Overlap-Free Words and Generalizations

by

Narad Rampersad

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The study of combinatorics on words dates back at least to the beginning of the 20th century and the work of Axel Thue. Thue was the first to give an example of an infinite word over a three letter alphabet that contains no squares (identical adjacent blocks) xx. This result was eventually used to solve some longstanding open problems in algebra and has remarkable connections to other areas of mathematics and computer science as well.

This thesis will consider several different generalizations of Thue's work. In particular we shall study the properties of infinite words avoiding various types of repetitions.

In Chapter 1 we introduce the theory of combinatorics on words. We present the basic definitions and give an historical survey of the area.

In Chapter 2 we consider the work of Thue in more detail. We present various well-known properties of the Thue–Morse word and give some generalizations. We examine Fife's characterization of the infinite overlap-free words and give a simpler proof of this result. We also present some applications to transcendental number theory, generalizing a classical result of Mahler.

In Chapter 3 we generalize a result of Séébold by showing that the only infinite 7/3-power-free binary words that can be obtained by iterating a morphism are the Thue–Morse word and its complement.

In Chapter 4 we continue our study of overlap-free and 7/3-power-free words. We discuss the squares that can appear as subwords of these words. We also show that it is possible to construct infinite 7/3-power-free binary words containing infinitely many overlaps.

In Chapter 5 we consider certain questions of language theory. In particular, we examine the context-freeness of the set of words containing overlaps. We show that over a three-letter alphabet, this set is not context-free, and over a two-letter alphabet, we show that this set cannot be unambiguously context-free.

In Chapter 6 we construct infinite words over a four-letter alphabet that avoid squares in any arithmetic progression of odd difference. Our constructions are based on properties of the paperfolding words. We use these infinite words to construct non-repetitive tilings of the integer lattice.

In Chapter 7 we consider approximate squares rather than squares. We give constructions of infinite words that avoid such approximate squares.

In Chapter 8 we conclude the work and present some open problems.

Acknowledgments

I would like to express my gratitude to my advisor Jeffrey Shallit for his guidance and support over the years. I would particularly like to thank him for introducing me to the area of combinatorics on words and for providing me with many interesting research problems on which to work.

I would like to thank the members of my examining committee, Daniel Brown, Jonathan Buss, Kevin Hare, and Tero Harju, for having generously agreed to review the current work. It was a particular honour to have Tero Harju as the external examiner for this thesis.

I would also like to thank my co-authors: Boris Adamczewski, Shandy Brown, James Currie, Jui-Yi Kao, Dalia Krieger, Pascal Ochem, Manuel Silva, and Troy Vasiga. Special thanks to Dalia: it was a pleasure to have her as, not only an officemate and a colleague, but also as a friend. I also thank Nicolae Santean for many interesting discussions on automata and other subjects.

The research presented here was supported financially by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Contents

Chapter 1

Combinatorics on Words

1.1 Introduction

The study of combinatorics on words dates back at least to the beginning of the 20$^{\text{th}}$ century and the work of Axel Thue [233, 234]. Unfortunately, Thue's work was published in a relatively obscure journal and therefore remained largely unknown for several decades. Many of his results were independently rediscovered in subsequent decades, for instance, by chess master Max Euwe [91], Aršon [26], and Morse and Hedlund [179].

Thue's work on combinatorics on words was largely concerned with repetitions in words, and his primary technique for studying such repetitions was the use of iterated morphisms (see Section 1.5 below).

The study of repetitions in words has several applications; perhaps the most famous application is in the work of Novikov and Adjan [184, 185, 186, 187] in solving the Burnside problem for groups [53] (see also Hall [115] and Adjan [10]). More recently, Rivest [208] applied techniques from combinatorics on words to strengthen cryptographic hash functions against certain types of attack.

Recently, combinatorial results regarding repetitions in words have been used to prove deep results in transcendental number theory. Two good surveys on these recent developments are Waldschmidt [240] and Adamczewski and Allouche [6].

Some of the concepts of combinatorics on words, such as avoiding repetitions, have also been applied to other combinatorial structures, for instance, in the study of non-repetitive tilings of the plane (see Carpi [55], Currie and Simpson [73], or Kao, Rampersad, Shallit, and Silva [135]) or non-repetitive colourings of graphs (see Alon, Grytczuk, Haluszczak, and Riordan [25] or Kündgen and Pelsmajer [153]).

For a more in-depth treatment of the subject of combinatorics on words, see one of the following: Allouche and Shallit [23], Berstel and Perrin [42], Berstel and Karhumäki [41], Choffrut and Karhumäki [62], or Lothaire [159, 160, 161].

In the next section we give an outline of the main results of the thesis.

1.2 Thesis outline

In Chapter 2 we recall the main results of Thue concerning the construction of infinite overlap-free and squarefree words, and we present several well-known properties of the Thue–Morse word. The main results of this chapter are a simpler proof of a characterization of the infinite overlap-free words due to Fife [96] and an application of the theory of overlap-free words to transcendental number theory, generalizing a classical result of Mahler [166]. This number-theoretic result can be found in Adamczewski and Rampersad [9].

In Chapter 3 we generalize a result of Séébold [220, 222] by showing that the only infinite 7/3-power-free binary words that can be obtained by iterating a morphism are the Thue–Morse word and its complement. The proof of this result turns out to be quite intricate. This is one of the main results of the thesis and can be found in Rampersad [202].

In Chapter 4 we continue our study of overlap-free and 7/3-power-free words. The main result of this chapter is a construction of an infinite 7/3-power-free binary word containing infinitely many overlaps. Most of the work in this chapter can be found in Currie, Rampersad, and Shallit [72].

In Chapter 5 we consider certain certain language-theoretic questions concerning overlaps and overlap-free words. The main results of this chapter are that over a three-letter alphabet, the set of words containing overlaps is not context-free, and over a two-letter alphabet, the set of words containing overlaps is not unambiguously context-free. This work can be found in Rampersad [203].

In Chapter 6 we use the well-studied paperfolding words to construct infinite words over a four-letter alphabet that avoid squares in any arithmetic progression of odd difference. Using these infinite words, we construct non-repetitive tilings of the integer lattice. The results of this chapter can be found in Kao, Rampersad, Shallit, and Silva [135].

In Chapter 7 we generalize the notion of a square by considering approximate squares, and we give constructions of infinite words that avoid approximate squares. The results of this chapter can be found in Krieger, Ochem, Rampersad, and Shallit [151].

Chapter 8 concludes the work and presents some directions for future research.

This concludes our summary of the main results of the thesis. The remainder of this chapter is devoted to an overview of some of the basic notions of combinatorics on words as well as an historical survey of the area.

1.3 Words

Let Σ be a finite, nonempty set called an *alphabet*; the elements of Σ are referred to as *symbols* or *letters*. We denote the set of all finite words over the alphabet Σ

by Σ^*. We also write Σ^+ to denote the set $\Sigma^* - \{\epsilon\}$, where ϵ is the empty word. Let Σ_k denote the alphabet $\{0, 1, \ldots, k-1\}$. The length of a word w is denoted $|w|$. For $a \in \Sigma$ and $w \in \Sigma^*$, the number of occurrences of a in w is denoted by $|w|_a$.

Let $\mathbb{N}$ denote the set $\{0, 1, 2, \ldots\}$. An *infinite word* is a map from $\mathbb{N}$ to Σ. If $\mathbf{w}$ is an infinite word, we often write

$$\mathbf{w} = w_0 w_1 w_2 \cdots ,$$

where each $w_i \in \Sigma$. A *bi-infinite word* is a map from $\mathbb{Z}$ to Σ.

The set of all infinite words over the alphabet Σ is denoted Σ^ω. We also write Σ^∞ to denote the set $\Sigma^* \cup \Sigma^\omega$. If $n \geq 0$ is an integer, then Σ^n denotes the set of all words over Σ of length n. If x is a finite word, then x^ω denotes the infinite word $xxx \cdots$.

A word $w' \in \Sigma^*$ is called a *subword* of $w \in \Sigma^\infty$ if w can be written in the form $uw'v$ for some $u \in \Sigma^*$ and $v \in \Sigma^\infty$. If such a decomposition exists where $u = \epsilon$ (resp., $v = \epsilon$), then w' is called a *prefix* (resp., *suffix*) of w.

For any word $w = w_0 w_1 \cdots w_n$, we denote by w^R the *reversal* of w, namely the word $w^R = w_n w_{n-1} \cdots w_0$. For any word w over the binary alphabet $\{0, 1\}$, we denote by $\overline{w}$ the *complement* of w, namely the word obtained from w by changing 0's to 1's and 1's to 0's.

Frequently we shall deduce the existence of an infinite word with a certain property from the existence of arbitrarily large finite words with the desired property. To pass from the finite to the infinite, we shall often rely (implicitly) on the following form of a result of König [147, 148] known as the "Infinity Lemma".

Theorem 1.1 (König's Infinity Lemma). *Let A be a an infinite subset of Σ^*. There exists an infinite word $\mathbf{w}$ such that every prefix of $\mathbf{w}$ is a prefix of at least one word in A.*

Proof. There must exist a letter $w_0 \in \Sigma$ such that infinitely many words in A begin with w_0. Similarly, there must exist a letter $w_1 \in \Sigma$ such that infinitely many words in A begin with $w_0 w_1$. Continuing in this fashion, one defines an infinite word $\mathbf{w} = w_0 w_1 w_2 \cdots$ such that every prefix of $\mathbf{w}$ is a prefix of at least one word in A. □

1.4 Periodicity

A infinite word $\mathbf{w}$ is *ultimately periodic* if we can write $\mathbf{w} = uv^\omega$ for some finite words u and v. If we may take $u = \epsilon$, then we say that $\mathbf{w}$ is *purely periodic*. An infinite word is *aperiodic* if it is not ultimately periodic.

An infinite word $\mathbf{w}$ is *recurrent* if every subword x of $\mathbf{w}$ occurs infinitely often in $\mathbf{w}$. If for every subword x of $\mathbf{w}$ there exists a positive integer k such every subword of length k of $\mathbf{w}$ contains an occurrence of x, then we say that $\mathbf{w}$ is *uniformly recurrent*.

Every purely periodic word is recurrent. The Thue–Morse word defined in Section 1.5 below is an example of a recurrent word that is aperiodic.

1.5 Morphisms

The notion of a *morphism* is fundamental to combinatorics on words in general, and in particular to the problems studied in this thesis. A map $h : \Sigma^* \rightarrow \Delta^*$ is called a *morphism* if h satisfies $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$. A morphism may be specified by providing the *image words* $h(a)$ for all $a \in \Sigma$. For example, we may define a morphism $h : \{0, 1, 2\}^* \rightarrow \{0, 1, 2\}^*$ by

$$\begin{array}{rcl} 0 & \rightarrow & 01201 \\ 1 & \rightarrow & 020121 \\ 2 & \rightarrow & 0212021. \end{array} \tag{1.1}$$

This definition is easily extended to (one-sided) infinite words.

