The relation to subword graphs of standard words

Let $\mathcal{G}_\infty$ be the infinite compacted subword graph of the standard word given by the infinite directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$. Let π be a path from the root to another node of $\mathcal{G}_\infty$ and let $\text{rep}(\pi) = (h_0, h_1, \ldots)$, where h_i is the number of the edges of the weight q_i on the path π.

The following fact is the interpretation of the corresponding result from [26] in terms of the dual Ostrowski numeration system.

Fact 5.5
Let the directive sequence γ, the graph $\mathcal{G}_\infty$, the path π and the sequence $\text{rep}(\pi)$ be defined as above.

(1) *The sequence $\text{rep}(\pi)$ is the representation of the length of the path π in the dual Ostrowski numeration system corresponding to γ.*

(2) *For each $k > 1$ there is exactly one fork-path of the length k in $\mathcal{G}_\infty$.*

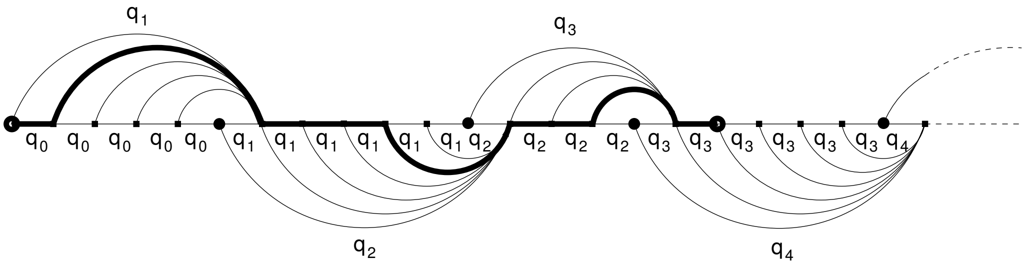
q1
q3
q0 q0 q0 q0 q0 q1 q1 q1 q1 q1 q1 q2 q2 q2 q2 q3 q3 q3 q3 q3 q3 q4
q2
q4

Figure 5.1: *The illustration of the point (1) of Fact 5.5. In this case the representation of the length of the path π in the dual Ostrowski numeration system is given by:* $\text{rep}(\pi) = (1, 4, 3, 2)$ *and* $|\pi| = 1{\cdot}|q_0|{+}4{\cdot}|q_1|{+}3{\cdot}|q_2|{+}2{\cdot}|q_3|$.

Proof
Point (1)
Let π be a path from the root to some internal node v in $\mathcal{G}_\infty$ – an infinite compacted subword graph corresponding to the standard word given by the directive sequence $\gamma = (\gamma_0, \gamma_1, \gamma_2, \ldots)$, and let $\text{rep}(\pi) = (h_0, h_1, \ldots)$ be defined as above. It is sufficient to prove that all requirements of the definition of the dual Ostrowski numeration system are satisfied.

The construction of the path π implies that

$$|\pi| \;=\; h_0 \cdot q_0 + h_1 \cdot q_1 + h_2 \cdot q_2 + \;\ldots \qquad \text{and} \qquad \forall_i \; 0 \le h_i \le \gamma_i.$$

Moreover, from the structure of $\mathcal{G}_\infty$ (see Figure 5.1), it is obvious that if for some i there is $h_i < \gamma_i$ (we have taken q_i less than γ_i times) and h_i is not the last non zero component of rep(π) then $h_{i+1} > 0$ (we must take at least one q_{i+1} to continue the construction of the path π). This concludes the proof of the point (1).

Point (2)
The thesis follows directly from the point (1) and the uniqueness of the representation in the dual Ostrowski numeration system. □

5.2 S-language and S-automaton

The S-language and the S-automaton are ideas related to the dual Ostrowski numeration system discussed in the previous section, but can be also defined independently. These objects were for the first time described in [6].