A morphism $h : \Sigma^* \rightarrow \Sigma^*$ such that $h(a) = ax$ for some $a \in \Sigma$ and $x \in \Sigma^*$ is said to be *prolongable on* a; we may then repeatedly iterate h to obtain the *fixed point*

$$h^\omega(a) = axh(x)h^2(x)h^3(x)\cdots.$$

Such words are sometimes also called *D0L words* (see Rozenberg and Salomaa [213]).

The morphism h given by (1.1) above is prolongable on 0, so we have the (infinite) fixed point

$$h^\omega(0) = 01201020121021202101201020121\cdots.$$

A morphism is *k-uniform* if $|h(a)| = k$ for all $a \in \Sigma$; it is *uniform* if it is k-uniform for some k. For example, if the morphism $\mu : \{0, 1\}^* \rightarrow \{0, 1\}^*$ is defined by

$$\begin{array}{rcl} 0 & \rightarrow & 01 \\ 1 & \rightarrow & 10, \end{array}$$

then μ is 2-uniform. This morphism is often referred to as the *Thue–Morse morphism*. Properties of this morphism will be critical to many of the results in this thesis.

The fixed point

$$\mathbf{t} = \mu^\omega(0) = 0110100110010110\cdots$$

is known as the *Thue–Morse word*. The morphism μ has another fixed point,

$$\mu^\omega(1) = 1001011001101001\cdots,$$

which is easily seen to be $\overline{\mathbf{t}}$, the complement of $\mathbf{t}$. We shall give many useful properties of the Thue–Morse word in Chapter 2.

1.6 Repetitions

Most of the work of this thesis is concerned with avoiding repetitions in words. The most basic type of repetition is the *square*, that is, a nonempty word of the form xx, where $x \in \Sigma^*$. An example of a square in English is the word `murmur`. We say a word w is *squarefree* (or *avoids squares*) if no subword of w is a square. It is easy to see that every word of length at least four over the alphabet $\{0, 1\}$ contains a square; it is therefore impossible to avoid squares in infinite binary words.

In 1906, Thue [233, Satz 5] proved the following fundamental result.

Theorem 1.2 (Thue)**.** *There exists an infinite squarefree word over an alphabet of size three.*

We shall present a proof of Theorem 1.2 in Section 2.2. In Chapter 7 we shall use the existence of infinite square-free words as a starting point for constructing words with even stronger avoidability properties.

The result of Theorem 1.2 has been independently rediscovered several times. For example, the following authors gave constructions of a squarefree ternary word: Aršon [26], Morse and Hedlund [179], Leech [155], and Istrail [130]. Berstel [35] showed that the constructions of Thue, Morse–Hedlund, and Istrail result in the same squarefree word.

Recall the morphism h defined by (1.1). Thue [234, Satz 18] (see also Pleasants [197]) showed that the infinite word $h^\omega(0)$ is squarefree. Later, Leech [155] constructed an infinite squarefree ternary word by iterating a uniform morphism.

By analogy with the definition of a square, a *cube* is a nonempty word of the form xxx, where $x \in \Sigma^*$. An *overlap* is a word of the form $axaxa$, where $a \in \Sigma$ and $x \in \Sigma^*$. An example of an overlap in English is the word `alfalfa`. A word w is *cubefree* (resp., *overlap-free*) if no subword of w is a cube (resp., overlap).

While it is impossible to avoid squares in infinite binary words, Thue [233, Satz 6] proved that the Thue–Morse word $\mathbf{t}$ defined in Section 1.5 is overlap-free, a result that was later rediscovered by Morse [178]. We shall discuss the Thue–Morse word in greater detail in Chapter 2.

For any positive integer $k \geq 2$, a *k-power* is a non-empty word of the form $xx\cdots x$ (k times), written for convenience as x^k. Thus a 2-power is a square, and a 3-power is a cube. A non-empty word that is not a k-power for any $k \geq 2$ is *primitive*. A word is *k-power-free* (or *avoids k-powers*) if none of its subwords are k-powers.

For a more general framework of pattern avoidance see Bean, Ehrenfeucht, and McNulty [31], Zimin [242], Roth [211, 212], and Cassaigne [59].

1.7 Fractional repetitions

We can generalize the notion of k-power to consider rational powers of words. Dejean [75] was the first to make such a generalization. Brandenburg [48] defined a *rational power* as follows. If α is a rational number, a word w is an *α-power* if there exist words x and x', with x' a prefix of x, such that $w = x^n x'$ and $\alpha = n + |x'|/|x|$. We refer to $|x|$ as a *period* of w. An *α^+-power* is a word that is a β-power for some $\beta > \alpha$. Note that by this definition, a square is a 2-power, and an overlap is a 2^+-power.

A word is *α-power-free* (resp., *α^+-power-free*) if none of its subwords is an β-power for some $\beta \geq \alpha$ (resp., $\beta > \alpha$).

Let us define the *repetition threshold* $\mathrm{RT}(k)$ as the infimum of all α such that there exists an infinite α-power-free word over a k-letter alphabet.

Thue [234] proved that $\mathrm{RT}(2) = 2$. Dejean [75] proved that $\mathrm{RT}(3) = 7/4$. She also conjectured that $\mathrm{RT}(4) = 7/5$ and $\mathrm{RT}(k) = k/(k-1)$ for all $k \geq 5$. Pansiot [194] proved that indeed $\mathrm{RT}(4) = 7/5$. Moulin-Ollagnier [182] verified Dejean's conjecture for $5 \leq k \leq 11$, and Mohammad-Noori and Currie [177] verified it for $7 \leq k \leq 14$. Recently, Carpi [58], in a remarkable paper, showed that Dejean's conjecture is true for all $k \geq 33$. The remaining cases are still open.

1.8 Some algorithmic results

Ehrenfeucht and Rozenberg [85] proved that if $\mathbf{w}$ is the infinite fixed point of a morphism h, then, given h, it is algorithmically decidable whether $\mathbf{w}$ is k-power-free for some positive integer k.

Berstel [34] proved that over a ternary alphabet, it is decidable whether $\mathbf{w}$ is squarefree, and Karhumäki [136] proved that over a binary alphabet, if is decidable whether $\mathbf{w}$ is cubefree. The strongest result in this regard is due to Mignosi and Séébold [176]:

Theorem 1.3 (Mignosi and Séébold). *If $\mathbf{w}$ is the infinite fixed point of a morphism h, then given h and a positive integer $k \geq 2$, it is algorithmically decidable whether $\mathbf{w}$ is k-power-free.*

Recent work of Krieger [150] generalizes some of this work to the case of fractional powers.

Pansiot [195] and Harju and Linna [118] independently proved the following result regarding the periodicity of fixed points. Honkala [123] gave a somewhat simpler proof.

Theorem 1.4 (Pansiot; Harju and Linna). *If* $\mathbf{w}$ *is the infinite fixed point of a morphism* h*, then given* h*, it is algorithmically decidable whether* $\mathbf{w}$ *is ultimately periodic.*

1.9 Repetition-free morphisms

Let α be a rational number, $\alpha > 1$. A morphism $h : \Sigma^* \to \Delta^*$ is said to be α-power-free (resp., α^+-power-free) if $h(w)$ is α-power-free (resp., α^+-power-free) whenever $w \in \Sigma^*$ is α-power-free (resp., α^+-power-free).

The properties of α-power-free morphisms were systematically studied by Bean, Ehrenfeucht, and McNulty [31] and by Brandenburg [48].

Thue [234, Satz 17], with improvements by Bean, Ehrenfeucht, and McNulty [31, Theorem 1], and Berstel [34], gave a criterion for a morphism to be squarefree. Karhumäki [136] gave a similar criterion for cubefree morphisms. Bean, Ehrenfeucht, and McNulty [31] proved the following theorem.

Theorem 1.5 (Bean, Ehrenfeucht, and McNulty). *For any alphabet* Σ *of size at least three, there exists a squarefree morphism* $h : \Sigma^* \to \{0, 1, 2\}^*$*. Further, for any alphabet* Δ *of size at least two, there exists a cubefree morphism* $g : \Delta^* \to \{0, 1\}^*$*.*

Note that if f is an α-power-free morphism and is prolongable on some letter a, then the infinite fixed point $f^\omega(a)$ is necessarily α-power-free.

Recall the Thue–Morse morphism μ defined in Section 1.5. Brandenburg [48] proved the following useful theorem, which was independently rediscovered by Shur [227].

Theorem 1.6 (Brandenburg; Shur). *Let* w *be a binary word and let* $\alpha > 2$ *be a real number. Then* w *is* α*-power-free if and only if* $\mu(w)$ *is* α*-power-free.*

We shall make frequent use of Theorem 1.6 in Chapters 2, 3, and 4.

1.10 Avoiding arbitrarily large repetitions

We have already noted that any binary word of length at least 4 must contain a square; however, Entringer, Jackson, and Schatz [87] constructed an infinite binary word containing no squares xx, where $|x| \geq 3$. The construction is as follows. Let

$\mathbf{w}$ be any infinite squarefree word over the alphabet $\{0, 1, 2\}$. Define the morphism h by

$$\begin{aligned} 0 &\rightarrow 1010 \\ 1 &\rightarrow 1100 \\ 2 &\rightarrow 0111. \end{aligned}$$

Then the word $h(\mathbf{w})$ has the desired properties.

Prodinger and Urbanek [199] gave an example of an infinite binary word whose only squares are of lengths 1, 3, or 5. The particular word studied by Prodinger and Urbanek is the well-known (ordinary) paperfolding word

$$0010011000110110 \cdots .$$

Paperfolding words will be studied in greater detail in Chapter 6.

Fraenkel and Simpson [98] constructed an infinite binary word containing only the squares 00, 11, and 0101. Rampersad, Shallit, and Wang [204], Harju and Nowotka [120], and Ochem [188] gave alternate proofs of this result.

The construction of Harju and Nowotka is particularly nice. As in the Entringer–Jackson–Schatz construction, let $\mathbf{w}$ be an infinite squarefree word over the alphabet $\{0, 1, 2\}$. Define the morphism g by

$$\begin{aligned} 0 &\rightarrow 1^3 0^3 1^2 0^2 1 0 1^2 0^3 1^3 0^2 1 0 \\ 1 &\rightarrow 1^3 0^3 1 0 1^2 0^3 1^3 0^2 1 0 1^2 0^3 1 0 \\ 2 &\rightarrow 1^3 0^3 1^2 0^2 1 0 1^2 0^3 1 0 1^3 0^2 1 0 1^2 0^2. \end{aligned}$$

Then the word $g(\mathbf{w})$ has the desired properties.

Dekking [76] showed that any infinite overlap-free binary word must contain arbitrarily large squares. He also constructed an infinite cubefree binary word containing no squares xx with $|x| \geq 4$, thus disproving a conjecture of Entringer, Jackson, and Schatz. Shallit [225] studied the general problem of constructing binary words that simultaneously avoid k-powers and all squares xx with $|x| \geq \ell$.

Shallit [225] extended the result of Dekking [76] by showing that for $2 < \alpha \leq 7/3$, every infinite α-power-free binary word contains arbitrarily large squares, but for $\alpha > 7/3$, there exist infinite α-power-free binary words that avoid arbitrarily large squares. Ilie, Ochem, and Shallit [128] gave additional related results.

1.11 Abelian repetitions

Another interesting generalization of *square* is the abelian square. An *abelian square* is a nonempty word xy, where y is a permutation of the symbols of x (i.e., x and y are anagrams of each other). For example, the English word `intestines` is an abelian square[1]. This concept was introduced by Erdős [88], who asked if there exists an infinite word over a finite alphabet that contains no abelian repetitions. Evdokimov [92] answered this question in the affirmative, constructing an infinite abelian squarefree word over a 25-letter alphabet.

Pleasants [197] subsequently lowered the alphabet size to 5. Justin [133, 134] showed that abelian 5-powers are avoidable over a binary alphabet. Dekking [78] gave a method of testing if a morphism is abelian k-power-free. He used this method to improve Justin's result by showing that abelian 4-powers are avoidable over a binary alphabet. Specifically, Dekking showed that the infinite fixed point of the morphism

$$\begin{array}{rcl} 0 & \rightarrow & 011 \\ 1 & \rightarrow & 0001 \end{array}$$

contains no abelian 4-powers. He also showed that the infinite fixed point of the morphism

$$\begin{array}{rcl} 0 & \rightarrow & 0012 \\ 1 & \rightarrow & 112 \\ 2 & \rightarrow & 022 \end{array}$$

contains no abelian cubes. Finally, in 1992 Keränen [138] improved the result of Pleasants by showing that abelian squares are avoidable over a 4-letter alphabet. The alphabet sizes in Dekking's and Keränen's results are best possible.

1.12 Circular words and unbordered words

One way to generalize the concepts discussed previously is to consider words as not necessarily being written linearly, but rather as possibly being written on a circle. We formalize this notion as follows. Finite words x and y are *conjugates* if $x = uv$ and $y = vu$ for some words u and v. A *circular k-power-free word* is a word w such that all of the conjugates of w are k-power-free.

Currie [69] proved that there are circular squarefree ternary words of length n for all $n \geq 18$.

Let

$$C = \{0, 1, 00, 11, 010, 101, 010010, 101101\}$$

and let

$$\mathcal{C} = \bigcup_{k \geq 0} \mu^k(C),$$

where μ is the Thue–Morse morphism introduced in Section 1.5. Thue [234, Satz 13] and Harju [117] gave the following characterization of the circular overlap-free binary words.

[1] This example is due to J. Shallit.

Theorem 1.7 (Thue; Harju)**.** *The circular overlap-free binary words are the conjugates of the words in $\mathcal{C}$.*

In particular, it is not the case that for every n sufficiently large, there exists a circular overlap-free binary word of length n. We shall use Theorem 1.7 in Section 5.8 to prove that the complement of the set of circular overlap-free binary words is not unambiguously context-free.

Aberkane and Currie [1] extended Theorem 1.7 as follows.

Theorem 1.8 (Aberkane and Currie)**.** *The circular $7/3$-power-free binary words are the conjugates of the words in $\mathcal{C}$.*

Again, it is not the case that for every n sufficiently large, there exists a circular $7/3$-power-free binary word of length n. However, Aberkane and Currie also proved that there are circular $(7/3)^+$-power-free words of length n for all $n \geq 210$.

There is an interesting connection between these results and the so-called unbordered words. A word w is *bordered* if $w = uvu$ for some words u and v with u non-empty. If w is not bordered, then it is *unbordered*. Harju and Nowotka [119] studied the binary words with the maximal number of unbordered conjugates. They first established that if w is a binary word with $|w| \geq 4$, then w can have at most $\lfloor |w|/2 \rfloor$ unbordered conjugates. They then characterized the binary words that reach this upper bound.

Let $C' = \{0, 1, 010, 101\}$ and let

$$\mathcal{C}' = \bigcup_{k \geq 0} \mu^k(C'),$$

so that $\mathcal{C}'$ is the set of primitive words in $\mathcal{C}$ (i.e., the words in $\mathcal{C}$ that are not squares).

Theorem 1.9 (Harju and Nowotka)**.** *Let $w \in \{0,1\}^*$ be a word with $|w| \geq 4$. Every second cyclic shift of w is unbordered if and only if w is a conjugate of a word in $\mathcal{C}'$.*

In particular, if $n \geq 1$, apart from the conjugates of ab^3 or a^3b, every binary word of length $2n$ with n unbordered conjugates is a conjugate of a word in $\mathcal{C}'$.

1.13 Enumerative combinatorics on words

Given a class of words—for instance, the squarefree words over a three-letter alphabet—it is natural to want to count the number of words of length n in the class. In this section, we discuss such results. For a survey of the area, see Berstel [40].

Kakutani (see Gottschalk and Hedlund [106]) showed that there are uncountably many infinite overlap-free binary words. This result may also be derived from the work of Fife [96] discussed in Chapter 2.7.

Restivo and Salemi [206] proved a remarkable factorization theorem for the overlap-free binary words.

Theorem 1.10 (Restivo and Salemi)**.** *Let $x \in \{0,1\}^*$ be overlap-free. Then there exist $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and an overlap-free $y \in \{0,1\}^*$ such that $x = u\mu(y)v$.*

For example, if $x = 0010011$, then by defining $u = 00$, $v = 1$, and $y = 10$, we see that

$$x = u\mu(y)v = 00\mu(10)1 = 0010011.$$

An easy consequence of this theorem is that there are at most $O(n^4)$ overlap-free binary words of length n. This bound was subsequently improved by Kfoury [139], Kobayashi [141], and Cassaigne [60].

Let a_n denote the number of overlap-free binary words of length n. Currently, the best known upper and lower bounds for a_n are due to Lepistö [158]:

$$a_n = \Omega(n^{1.217}) \quad \text{and} \quad a_n = O(n^{1.369}). \tag{1.2}$$

We shall use these estimates of Lepistö in Section 5.8 to prove that the complement of the set of overlap-free binary words is not unambiguously context-free.

Karhumäki and Shallit [137] generalized Theorem 1.10 as follows.

Theorem 1.11 (Karhumäki and Shallit)**.** *Let $x \in \{0,1\}^*$ be α-power-free, $2 < \alpha \leq 7/3$. Then there exist $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and an α-power-free $y \in \{0,1\}^*$ such that $x = u\mu(y)v$.*

Theorem 1.11 turns out to be very useful; many significant results can be derived from it. We shall use this result several times in Chapters 2, 3, and 4. Karhumäki and Shallit used Theorem 1.11 to deduce that there are only polynomially many 7/3-power-free binary words of length n. Furthermore, they showed that the threshold between polynomial and exponential growth in k-power-free words is $k = 7/3$. That is, there are only polynomially many 7/3-power-free binary words of length n, but there are exponentially many $(7/3)^+$-power-free words of length n.

Brandenburg [48] and Brinkhuis [50] were the first to give an exponential lower bound on the number of squarefree words over a ternary alphabet. Brandenburg showed that there are at least $2^{n/21}$ squarefree words of length n over a ternary alphabet. These bounds have been sucessively improved by Ekhad and Zeilberger [86] (to $2^{n/17}$), Grimm [110] (to $65^{n/40}$), and Sun [229] (to $110^{n/42}$). Further improvements have recently been given by Kolpakov [144].

Brandenburg [48] showed that there are exponentially many cubefree ternary words of length n. Ochem [188] showed that there are exponentially many $(7/4)^+$-power-free ternary words of length n and exponentially many $(7/5)^+$-power-free words of length n over a 4-letter alphabet. It is easy to verify by means of a backtracking computer search that there are only finitely many 7/4-power-free ternary words and only finitely many 7/5-power-free words over a 4-letter alphabet, a fact first observed by Dejean [75]. In particular, all ternary words of length ≥ 39 contain a 7/4-power and all words of length ≥ 122 over a 4-letter alphabet contain a 7/5-power.

Rampersad, Shallit, and Wang [204] proved exponential growth for binary words avoiding all squares xx with $|x| \geq 3$, as well as exponential growth for binary words containing only 00, 11, and 0101 as square subwords.

Carpi [57] showed that are exponentially many abelian squarefree words of length n over a 4-letter alphabet by giving an abelian squarefree substitution. Currie [70] generalized Dekking's test of abelian m-power-freeness from morphisms to substitutions. By giving an abelian 4-power-free substitution, he showed that there are exponentially many binary words of length n that avoid abelian fourth powers. Aberkane, Currie, and Rampersad [2] showed that there are exponentially many abelian cubefree words of length n over a 3-letter alphabet.

1.14 Letter frequencies in words

Another well-studied property of infinite words concerns the frequency with which each alphabet symbol occurs in the word. If $\mathbf{w}$ is an infinite word, the frequency of a letter a in $\mathbf{w}$ is given by

$$\lim_{n\to\infty} \frac{|w_0 w_1 \cdots w_{n-1}|_a}{n},$$

provided the limit exists (recall that $|x|_a$ denotes the number of occurrences of a in x).

Saari [215] used Perron–Frobenius non-negative matrix theory (see Horn and Johnson [125, Chapter 8]) to show that letter frequencies always exist for fixed points of binary morphisms. He also [216] gave a criterion to determine when letter frequencies exist in fixed points of morphisms over larger alphabets. Earlier results in this vein were obtained by Cobham [67] and Michel [173, 174] (see also Queffélec [201] or Allouche and Shallit [23, Chapter 8]). Adamczewski [4, 5] also gave some interesting results related to letter frequencies in infinite words.

Tarannikov [231] proved that no infinite squarefree ternary word can have a letter occur with frequency less than $1780/6481 = 0.27464897\cdots$. Ochem [189] improved this bound to $1000/3641 = 0.27464982\cdots$. Ochem also constructed an infinite squarefree ternary word with a minimal letter frequency of $883/3215 = 0.27465007\cdots$. Khalyavin [140] recently proved that the value of 883/3215 is optimal.