Recall that $|w|_a$ denotes the number of occurrences of the letter a in the word w. For a directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ and the alphabet $A = \{a_0, \ldots, a_n\}$ we define the S-language $L = \text{S-lan}(\gamma)$ as follows:

- if $\gamma_n = 1$ then L is the set of all subsequences u of the word $a_0^{\gamma_0} a_1^{\gamma_1} \ldots a_n^{\gamma_n}$, which satisfy:
  - $|u|_{a_n} \; = \; 1$,
  - $\forall_{0<i<n} \; |u|_{a_i} < \gamma_i \implies |u|_{a_{i+1}} > 0$,
- if $\gamma_n > 1$ then $L \; = \; \text{S-lan}(\gamma_0, \gamma_1, \ldots \gamma_{n-1}, \gamma_n - 1, 1)$.

The S-automaton, denoted by S-aut(γ), is defined as the minimal deterministic automaton accepting the language S-lan(γ). In the automaton S-aut(γ) we exclude the *dead state* – the nonaccepting state, which *loops itself* (each transition from this state goes back to itself). The missing edges in the graph of the automaton are assumed to go to the *dead* state.

Recall that a word $z \in \{a, b\}^*$ is a special prefix of a word $w \in \{a, b\}^*$ if both za and zb are subwords of w. Recall also the structure of compacted subword graphs of standard words (see the chapter 2). The following fact is a direct implication of Fact 2.2.

Fact 5.6
The minimal S-automaton (without the dead state) for a directive sequence γ is isomorphic as a graph with the compacted directed acyclic subword graph of the standard Sturmian word Sw(γ).

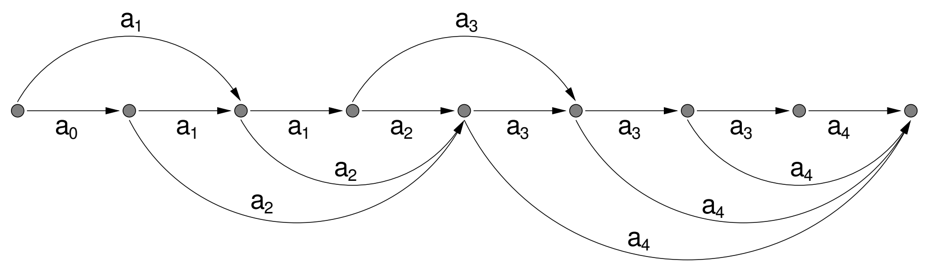

Figure 5.2: *The S-automaton (the minimal deterministic automaton, without the dead state) S-aut*$(1,2,1,3,1)$*. The only accepting state is the* sink *node.*

A prefix u of a word w is called *maximal* if u is not a proper prefix of another prefix of w. Recall that the *basic subword* y_k is defined as the reverse of x_k, where x_k is as in the equation (1.1), and $\hat{w}$ – as the prefix of w of the size 2.

For the directive sequence $\gamma = (\gamma_0, \dots, \gamma_n)$ and the alphabet $A = \{a_0, \dots, a_n\}$ we define the following morphism h_γ:

- If $\gamma_n = 1$ then $h_\gamma(a_i) = y_i$, for $0 \le i < n$ and $h_\gamma(a_n) = \hat{y}_n$.
- If $\gamma_n > 1$ then $h_\gamma(a_i) = y_i$, for $0 \le i \le n$, and $h_\gamma(a_{n+1}) = \hat{y}_{n+1}$.

The morphic image of a language is meant in the usual sense and the morphic image of an automaton results by changing the label of each edge of this automaton using this morphism.

The following results are implied by Fact 2.2 and Fact 2.6.

Fact 5.7

Let γ *be a directive sequence and* $\mathrm{Sw}(\gamma)$ *be a standard word.*

(1) *The set of maximal prefixes of* $\mathrm{Sw}(\gamma)$ *equals* $h_\gamma\big(S\text{-}lan(\gamma)\big)$ *(it is the morphic image of the S-language for* γ*).*

(2) *The compacted subword graph of* $\mathrm{Sw}(\gamma)$ *is the image of the S-automaton S-aut*(γ) *under the morphism* h_γ*.*

Conclusions and final remarks

The aim of the thesis was to study some problems related to repetitions and the combinatorial structure for one of the most thoroughly investigated class of strings in combinatorics on words – the standard Sturmian words.

The detailed analysis of the subword graphs structure of those words, done in the chapter 2, leads to simple alternative graph-based proofs of several known facts and to special easy algorithms computing some properties of standard words. It also implies an interesting interpretation of the representation of integer numbers in the dual Ostrowski numeration system.

The matter of the chapters 3 and 4 was the investigation of the structure of repetitions in standard words. We have presented the formulas for the numbers of runs and distinct squares along with the detailed analysis of their asymptotic behaviour. The complete understanding of their combinatorial structure for a large class of complicated words is a step towards a better understanding of this problem in general.

The maximal repetition ratio 0.8 and the square ratio 0.9 for standard words has been first discovered by us doing experiments with very long strings. Similarly, we were tuning many intermediate formulas with the assistance of the computer. Our algorithms for computing the number of runs and and the number of distinct squares in standard words are examples of the very fast computation on highly compressed texts in linear time with respect to the size of their compressed representation.

For the Fibonacci words the number of distinct squares is only one more than the number of runs. The results of this thesis show that those numbers are not so closely related in general. In case of well structured words (Sturmian words) the *density ratio* of the distinct squares (the asymptotic quotient of the maximal number of squares by the length of the string) and the *density ratio* of the maximal repetitions are close, however both limits could be reached for different types of words.

Bibliography

[1] J. Allouche and J. Shallit. *Automatic Sequences. Theory, Applications, Generalizations.* Cambridge University Press, 2003.

[2] A. Apostolico and F. P. Preparata. Optimal off-line detection of repetitions in strings. *Theoretical Computer Science*, 22:297–315, 1983.

[3] J. D. Barrow. Chaos in numberland: The secret life of continued fractions. http://plus.maths.org/issue11/features/cfractions, 2000.

[4] P. Baturo, K. Czarkowski, M. Pilichowski, **M. Piątkowski**, and W. Rytter. Suffix arrays of Fibonacci words and lexicographic properties of the Fibonacci number system. In *Proceedings of the Conference on Combinatorics, Automata and Number Theory.* University of Liege, 2006.

[5] P. Baturo, **M. Piątkowski**, and W. Rytter. The number of runs in Sturmian words. In *Proceedings of the 13th international conference on Implementation and Applications of Automata*, volume 5148 of *Lecture Notes in Computer Science*, pages 252–261. Springer, 2008.

[6] P. Baturo, **M. Piątkowski**, and W. Rytter. Usefulness of directed acyclic subword graphs in problems related to standard Sturmian words. *International Journal of Foundations of Computer Science*, 20(6):1005–1023, 2009.

[7] P. Baturo and W. Rytter. Compressed string-matching in standard Sturmian words. *Theoretical Computer Science*, 410(30–32):2804–2810, 2009.

[8] J. Bernoulli. Recueil pour les astronomes. *Sur une nouvelle espece de calcul*, I:255–284, 1772. Berlin.

[9] J. Berstel. Sturmian and Episturmian words: a survey of some recent results. In *Proceedings of the 2nd international conference on Algebraic informatics*, volume 4728 of *Lecture Notes in Computer Science*, pages 23–47. Springer, 2007.

[10] J. Berstel and J. Karhumaki. Combinatorics on words: a tutorial. *Bulletin of the EATCS*, 79:178–228, 2003.

[11] 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.

[12] J. Berstel and D. Perrin. The origins of combinatorics on words. *European Journal of Combinatorics*, 28(3):996–1022, 2007.

[13] S. Brlek, J. O. Lachaud, X. Provençal, and C. Reutenauer. Lyndon + Christoffel = digitally convex. *Pattern Recognition*, 42(10):2239–2246, 2009.

[14] A. Cayley. Desiderata and suggestions: No. 2. the theory of groups: graphical representation. *American Journal of Mathematics*, 2:174–176, 1878.

[15] E. B. Christoffel. Observatio arithmetica. *Mathematische Annalen*, 6:145–152, 1875.

[16] E. B. Christoffel. Lehrsätze über arithmetische eigenshaften du irrationnalzahlen. *Annali di Mathematica Pura ed Applicata, Series II*, 15:253–276, 1888.

[17] M. Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12(5):244–250, 1981.

[18] M. Crochemore and L. Ilie. Analysis of maximal repetitions in strings. In *Proceedings of the 32nd International Conference on Mathematical Foundations of Computer Science*, volume 4708 of *Lecture Notes in Computer Science*, pages 465–476. Springer, 2007.