From Theorem 1.10 one easily concludes that the frequencies of 0 and 1 in any infinite overlap-free binary word are both 1/2. Indeed, from Theorem 1.11 one deduces the same result for 7/3-power-free words, a fact observed by Kolpakov, Kucherov, and Tarannikov [145]. Kolpakov, Kucherov, and Tarannikov also proved that for $\alpha > 7/3$, it is possible to construct an infinite α-power-free binary word where the frequency of 0 is less than 1/2.

1.15 Subword complexity

For an infinite word $\mathbf{w}$, let $p_{\mathbf{w}}(n)$ denote the *subword complexity function* of $\mathbf{w}$. That is, the value of $p_{\mathbf{w}}(n)$ is equal to the number of subwords of length n that occur in $\mathbf{w}$. This notion of subword complexity will be quite important for several results discussed in Chapter 5.

The following theorem (for which see Coven and Hedlund [68]) characterizes the ultimately periodic words.

Theorem 1.12. *Let $\mathbf{w}$ be an infinite word. The following are equivalent:*

- *$\mathbf{w}$ is ultimately periodic;*
- *$p_{\mathbf{w}}(n)$ is a bounded function;*
- *$p_{\mathbf{w}}(n+1) = p_{\mathbf{w}}(n)$ for some $n \geq 1$;*
- *$p_{\mathbf{w}}(n) \leq n$ for some $n \geq 1$.*

Ehrenfeucht, Lee, and Rozenberg [83] proved that if $\mathbf{w}$ is the fixed point of a morphism, then $p_{\mathbf{w}}(n) = O(n^2)$. Pansiot [193] improved this result by precisely determining the possible orders of growth for $p_{\mathbf{w}}(n)$.

Theorem 1.13 (Pansiot)**.** *Let $\mathbf{w}$ be an infinite word obtained by iterating a morphism. There exists constants c_1 and c_2, $0 < c_1 \leq c_2$, such that $c_1 \cdot g(n) \leq p_{\mathbf{w}}(n) \leq c_2 \cdot g(n)$, where $g(n)$ is one of the functions 1, n, $n \log n$, $n \log \log n$, or n^2.*

The following result of Cobham [67] applies only to uniform morphisms.

Theorem 1.14 (Cobham)**.** *Let $\mathbf{w}$ be an aperiodic infinite word obtained by iterating a uniform morphism. There exists constants c_1 and c_2, $0 < c_1 \leq c_2$, such that $c_1 \cdot n \leq p_{\mathbf{w}}(n) \leq c_2 \cdot n$.*

Avgustinovich, Fon-Der-Flaass, and Frid [30] generalized the concept of subword complexity by considering the *arithmetical complexity* of a word. The arithmetical complexity function of a word w is the function $p_w^A(n)$ that counts the total number of distinct subwords of length n that appear in all subsequences of w indexed by arithmetic progressions. Avgustinovich, Fon-Der-Flaass, and Frid showed that the words with lowest arithmetical complexity come from a class of words known as

Toeplitz words, of which the paperfolding words form a special class. We shall make use of the results of Avgustinovich, Fon-Der-Flaass, and Frid [30] in Chapter 6.

This concludes our brief introduction to the area of combinatorics on words. In the next chapter we begin our study of the overlap-free words by first considering various properties of the Thue–Morse word.

Chapter 2

The Thue–Morse Word and Overlap-Free Words

There is a considerable body of literature regarding the Thue–Morse word, which we defined in Section 1.5. We begin this chapter by giving some of the more notable properties of this well-studied word. For a survey, see Allouche and Shallit [21]. We also present a proof of Fife's characterization of the infinite binary overlap-free words. Finally, we give some applications to transcendental number theory.

2.1 The standard definitions

First, we should note that although the infinite word

$$0110100110010110100101100110100 1\cdots$$

is named after Thue, who studied its properties in a 1906 paper, and Morse, who rediscovered it in the 1920's, the Thue–Morse word occurs in a much earlier communication of Prouhet [200] to the French Academy of Sciences in 1851[1]. Indeed, Prouhet gave a more general construction, yielding not only the Thue–Morse word, but a family of words over larger alphabets having several interesting properties. These words are sometimes referred to as either *generalized Thue–Morse words* or *Prouhet words*.

Let $k \geq 2$ and $m \geq 2$ be integers. Let $s_k(n)$ denote the sum of the digits in the base-k expansion of n. It is well known that the Thue–Morse word $\mathbf{t} = t(0)t(1)t(2)\cdots$ can be defined by $t(n) = s_2(n) \bmod 2$. We define the *generalized Thue–Morse word* $\mathbf{t}_{k,m}$ by $\mathbf{t}_{k,m}(n) = s_k(n) \bmod m$, so that $\mathbf{t} = \mathbf{t}_{2,2}$. For example,

$$\mathbf{t}_{3,4} = 012123230123230301\cdots.$$

[1] The original communication of Prouhet is transcribed in Appendix A.

In Section 1.5 we defined the Thue–Morse word as the fixed point of the morphism μ defined by $\mu(0) = 01$ and $\mu(1) = 10$. Similarly, the generalized Thue–Morse word $\mathbf{t}_{k,m}$ can be defined as the fixed point of the morphism $\mu_{k,m}$, where for $a \in \{0, 1, \ldots, m-1\}$, $\mu_{k,m}(a) = a(a+1)(a+2)\cdots(a+k-1)$ and the sums are taken modulo m. For example, $\mu_{3,4}$ is the morphism

$$\begin{aligned} 0 &\rightarrow 012 \\ 1 &\rightarrow 123 \\ 2 &\rightarrow 230 \\ 3 &\rightarrow 301. \end{aligned}$$

We have already noted in Section 1.6 the following seminal result of Thue [233, Satz 6].

Theorem 2.1 (Thue)**.** *The Thue–Morse word* $\mathbf{t}$ *is overlap-free.*

Allouche and Shallit [22] proved the following generalization.

Theorem 2.2 (Allouche and Shallit)**.** *For* $k \geq 2$ *and* $m \geq 2$*, the generalized Thue–Morse word* $\mathbf{t}_{k,m}$ *is overlap-free if and only if* $m \geq k$*.*

Morton and Mourant [180] proved that $\mathbf{t}_{k,m}$ is ultimately periodic if and only if $m|(k-1)$. Blondin–Massé, Brlek, Glen, and Labbé [46] determined the critical exponent of $\mathbf{t}_{k,m}$ for all values of k and m. The *critical exponent* of an infinite word $\mathbf{w}$ is the quantity

$$\sup\{q \in \mathbb{Q} : \text{there exists a non-empty word } x \text{ such that } x^q \text{ is a subword of } \mathbf{w}\}.$$

Séébold [223, 224] gave additional results concerning the generalized Thue–Morse words, and Frid [102] studied a further generalization of such words. Tompkins [236] generalized the morphisms $\mu_{k,m}$ by defining morphisms whose image words are given by the rows of a Latin square, and he proved that the infinite words generated by iterating these morphisms are overlap-free.

2.2 A connection to squarefree words

Recall from Section 1.6 Thue's result that there exist infinite squarefree words over an alphabet of size three. We may derive this result as a consequence of Theorem 2.1 (see, for example, Allouche and Shallit [23, Theorem 1.6.2]).

Proof of Theorem 1.2. Let $\mathbf{t} = t_0t_1t_2\cdots$ be the Thue–Morse word. Define a word $\mathbf{x} = x_1x_2x_3\cdots$ such that for all $n \geq 1$, x_n is the number of 1's between the n-th and $(n+1)$-st occurrences of 0 in $\mathbf{t}$. We claim that

$$\mathbf{x} = 210201210120210\cdots$$

is squarefree.

Suppose to the contrary that $\mathbf{x}$ contains a square yy, where $y = y_1 y_2 \cdots y_m$. Then by the definition of $\mathbf{x}$, $\mathbf{t}$ contains a subword

$$01^{y_1}01^{y_2}0 \cdots 01^{y_m}01^{y_1}01^{y_2}0 \cdots 01^{y_m}0,$$

which is an overlap, contrary to Theorem 2.1. We conclude that $\mathbf{x}$ is squarefree, as required. □

There is a more general connection between a certain class of ternary squarefree words and the binary overlap-free words. Define a morphism τ by $\tau(0) = 011$, $\tau(1) = 01$, and $\tau(2) = 0$.

The following theorem is due to Thue [234] (see Problem 2.3.7 of Lothaire [159]).

Theorem 2.3 (Thue)**.** *An infinite word* $\mathbf{a}$ *over* $\{0, 1, 2\}$ *avoids squares and the subwords* 010 *and* 02120 *if and only if* $\mathbf{b} = \tau(\mathbf{a})$ *is overlap-free.*

In Section 2.9 we shall use Theorem 2.3 to derive a result concerning the ternary expansions of algebraic numbers.

2.3 Properties of the Thue–Morse word

Thue [234] (see also Gottschalk and Hedlund [107]) characterized the bi-infinite overlap-free words as follows. We first define the bi-infinite Thue–Morse word

$$\begin{aligned} \mathbf{t}^R\mathbf{t} &= \cdots t_2 t_1 t_0 t_0 t_1 t_2 \cdots \\ &= \cdots 1001100101100110100110 01 \cdots . \end{aligned}$$

Theorem 2.4 (Thue; Gottschalk and Hedlund)**.** *The set of bi-infinite overlap-free words over* $\{0, 1\}$ *is exactly the set of shifts of the bi-infinite Thue–Morse word.*

To characterize the one-sided infinite overlap-free words is much more difficult, and the finite overlap-free words even harder still. We shall discuss this problem in Section 2.7.

Shur [227] gave the following analogue of Theorem 2.4.

Theorem 2.5 (Shur)**.** *The set of bi-infinite* $7/3$*-power-free words over* $\{0, 1\}$ *is exactly the set of shifts of the bi-infinite Thue–Morse word.*

Furthermore, Shur showed that the number $7/3$ is best possible; *i.e.*, the result no longer holds if the number $7/3$ is replaced by a larger number. In other words, for $\alpha > 7/3$, there exist bi-infinite α-power-free words over $\{0, 1\}$ that are not shifts of the bi-infinite Thue–Morse word.

Pansiot [192] proved the following result concerning the one-sided Thue–Morse word.

Theorem 2.6 (Pansiot)**.** *Let h be a morphism and let* $\mathbf{t}$ *be the Thue–Morse word. If* $h(\mathbf{t}) = \mathbf{t}$*, then h is a power of the Thue–Morse morphism* μ*.*

Séébold [220, 222] strengthened this result as follows.

Theorem 2.7 (Séébold)**.** *The only infinite overlap-free words over* $\{0, 1\}$ *generated by iterated morphisms are* $\mathbf{t}$ *and its complement* $\overline{\mathbf{t}}$*.*

Berstel and Séébold [43] gave an alternate proof of this result. In Chapter 3 we shall give a generalization of this result.

2.4 Squares in the Thue–Morse word

In this section we present some results concerning the squares that occur as subwords of the Thue–Morse word.

Let

$$A = \{00, 11, 010010, 101101\}$$

and let

$$\mathcal{A} = \bigcup_{k \geq 0} \mu^k(A).$$

Pansiot [192] and Brlek [51] gave the following characterization of the squares in $\mathbf{t}$.

Theorem 2.8 (Pansiot; Brlek)**.** *The set of squares in* $\mathbf{t}$ *is exactly the set* $\mathcal{A}$*.*

We can use this result to prove the following proposition, for which we shall find an interesting application to a language theoretic problem in Chapter 5. We first state the following lemma, which is easy to prove.

Lemma 2.9. *Let x and y be binary words. Then x is a prefix (resp. suffix) of y if and only if $\mu(x)$ is a prefix (resp. suffix) of $\mu(y)$.* □

Proposition 2.10. *For any position i, there is at most one square in* $\mathbf{t}$ *beginning at position i.*

Proof. Suppose to the contrary that there exist distinct squares x and y that begin at position i. Without loss of generality, suppose that x and y begin with 0. Then by Theorem 2.8, $x = \mu^p(u)$ and $y = \mu^q(v)$, for some p, q and $u, v \in \{00, 010010\}$. Suppose $p \leq q$ and let $w = \mu^{q-p}(v)$. By Lemma 2.9, either u is a proper prefix of w or w is a proper prefix of u, neither of which is possible for any choice of $u, v \in \{00, 010010\}$. □

Brown, Rampersad, Shallit, and Vasiga [52] gave additional results concerning the squares in the Thue–Morse word. In particular, they studied properties of the sequences $A(i)$, $B(i)$, and $C(i)$, where for $i \geq 0$

- $A(i)$ counts the number of squares that begin at a position $\leq i$ in the Thue–Morse word;
- $B(i)$ counts the number of positions $p \leq i$ such that the Thue–Morse word contains a square beginning at position p; and,
- $C(i)$ is 0 if the Thue–Morse word contains no square xx beginning at position i, and $|x|$ otherwise.

Note that by Proposition 2.10 we have $A(i) = B(i)$ for all $i \geq 0$.

Aberkane, Linek, and Mor [3] characterized the set of all rational numbers α such that the Thue–Morse word contains an α-power. Saari [217] proved that at every position of the Thue–Morse word there begins a 5/3-power, and the constant 5/3 is optimal.

2.5 Subword complexity of the Thue–Morse word

Let $\mathbf{t}$ denote the Thue–Morse word and recall that $p_{\mathbf{t}}(n)$ denotes the number of subwords of $\mathbf{t}$ of length n. Brlek [51] and de Luca and Varricchio [165] (see also the subsequent work of Avgustinovich [29], Tapsoba [230], Frid [99], and Tromp and Shallit [238]) determined that

$$p_{\mathbf{t}}(n+1) = \begin{cases} 2, & \text{if } n = 0; \\ 4, & \text{if } n = 1; \\ 4n - 2^a, & \text{if } n = 2^a + b, \\ & \text{where } a \geq 1,\ 0 \leq b < 2^{a-1}; \\ 4n - 2^a - 2b, & \text{if } n = 2^a + 2^{a-1} + b, \\ & \text{where } a \geq 1,\ 0 \leq b < 2^{a-1}. \end{cases} \tag{2.1}$$

The Thue–Morse word thus has $O(n)$ subword complexity, as it must by Theorem 1.14. We shall use this characterization of the subword complexity of the Thue–Morse word in Section 5.8.

Clearly the Thue–Morse word must avoid, in addition to all overlaps, many other subwords. It is possible to describe precisely the set of subwords that do not appear in the Thue–Morse word in terms of the set of its minimal forbidden subwords. A word w is a *minimal forbidden subword* of an infinite word $\mathbf{x}$ if w does not occur in $\mathbf{x}$ but every proper subword w' of w does occur in $\mathbf{x}$. Using results of Mignosi, Restivo, and Sciortino [175], along with the work of de Luca and Mione [164], one can describe the set M of minimal forbidden subwords of $\mathbf{t}$ as follows. For $w \in \{010, 101\}$ define the sets

$$M_{1,w} = \bigcup_{k \geq 1} 0\mu^{2k-1}(w)0, \quad M_{2,w} = \bigcup_{k \geq 1} 1\mu^{2k-1}(w)1,$$

$$M_{3,w} = \bigcup_{k\geq 1} 0\mu^{2k}(w)1, \quad \text{and} \quad M_{4,w} = \bigcup_{k\geq 1} 1\mu^{2k}(w)0.$$

Then

$$M = \bigcup_{i=1}^{4} (M_{i,010} \cup M_{i,101}).$$

Next we consider the generalized Thue–Morse words. Tromp and Shallit [238] characterized the subword complexity of these words as follows:

$$p_{\mathbf{t}_{2,m}}(n+1) = \begin{cases} m, & \text{if } n = 0; \\ m^2, & \text{if } n = 1; \\ m(mn - 2^{a-1}), & \text{if } n = 2^a + b, \\ & \text{where } a \geq 1,\ 0 \leq b < 2^{a-1}; \\ m(mn - 2^{a-1} - b), & \text{if } n = 2^a + 2^{a-1} + b, \\ & \text{where } a \geq 1,\ 0 \leq b < 2^{a-1}. \end{cases} \tag{2.2}$$

To the best of our knowledge, no one has characterized the minimal forbidden subwords of the generalized Thue–Morse words. We therefore have the following open problem.

Problem 2.11. *Determine the minimal forbidden subwords of the generalized Thue–Morse words.*

2.6 Finite modifications of the Thue–Morse word

In this section we consider the question of whether or not the Thue–Morse word remains overlap-free if a finite number of its symbols are changed. Specifically, we show that every word obtained by changing a finite number of bits of the Thue–Morse word contains an overlap[2]. The proof of this result provides a nice illustration of the usefulness of Theorem 1.10 in proving properties of overlap-free words. It also allows us to introduce a set of identities concerning overlap-free words that we shall use extensively in Section 2.7.

Let $\mathcal{PF}(2^+)$ denote the set of infinite overlap-free (2^+-power-free) words over $\{0,1\}$. Allouche, Currie, and Shallit [19] proved the following identities for infinite overlap-free words.

Theorem 2.12 (Allouche, Currie, and Shallit). *For $a \in \{0,1\}$ and $\mathbf{x} \in \{0,1\}^\omega$:*

(a) $\mathbf{x} \in \mathcal{PF}(2^+) \iff \mu(\mathbf{x}) \in \mathcal{PF}(2^+)$

(b) $a\mathbf{x} \in \mathcal{PF}(2^+) \iff \overline{a}\mu(\mathbf{x}) \in \mathcal{PF}(2^+)$

[2] This result can be found in Brown, Rampersad, Shallit, and Vasiga [52].

(c) $a\mathbf{x} \in \mathcal{PF}(2^+)$ *and* $\mathbf{x}$ *starts* $a\bar{a}a \iff \overline{aa}\mu(\mathbf{x}) \in \mathcal{PF}(2^+)$.

Allouche, Currie, and Shallit used these identities to prove that $001001\bar{\mathbf{t}}$ is the lexicographically least infinite overlap-free binary word. Note that $\mathbf{x}$ must be infinite in Theorem 2.12; the result no longer holds if $\mathbf{x}$ is a finite word.

Theorem 2.13. *Let* $\mathbf{t}'$ *be a word obtained from* $\mathbf{t}$ *by changing* $k > 0$ *bits. Then* $\mathbf{t}'$ *contains an overlap.*

Proof. The proof is by contradiction. Let k be minimal such that $\mathbf{t}'$ is overlap-free. Suppose $k = 1$. Changing the first, second, or third bit of $\mathbf{t}$ creates the overlaps 111, 01010, or 1001001 respectively. Furthermore, there are 22 words of length 8 in $\mathbf{t}$, and changing the fourth bit of any such word creates an overlap, as shown in Table 2.1.

Original subword	Modified subword	Original subword	Modified subword
00101100	00**111**100	10010110	1**000**0110
00101101	00**111**101	10011001	1**000**1001
00110010	001**000**10	10011010	1**000**1010
00110100	**0010010**0	10100101	101**10101**
01001011	0**1011011**	10100110	**1011011**0
01001100	010**111**00	10110011	101**000**11
01011001	**0100100**1	10110100	1**0100100**
01011010	010**01010**	11001011	**1101101**1
01100101	0**111**0101	11001101	110**111**01
01100110	0**111**0110	11010010	11**000**010
01101001	0**111**1001	11010011	11**000**011

Table 2.1: Changing a bit in the Thue-Morse word

We assume then that $k > 1$. By Theorem 1.10 we can write $\mathbf{t}' = x\mu(\mathbf{y})$, where $x \in \{\epsilon, 0, 1, 00, 11\}$ and $\mathbf{y}$ is overlap-free. We have three cases, $x \in \{\epsilon, 0, 00\}$. (The cases where $x \in \{1, 11\}$ are similar to those where $x \in \{0, 00\}$.)

Case 1: $x = \epsilon$, $\mathbf{t}' = \mu(\mathbf{y})$. Then $\mathbf{y}$ differs from $\mathbf{t}$ in fewer than k bits and is overlap-free, contradicting the minimality of k.

Case 2: $x = 0$, $\mathbf{t}' = 0\mu(\mathbf{y})$. The definition of $\mathbf{t}$ as the fixed point of μ implies that all occurrences of 00 in $\mathbf{t}$ begin at an odd position. It follows that somewhere after the last bit changed in $\mathbf{t}'$ there must be an occurrence of 00 that begins in an odd position. But then 00 must be the image under μ of either 0 or 1, which is impossible.

Case 3: $x = 00$, $\mathbf{t}' = 00\mu(\mathbf{y})$. By Theorem 2.12, if $00\mu(\mathbf{y})$ is overlap-free, then $1\mathbf{y}$ is overlap-free. But $1\mathbf{y}$ differs from $\mathbf{t}$ in fewer than k bits, contradicting the minimality of k. □

2.7 Fife's Theorem

The properties of infinite overlap-free words have been studied extensively (see, for example, the survey by Séébold [221]). In this section we discuss Fife's characterization of the infinite overlap-free binary words [96]. The original proof of Fife is somewhat difficult to follow; Berstel [39] gave a simpler proof of Fife's result. Cassaigne [60] and Carpi [56] further extended the work of Fife by characterizing the set of finite overlap-free binary words by means of a regular language. Our goal in this section is to give yet another proof of Fife's theorem based on the identites of Theorem 2.12. However, we shall rely heavily on the structure and notation of Berstel's proof.