[19] M. Crochemore and L. Ilie. Maximal repetitions in strings. *Journal of Computer and System Sciences*, 74(5):796–807, 2008.

[20] 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*, volume 5029 of *Lecture Notes in Computer Science*, pages 290–302. Springer, 2008.

[21] M. Crochemore and D. Perrin. Two-way string matching. *Journal of ACM*, 38(3):651–673, 1991.

[22] M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13(5):405–425, 1995.

[23] M. Crochemore and W. Rytter. *Jewels of Stringology: text algorithms.* World Scientific, 2003.

[24] D. Damanik and D. Lenz. Powers in Sturmian sequences. *European Journal of Combinatorics*, 24(4):377–390, 2003.

[25] I. Debled-Rennesson. Éléments de géométrie discrete: Vers une étude des structures descrètes bruitées. Habilitation à diriger des recherches, Université Henri Poincaré, Nancy I, 2007.

[26] C. Epifanio, F. Mignosi, J. Shallit, and I. Venturini. Sturmian graphs and the conjecture of Moser. In *Proceedings of the 8th International Conference on Developments in Language Theory*, volume 3340 of *Lecture Notes in Computer Science*, pages 175–187. Springer, 2004.

[27] A. S. Fraenkel and J. Simpson. How many squares can a string contain? *Journal of the Combinatorial Theory Series A*, 82:112–120, 1998.

[28] A. S. Fraenkel and J. Simpson. The exact number of squares in Fibonacci words. *Theoretical Computer Science*, 218(1):95–106, 1999.

[29] F. Franek, A. Karaman, and W. F. Smyth. Repetitions in Sturmian strings. *Theoretical Computer Science*, 249(2):289–303, 2000.

[30] 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*, pages 26–35, 2003.

[31] 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):195–203, 2008.

[32] M. Giraud. Not so many runs in strings. In *Proceedings of the Second International Conference on Language and Automata Theory and Applications*, volume 5196 of *Lecture Notes in Computer Science*, pages 232–239. Springer, 2008.

[33] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Adison-Wesley, 2006.

[34] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[35] T. Harju and D. Nowotka. Density of critical factorizations. *Informatique Théoretique et Applications*, 36(3):315–327, 2002.

[36] G. Held and T. Marshall. *Data compression: Techniques and Applications – Hardware and Software Considerations*. John Wiley and Sons, 1991.

[37] D. Hensley. *Continued Fractions*. World Scientific, 2006.

[38] L. Ilie. A simple proof that a word of length n has at most $2n$ distinct squares. *Journal of Combinatorial Theory Series A*, 112:163–164, 2005.

[39] L. Ilie. A note on the number of squares in a word. *Theoretical Computer Science*, 380:373–376, 2007.

[40] C. S. Iliopoulos, D. Moore, and W. F. Smyth. A characterization of the squares in a Fibonacci string. *Theoretical Computer Science*, 172(1–2):281–291, 1997.

[41] N. C. Jones and P. A. Pevzner. *An Introduction to Bioinformatics Algorithms*. The MIT Press, 2004.

[42] W. B Jones and W. J. Thorn. *Continued Fractions. Analytic Theory and Applications*, volume 11 of *Encyclopedia of Mathematics and its Applications*. Adison-Wesley, 1980.

[43] A. Karaman and W. F. Smyth. A representation of Sturmian strings. Tech. Rep. CAS 98-04, Department of Computing & Software, McMasters University, 1998.

[44] J. Karhumaki. Combinatorics on words (notes in pdf). http://www.math.utu.fi/en/home/karhumak/combwo.pdf.

[45] C. O. Kisselman. Digital geometry and mathematical morphology. Lecture notes, Uppsala University, http://www.math.uu.se/ kiselman, 2004.

[46] D. E. Knuth. *The art of computer programming.* Addison-Wesley, 1997.

[47] R. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 596–604. IEEE Computer Society, 1999.

[48] R. M. Kolpakov and G. Kucherov. On maximal repetitions in words. In *Proceedings of 12th International Symposium on Fundamentals of Computation Theory*, volume 1684 of *Lecture Notes in Computer Science*, pages 374–385. Springer, 1999.

[49] 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.

[50] M. Lothaire. *Combinatorics on Words.* Cambridge University Press, 1997.