Recall that we denote the complement of any binary word x by $\overline{x}$. Let

$$X = \{\mu^n(0) : n \geq 0\} \cup \{\mu^n(1) : n \geq 0\}.$$

The words in X are called *Morse blocks*. Let w be any binary word ending in 01 or 10. We define mappings α, β, and γ on w as follows. Let x be the longest word in X such that $w = yx\overline{x}$ for some y. Then

$$\begin{aligned} \alpha(w) &= wxx\overline{x} \\ \beta(w) &= wx\overline{x}\overline{x}x \\ \gamma(w) &= w\overline{x}x. \end{aligned}$$

For example, if $w = 011001$, then $w = yx\overline{x}$, where $y = 01$, $x = 10$, and $\overline{x} = 01$. Further, we have

$$\begin{aligned} \alpha(w) &= 011001\,101001 \\ \beta(w) &= 011001\,10010110 \\ \gamma(w) &= 011001\,0110. \end{aligned}$$

Let $\mathbf{f} = f_0f_1f_2\cdots$ be a infinite word over the alphabet $B = \{\alpha, \beta, \gamma\}$. Fife showed that the infinite composition

$$\mathbf{w} = (\cdots \circ f_2 \circ f_1 \circ f_0)(01),$$

which for convenience we shall write as $\mathbf{w} = 01 \bullet \mathbf{f}$, is overlap-free if and only if $\mathbf{f}$ contains no subword

$$f \in \{\alpha, \beta\}(\gamma\gamma)^*\{\beta\alpha, \gamma\beta, \alpha\gamma\}.$$

For example, the Thue–Morse word $\mathbf{t}$ is defined by

$$\mathbf{t} = (\cdots \circ \gamma \circ \gamma \circ \gamma)(01) = 01 \bullet \gamma^\omega.$$

Let $\mathbf{f} \in B^\omega$ such that $01 \bullet \mathbf{f} = \mathbf{x}$ is overlap-free. Then we have the following:

$$01 \bullet \alpha\mathbf{f} = 0\mu(\overline{\mathbf{x}}), \quad 01 \bullet \beta\mathbf{f} = \mu(0\mathbf{x}), \quad 01 \bullet \gamma\mathbf{f} = \mu(\mathbf{x}). \tag{2.3}$$

As in Berstel [39], we define the sets I and F by

$$I = \{\alpha, \beta\}(\gamma\gamma)^*\{\beta\alpha, \gamma\beta, \alpha\gamma\}$$

and

$$F = B^\omega \setminus B^* I B^\omega.$$

Fife's theorem, which characterizes the infinite overlap-free binary words beginning with 0, is the following.

Theorem 2.14 (Fife)**.** *Let* $\mathbf{x}$ *be an infinite overlap-free word over* $\{0, 1\}$*.*

1. *If* $\mathbf{x}$ *begins with* 01*, then* $\mathbf{x}$ *is overlap-free if and only if* $\mathbf{x} = 01 \bullet \mathbf{f}$ *for some* $\mathbf{f} \in F$*.*
2. *If* $\mathbf{x}$ *begins with* 001*, then* $\mathbf{x}$ *is overlap-free if and only if* $\mathbf{x} = 001 \bullet \mathbf{f}$ *for some* $\mathbf{f}$ *such that* $\beta\mathbf{f} \in F$*.*

We first show that part 2 follows from part 1. Note that

$$\mu(001 \bullet \mathbf{f}) = 010110 \bullet \mathbf{f} = 01 \bullet \beta\mathbf{f}.$$

By Theorem 1.6, $001 \bullet \mathbf{f}$ is overlap-free if and only if $\mu(001 \bullet \mathbf{f})$ is overlap-free. However, by part 1, $\mu(001 \bullet \mathbf{f})$ is overlap-free if and only if $\beta\mathbf{f} \in F$. It thus suffices to prove part 1. Let

$$W = \{\mathbf{f} \in B^\omega : 01 \bullet \mathbf{f} \in \mathcal{PF}(2^+)\}.$$

To prove Theorem 2.14 it is enough to prove that $W = F$.

Let $L \subseteq \Sigma^\omega$ and let $x \in \Sigma^*$. We define the *(left) quotient* $x^{-1}L$ by

$$x^{-1}L = \{\mathbf{y} \in \Sigma^\omega : x\mathbf{y} \in L\}.$$

The next proposition (compare Berstel [39, Proposition 5.2]) establishes several identities concerning quotients of the set W. These identities demonstrate that the set W is precisely the set of infinite labeled paths through the automaton given in Figure 2.1; or, equivalently, that $W = F$. Thus, proving Proposition 2.15 completes the proof of Theorem 2.14.

Proposition 2.15. *The following identities hold:*

(a) $W = \gamma^{-1}W$*;*

(b) $\alpha^{-1}W = \beta^{-1}W = (\alpha\gamma\alpha)^{-1}W = (\alpha\gamma\gamma)^{-1}W$*;*

(c) $(\alpha^2)^{-1}W = (\alpha^3)^{-1}W$*;*

(d) $(\alpha\beta)^{-1}W = (\alpha^2\beta)^{-1}W$*;*

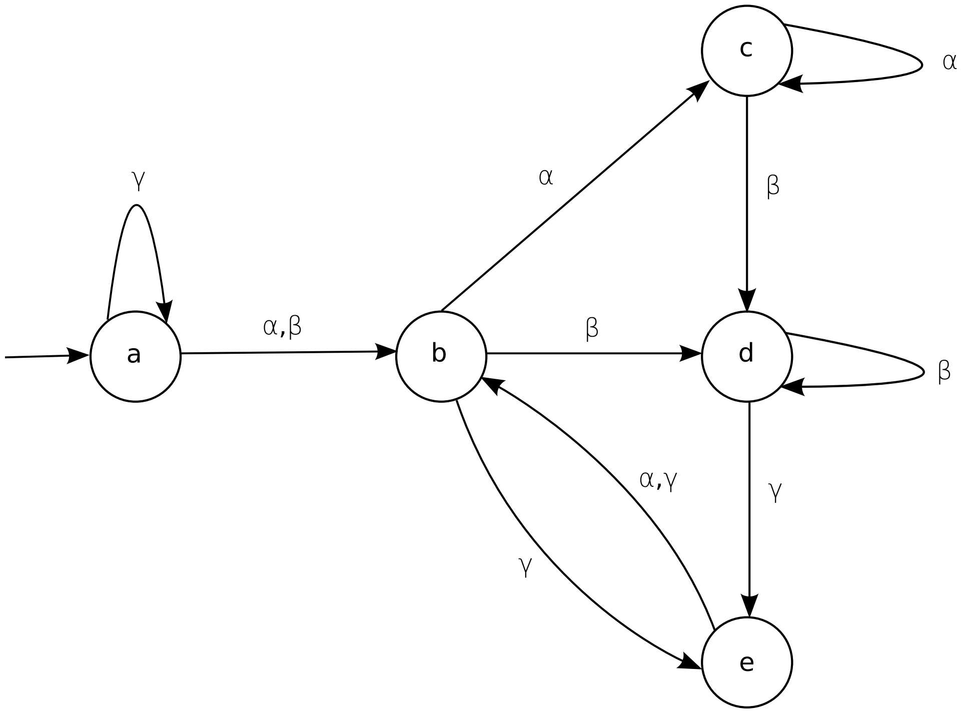

Figure 2.1: The Fife automaton M

(e) $(\alpha\gamma)^{-1}W = (\alpha\beta\gamma)^{-1}W$;

(f) $(\alpha^2\gamma)^{-1}W = (\alpha\beta\alpha)^{-1}W = (\alpha\gamma\beta)^{-1}W = \emptyset$.

Proof. Let $01 \bullet \mathbf{f} = \mathbf{x}$ be overlap-free. For the proof we adopt a somewhat terse notation that we hope is clear nevertheless. In the expressions below, the occurrence of (a), (b), or (c) above an equivalence sign ($\Longleftrightarrow$) indicates that the equivalence holds because of part (a), (b), or (c) respectively of Theorem 2.12. The occurrence of (2.3) above an equivalence sign indicates that the equivalence holds due to one or more of the identities of Eq. (2.3).

(a)

$$\gamma\mathbf{f} \in W \iff 01 \bullet \gamma\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} \mu(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(a)}{\iff}$$
$$\mathbf{x} \in \mathcal{PF}(2^+) \iff 01 \bullet \mathbf{f} \in \mathcal{PF}(2^+) \iff \mathbf{f} \in W$$

Thus, $W = \gamma^{-1}W$, as required.

(b)

$$\alpha\mathbf{f} \in W \iff 01 \bullet \alpha\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff} 0\mathbf{x} \in \mathcal{PF}(2^+)$$

$$\beta\mathbf{f} \in W \iff 01 \bullet \beta\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} \mu(0\mathbf{x}) \in \mathcal{PF}(2^+) \iff 0\mathbf{x} \in \mathcal{PF}(2^+)$$

$$\alpha\gamma\alpha\mathbf{f} \in W \iff 01 \bullet \alpha\gamma\alpha\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 0\mu^2(1\mu(\mathbf{x})) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$1\mu(1\mu(\mathbf{x})) \in \mathcal{PF}(2^+) \overset{(b)}{\iff} 01\mu(\mathbf{x}) \in \mathcal{PF}(2^+) \iff$$
$$\mu(0\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(a)}{\iff} 0\mathbf{x} \in \mathcal{PF}(2^+)$$

$$\alpha\gamma\gamma\mathbf{f} \in W \iff 01 \bullet \alpha\gamma\gamma\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu^3(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$0\mu^2(\mathbf{x}) \overset{(b)}{\iff} 1\mu(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff} 0\mathbf{x} \in \mathcal{PF}(2^+)$$

Thus,

$$\alpha\mathbf{f} \in W \iff \beta\mathbf{f} \in W \iff \alpha\gamma\alpha\mathbf{f} \in W \iff \alpha\gamma\gamma\mathbf{f} \in W,$$

so that

$$\alpha^{-1}W = \beta^{-1}W = (\alpha\gamma\alpha)^{-1}W = (\alpha\gamma\gamma)^{-1}W,$$

as required.

(c)

$$\alpha^2\mathbf{f} \in W \iff 01 \bullet \alpha^2\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 0\mu(1\mu(\mathbf{x})) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$11\mu(\mathbf{x}) \in \mathcal{PF}(2^+)$$

$$\alpha^3\mathbf{f} \in W \iff 01 \bullet \alpha^3\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu(0\mu(1\mu(\mathbf{x}))) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$00\mu(1\mu(\mathbf{x})) \overset{(c)}{\iff} 11\mu(\mathbf{x}) \in \mathcal{PF}(2^+)$$

Thus, $\alpha^2\mathbf{f} \in W \iff \alpha^3\mathbf{f} \in W$, so that $(\alpha^2)^{-1}W = (\alpha^3)^{-1}W$, as required.

(d)

$$\alpha\beta\mathbf{f} \in W \iff 01 \bullet \alpha\beta\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu^2(0\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$0\mu(0\mathbf{x}) \in \mathcal{PF}(2^+)$$

$$\alpha^2\beta\mathbf{f} \in W \iff 01 \bullet \alpha^2\beta\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 0\mu(1\mu^2(0\mathbf{x})) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$11\mu^2(0\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(c)}{\iff} 0\mu(0\mathbf{x}) \in \mathcal{PF}(2^+)$$

Thus, $\alpha\beta\mathbf{f} \in W \iff \alpha^2\beta\mathbf{f} \in W$, so that $(\alpha\beta)^{-1}W = (\alpha^2\beta)^{-1}W$, as required.

(e)

$$\alpha\gamma\mathbf{f} \in W \iff 01 \bullet \alpha\gamma\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu^2(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$0\mu(\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(b)}{\iff} 1\mathbf{x} \in \mathcal{PF}(2^+)$$

$$\alpha\beta\gamma\mathbf{f} \in W \iff 01 \bullet \alpha\beta\gamma\mathbf{f} \in \mathcal{PF}(2^+) \overset{(2.3)}{\iff} 1\mu^2(0\mu(\mathbf{x})) \in \mathcal{PF}(2^+) \overset{(b)}{\iff}$$
$$0\mu(0\mu(\mathbf{x})) \overset{(b)}{\iff} 10\mu(\mathbf{x}) \in \mathcal{PF}(2^+) \iff \mu(1\mathbf{x}) \in \mathcal{PF}(2^+) \overset{(c)}{\iff} 1\mathbf{x} \in \mathcal{PF}(2^+)$$

Thus, $\alpha\gamma\mathbf{f} \in W \iff \alpha\beta\gamma\mathbf{f} \in W$, so that $(\alpha\gamma)^{-1}W = (\alpha\beta\gamma)^{-1}W$, as required.

(f) It suffices to observe that each of the words

$$\begin{aligned} 01 \bullet \alpha^2\gamma &= 010011\,010011\,0\,010110 \\ 01 \bullet \alpha\beta\alpha &= 010011001011\,010011001011\,0 \\ 01 \bullet \alpha\gamma\beta &= 01001011\,01001011\,0\,01101001 \end{aligned}$$

contains an overlap. □

Observe that each state of the Fife automaton M corresponds to one of the sets identified in parts (a)–(f) of Proposition 2.15 (the state corresponding to (f) is not shown in the figure). Note also that there are uncountably many distinctly labeled infinite paths through the automaton M, whence we deduce the following result mentioned in Section 1.13.

Theorem 2.16 (Kakutani). *There are uncountably many infinite overlap-free words over* $\{0, 1\}$.

2.8 Generalizing Fife's Theorem

Computer experiments suggest that a characterization similar to that of Theorem 2.14 may hold for the infinite 7/3-power-free binary words as well. We do not currently have a proof of this, but we are able to prove some identities along the lines of Proposition 2.15.

Let $\mathcal{PF}(7/3)$ denote the set of infinite 7/3-power-free words over $\{0, 1\}$ and let

$$V = \{\mathbf{f} \in B^\omega : 01 \bullet \mathbf{f} \in \mathcal{PF}(7/3)\}.$$

By the same argument as in Section 2.7, $001 \bullet \mathbf{f} \in \mathcal{PF}(7/3)$ if and only if $\beta\mathbf{f} \in V$, so it suffices to consider the set V.

Proposition 2.17. *The following identities hold:*

(a) $V = \gamma^{-1}V$;

(b) $\beta^{-1}V = (\beta\gamma\beta)^{-1}V$;

(c) $(\beta\alpha)^{-1}V = (\beta\beta\alpha)^{-1}V$;

(d) $(\beta\gamma\alpha)^{-1}V = (\beta\beta\gamma\alpha)^{-1}V$.

Proof. Let $01 \bullet \mathbf{f} = \mathbf{x}$ be 7/3-power-free. We now omit the superscripts above the equivalence symbols and we apply the identities of Eq. (2.3) implicitly.

(a)

$$\gamma\mathbf{f} \in V \iff 01 \bullet \gamma\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(x) \in \mathcal{PF}(7/3) \iff \\ \mathbf{x} \in \mathcal{PF}(7/3) \iff \mathbf{f} \in V$$

Thus, $V = \gamma^{-1}V$, as required.

(b)

$$\beta\mathbf{f} \in V \iff 01 \bullet \beta\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(0\mathbf{x}) \in \mathcal{PF}(7/3) \iff 0\mathbf{x} \in \mathcal{PF}(7/3)$$

$$\beta\gamma\beta\mathbf{f} \in V \iff 01 \bullet \beta\gamma\beta\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(0\mu^2(0\mathbf{x})) \in \mathcal{PF}(7/3) \iff \\ 0\mu^2(0\mathbf{x}) \in \mathcal{PF}(7/3) \iff 0\mathbf{x} \in \mathcal{PF}(7/3)$$

One direction of the last equivalence is clear. To see the other, suppose that $0\mathbf{x} \in \mathcal{PF}(7/3)$ but $\mathbf{y} = 0\mu^2(0\mathbf{x}) \notin \mathcal{PF}(7/3)$. Any 7/3-power in $\mathbf{y}$ is necessarily a prefix of $\mathbf{y}$. Note that $\mathbf{y}$ begins with 001100110. This prefix cannot occur elsewhere in $\mathbf{y}$ as that would imply that $\mu^2(0\mathbf{x})$ contains either the cube 000 or the 7/3-power 1001100110, a contradiction. Thus the period of any 7/3-power in $\mathbf{y}$ is at most 8. It is easy to verify that no such 7/3-power exists.

Thus, $\beta\mathbf{f} \in V \iff \beta\gamma\beta\mathbf{f} \in V$, so that $\beta^{-1}V = (\beta\gamma\beta)^{-1}V$, as required.

(c)

$$\beta\alpha\mathbf{f} \in V \iff 01 \bullet \beta\alpha\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(00\mu(\overline{\mathbf{x}})) \in \mathcal{PF}(7/3) \iff \\ \mu(11\mu(\mathbf{x})) \in \mathcal{PF}(7/3) \iff 11\mu(\mathbf{x}) \in \mathcal{PF}(7/3)$$

$$\beta\beta\alpha\mathbf{f} \in V \iff 01 \bullet \beta\beta\alpha\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(0\mu(00\mu(\overline{\mathbf{x}}))) \in \mathcal{PF}(7/3) \iff \\ \mu(1\mu(11\mu(\mathbf{x}))) \in \mathcal{PF}(7/3) \iff 1\mu(11\mu(\mathbf{x})) \in \mathcal{PF}(7/3) \iff 11\mu(\mathbf{x}) \in \mathcal{PF}(7/3)$$

One direction of the last equivalence is clear. To see the other, suppose that $\mathbf{z} = 11\mu(\mathbf{x}) \in \mathcal{PF}(7/3)$ but $\mathbf{y} = 1\mu(\mathbf{z}) \notin \mathcal{PF}(7/3)$. Any 7/3-power in $\mathbf{y}$ is necessarily a prefix of $\mathbf{y}$. Note that $\mathbf{y}$ begins with 1101001101001. In fact, $\mathbf{y}$ must begin with 110100110100101; otherwise, $\mu(\mathbf{z})$ contains the 7/3-power 10100110100110. The prefix 1101001101001 cannot occur elsewhere in $\mathbf{y}$ as that would imply that $\mu(\mathbf{z})$ contains either the cube 111 or the 7/3-power 01101001101001, a contradiction. Thus the period of any 7/3-power in $\mathbf{y}$ is at most 12. It is easy to verify that no such 7/3-power exists.

Thus, $\beta\alpha\mathbf{f} \in V \iff \beta\beta\alpha\mathbf{f} \in V$, so that $(\beta\alpha)^{-1}V = (\beta\beta\alpha)^{-1}V$, as required.

(d)

$$\beta\gamma\alpha\mathbf{f} \in V \iff 01 \bullet \beta\gamma\alpha\mathbf{f} \in \mathcal{PF}(7/3) \iff \mu(0\mu(1\mu(\mathbf{x}))) \in \mathcal{PF}(7/3) \iff \\ 0\mu(1\mu(\mathbf{x})) \in \mathcal{PF}(7/3)$$

$$\begin{gathered}\beta\beta\gamma\alpha\mathbf{f} \in V \iff 01 \bullet \beta\beta\gamma\alpha\mathbf{f} \in \mathcal{PF}(7/3) \iff \\ \mu(0\mu(0\mu(1\mu(\mathbf{x})))) \in \mathcal{PF}(7/3) \iff 0\mu(0\mu(1\mu(\mathbf{x}))) \in \mathcal{PF}(7/3) \iff \\ 0\mu(1\mu(\mathbf{x})) \in \mathcal{PF}(7/3)\end{gathered}$$

One direction of the last equivalence is clear. To see the other, suppose that $\mathbf{z} = 0\mu(1\mu(\mathbf{x})) \in \mathcal{PF}(7/3)$ but $\mathbf{y} = 0\mu(z) \notin \mathcal{PF}(7/3)$. Any 7/3-power in $\mathbf{y}$ is necessarily a prefix of $\mathbf{y}$. Note that $\mathbf{y}$ begins with

$$0011001011010011001011 0.$$

In fact, $\mathbf{y}$ must begin with

$$00110010110100110010110100101;$$

otherwise, $\mu(z)$ contains the 7/3-power

$$0110010110100110010110100110.$$

The prefix

$$00110010110100110010110$$

cannot occur elsewhere in $\mathbf{y}$ as that would imply that $\mu(\mathbf{z})$ contains either the cube 000, one of the 5/2-powers 10101 or 0110011001, or the 7/3-power

$$0110100110010110100110010110,$$

a contradiction. Thus the period of any 7/3-power in y is at most 22. It is easy to verify that no such 7/3-power exists.

Thus, $\beta\gamma\alpha\mathbf{f} \in V \iff \beta\beta\gamma\alpha\mathbf{f} \in V$, so that $(\beta\gamma\alpha)^{-1}V = (\beta\beta\gamma\alpha)^{-1}V$, as required. □

These are only partial results; the following problem remains open.

Problem 2.18. *Does there exist a characterization similar to that of Theorem 2.14 of the infinite* 7/3*-power-free binary words?*

2.9 Transcendence of overlap-free base-b expansions

We have now developed enough of the theory of combinatorics on words to consider some applications to problems in number theory. In this section we consider infinite words as base-b expansions of some real number, where $b \geq 2$ is an integer (we shall give a result involving an expansion in an irrational base in Section 6.8)[3]. If

[3] The results in this section can be found in Adamczewski and Rampersad [9].