[51] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of mathematics and its application.* Cambridge University Press, 2002.

[52] M. Lothaire. *Applied Combinatorics on Words*, volume 105 of *Encyclopedia of Mathematics and its Application.* Cambridge University Press, 2005.

[53] W. Magnus, A. Karrass, and D. Solitar. *Combinatorial Group Theory: Presentations of Groups in Terms of Generators and Relations*, volume XIII of *Pure and Applied Mathematics.* Interscience Publishers, 1966.

[54] M. G. Main. Detecting leftmost maximal periodicities. *Discrete Applied Mathematics*, 25(1–2):145–153, 1989.

[55] M. G. Main and J. Lorentz. An $o(n \log n)$ algorithm for finding all repetitions in a string. *Journal of Algorithms*, 5(3):422–432, 1984.

[56] A. Markoff. Sur une question de Jean Bernoulli. *Mathematische Annalen*, 19:27–36, 1882.

[57] M. Morse and G. A. Hedlund. Symbolic dynamics II. sturmian trajectories. *American Journal of Mathematics*, 62:1–42, 1950.

[58] M. Nelson and J.-I Gailly. *The Data Compression Book.* M& T Books, 1995.

[59] N. A. Parshin and I. R Shafarevich, editors. *Combinatorial Group Theory. Applications to Geometry*, volume 58 of *Encyclopaedia of Mathematical Sciences.* Springer-Verlag, 1993.

[60] **M. Piątkowski**. Stringological applets . http://www.mat.umk.pl/~martinp/stringology/applets.

[61] **M. Piątkowski** and W. Rytter. Asymptotic behaviour of the maximal number of squares in standard Sturmian words. In *Proceedings of the 14-th Prague Stringology Conference*, pages 237–248. Czech Technical University, 2009. accepted to International Journal of Foundations of Computer Science.

[62] S. J. Puglisi, J. Simpson, and W. F. Smyth. How many runs can a string contain? *Theoretical Computer Science*, 4001(1–3):165–171, 2008.

[63] N. Pytheas Fogg. *Substitutions in Dynamics, Arithmetics and Combinatorics*, volume 1794 of *Lecture Notes in Mathematics.* Springer, 2002.

[64] A. M. Rocket and P. Szüsz. *Continued Fractions.* World Scientific, 1992.

[65] W. Rytter. Grammar compression, LZ-encodings, and string algorithms with implicit input. In *Proceedings of the 31st International Colloquium on Automata, Languages and Programming*, volume 3142 of *Lecture Notes in Computer Science*, pages 15–27. Springer, 2004.

[66] W. Rytter. The number of runs in a string: Improved analysis of the linear upper bound. In *Proceedings of the 23rd Annual Symposium on Theoretical Aspects of Computer Science*, volume 3884 of *Lecture Notes in Computer Science*, pages 184–195. Springer, 2006.

[67] W. Rytter. The structure of subword graphs and suffix trees of Fibonacci words. *Theoretical Computer Science*, 363(2):211–223, 2006.

[68] W. Rytter. The number of runs in a string. *Information and Computation*, 205(9):1459–1469, 2007.

[69] K. Sayood. *Introduction to data compression.* Morgan Kaufmann Publishers, 2000.

[70] M. Sciortino and L. Zamboni. Suffix automata and standard Sturmian words. In *Proceedings of the 11th International Conference on Developments in Language Theory*, volume 4588 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2007.

[71] J. Shallit. Characteristic words as fixed points of homomorphisms. Technical Report CS-91-72, University of Waterloo, Department of Computer Science, 1991.

[72] H. J. S Smith. Note on continued fractions. *Messenger of Mathematics*, 1874.

[73] A. Thue. Über unendliche zeichenreihen. *Norske Vid. Selsk. Skr. I Math-Nat.*, Christiana 7:1–22, 1906.

[74] A. Thue. Über die gegenseitige lage gleicher teile gewisser zeichenreihen. *Norske Vid. Selsk. Skr. I Math-Nat.*, Christiana 10:1–67, 1912.

[75] H. Uscka-Wehlou. *Digital lines, Sturmian words, and continued fractions.* PhD thesis, Department of Mathematics, Upspsala University, 2009.