$\mathbf{a} = a_0a_1a_2\cdots$ is an infinite word over the alphabet $\{0, 1, \ldots, b-1\}$, then let $\xi_{\mathbf{a}}$ be the real number defined by

$$\xi_{\mathbf{a}} = \sum_{n\geq 0} a_n b^{-n}.$$

We are particularly interested in what properties of $\mathbf{a}$ imply the transcendence of $\xi_{\mathbf{a}}$. Recall that a real number is *transcendental* if it is not the root of some polynomial with integer coefficients. For a general introduction to number theory, see Hardy and Wright [116]; for an overview of the application of combinatorics on words to Diophantine approximation, see the survey of Waldschmidt [240].

Mahler [166] showed that the real number

$$\tau = \sum_{n\geq 0} t_n 2^{-n} \approx 0.824890137\cdots,$$

where $\mathbf{t} = t_0t_1t_2\cdots$ is the Thue–Morse word, is transcendental. For a proof of this result, see Nishioka [183, Example 1.3.1]. Allouche and Shallit [23, Theorem 13.4.2] gave a corrected version of a proof of this result due to Dekking [77].

Ferenczi and Mauduit [95] used Ridout's p-adic generalization [207] of the Thue–Siegel–Roth Theorem [210] to prove that if $\mathbf{a}$ is binary and for all k, $\mathbf{a}$ has $k+1$ subwords of length k, then $\xi_{\mathbf{a}}$ is transcendental for any base b.

Allouche and Zamboni [24] combined the argument of Ferenczi and Mauduit with the result of Theorem 2.7 to prove that if $\mathbf{a}$ is the infinite fixed point of a uniform binary morphism, then $\xi_{\mathbf{a}}$ is either rational or transcendental for any base b.

Adamczewski, Bugeaud, and Luca [8] used Schlickewei's p-adic generalization [218] of Schmidt's Subspace Theorem (see Schmidt [219]) to give a remarkably short and elegant proof that the result of Allouche and Zamboni can be extended to fixed points of uniform morphisms over any alphabet. Adamczewski and Bugeaud [7] gave an English version of this proof. Loxton and van der Poorten [162, 163] had previously (and incorrectly—see Becker [33]) claimed a proof.

Adamczewski, Bugeaud, and Luca [8] gave a general criterion for the transcendence of $\xi_{\mathbf{a}}$ that relies on the existence of arbitrarily large repetitions occurring near the beginning of $\mathbf{a}$. We state this criterion as the following theorem.

Theorem 2.19 (Adamczewski, Bugeaud, and Luca)**.** *Suppose there exists a real $\alpha > 1$ such that for all $n \geq 1$ the following conditions hold:*

- $\mathbf{a}$ *begins with a prefix of the form $U_nV_n^{\alpha}$;*
- *the sequence $(|U_n|/|V_n|)_{n\geq 1}$ is bounded; and,*
- *the sequence $(|V_n|)_{n\geq 1}$ is strictly increasing.*

Then $\xi_{\mathbf{a}}$ is either rational or transcendental for any base b.

Combining the result of this theorem with the result of Theorem 1.10 leads to the following theorem regarding overlap-free words. This result generalizes the earlier work of Mahler [166].

Theorem 2.20. *Let $b \geq 2$ be an integer. If $\mathbf{a}$ is an infinite overlap-free word over any binary subset of $\{0, 1, \ldots, b-1\}$, then $\xi_{\mathbf{a}}$ is transcendental in base b.*

Proof. For any $k \geq 1$, we may iterative apply Theorem 1.10 k times to $\mathbf{a}$ to obtain a factorization

$$\mathbf{a} = u_1\mu(u_2)\mu^2(u_3)\cdots\mu^{k-1}(u_k)\mu^k(\mathbf{y}'),$$

where each u_i has length at most 2 and $\mathbf{y}'$ is an infinite overlap-free word. But now the prefix of length 4 of $\mathbf{y}'$ must contain a square xx, where the length of x is either 1 or 2. Thus, $\mu^k(\mathbf{y}')$ contains a square $V_k^2 = \mu^k(xx)$ with $|V_k|$ either 2^k or 2^{k+1}. If we write

$$U_k = u_1\mu(u_2)\mu^2(u_3)\cdots\mu^{k-1}(u_k),$$

then $|U_k| \leq 2^{k+2}$. We thus have $|U_k|/|V_k| \leq 4$. Without loss of generality we may take $(|V_k|)_{k\geq 1}$ to be increasing (if not, we simply consider a subsequence). Thus for $\alpha = 2$, $\mathbf{a}$ satisfies the conditions of Theorem 2.19. We conclude that $\xi_{\mathbf{a}}$ is transcendental. □

If we apply Theorem 1.11 in the proof of Theorem 2.20, instead of Theorem 1.10, we obtain the following stronger result.

Theorem 2.21. *If $\mathbf{a}$ is an infinite $7/3$-power-free word over any binary subset of $\{0, 1, \ldots, b-1\}$, then $\xi_{\mathbf{a}}$ is transcendental for any base b.*

This result can be used to derive some information concerning the occurrences of certain patterns of symbols in the binary expansions of algebraic numbers. The study of such occurrences of symbols was initiated by Borel [47]. It is widely believed that for any base b, each subword of length n over the alphabet $\{0, 1, \ldots, b-1\}$ should occur with limiting frequency $1/b^n$ in the base-b expansion of any algebraic irrational number. This conjecture, however, seems to be far out of the reach of current mathematics. It remains one of the important open problems in number theory[4].

In this regard we have the following corollary of Theorem 2.21.

Corollary 2.22. *The binary expansion of an algebraic number contains infinitely many occurrences of $7/3$-powers.*

[4] According to Borel [47]: "En définitive, le problème de savoir si les chiffres d'un nombre tel que $\sqrt{2}$ satisfont ou non à *toutes* les lois que l'on peut énoncer pour des chiffres choisis au hasard me paraît toujours être un des problèmes les plus importants qui se posent aux mathématiciens."

"All things considered, the problem of knowing whether or not the digits of a number such as $\sqrt{2}$ satisfies *all* the laws that one can state for numbers selected at random seems to me still to be one of the most important problems facing mathematicians."

When considering the ternary expansion of an algebraic number, we would like to be able to prove that such an expansion must contain infinitely many occurrences of squares; however, we are only able to prove the following weaker result.

Theorem 2.23. *The ternary expansion of an algebraic number contains either infinitely many occurrences of squares or infinitely many occurrences of one of the blocks* 010 *or* 02120.

Proof. Every rational number contains infinitely many squares in its ternary expansion, so it suffices to consider algebraic irrational numbers. To prove the desired result it is enough to show that any infinite ternary word that avoids squares as well as the subwords 010 and 02120 satisfies the conditions of Theorem 2.19.

Let $\mathbf{a} = a_0 a_1 a_2$ be a ternary word avoiding squares as well as 010 and 02120. Define a morphism τ by $\tau(0) = 011$, $\tau(1) = 01$, and $\tau(2) = 0$. From Theorem 2.3 we see that $\mathbf{b} = \tau(\mathbf{a})$ is overlap-free. We may thus apply the same argument as in the proof of Theorem 2.20 to show that for every integer $k \geq 1$, $\mathbf{b}$ begins with a prefix $U_k V_k^2$, where $|U_k| \leq 2^{k+2}$ and $|V_k| \geq 2^k$.

Observe that for any overlap-free binary word, between any two successive occurrences of 0 there can be at most two 1's; otherwise, we would have an occurrence of the overlap 111. By considering the number of 1's between every two successive occurrences of 0, we see that the word $\mathbf{b}$ has a unique factorization into the blocks 011, 01, and 0.

Let x and y be the shortest words such that xV_k^2y is a subword of $\mathbf{b}$ and xV_k^2y begins and ends with 0. Now write

$$xV_k^2y = 01^{i_1}01^{i_2} \cdots 01^{i_\ell}0,$$

where $i_j \in \{0, 1, 2\}$ for $j = 1, \ldots, \ell$. Set $W_k = w_1 w_2 \ldots w_\ell$, where $w_j = (2 - i_j)$ for $1 \leq j \leq \ell$. Then, W_k is a subword of $\mathbf{a}$ and has either the form $XaXb$ or $bXaX$, where $a, b \in \{0, 1, 2\}$ and $a \neq b$.

It follows that for every $k \geq 1$, there exist words Y_k and Z_k, and a letter $a \in \{0, 1, 2\}$, such that $\mathbf{a}$ begins with $Y_k Z_k a Z_k$, where

$$|Y_k| \leq 2^{k+2} \text{ and } |Z_k| \geq \frac{2^k}{3}.$$

Thus $\mathbf{a}$ satisfies the conditions of Theorem 2.19 for every α, $1 < \alpha < 2$. This concludes the proof. □

In Section 6.8 we shall see another application of the theory of combinatorics on words to number theory. In the next chapter, however, we return to the study of words, in particular, to the study of 7/3-power-free words.

Chapter 3

Words Avoiding 7/3-powers and the Thue–Morse Morphism

3.1 Introduction

In this chapter we present one of the main results of the thesis[1]. Our main result is a generalization of Theorem 2.7: we show that the Thue–Morse word and its complement are the only infinite 7/3-power-free binary words that can be obtained by iteration of a morphism. At first glance, it may seem that this is an immediate consequence of Theorem 1.6; however, this is not necessarily so, as there are infinite 7/3-power-free binary words that cannot be extended to the left to form bi-infinite 7/3-power-free binary words. For example, if we denote the complement of the Thue–Morse word by $\overline{\mathbf{t}}$, the infinite binary word $\mathbf{s} = 001001\overline{\mathbf{t}}$ was shown by Allouche, Currie, and Shallit [19] to be the lexicographically least infinite overlap-free binary word; however, $\mathbf{s}$ cannot be extended to the left to form a 7/3-power-free word: prepending a 0 creates the cube 000, and prepending a 1 creates the 7/3-power 1001001.

3.2 Preliminary lemmas

We begin by establishing a few preliminary lemmas. Lemma 3.1 is analogous to a similar lemma for overlap-free words given in Allouche and Shallit [23, Lemma 1.7.6]. This result for overlap-free words was also stated without formal proof by Berstel and Séébold [43].

Lemma 3.1. *Let $w \in \{0,1\}^*$ be a 7/3-power-free word with $|w| \geq 52$. Then w contains both $\mu^3(0) = 01101001$ and $\mu^3(1) = 10010110$ as subwords.*

[1] The results in this chapter can be found in Rampersad [202].

Proof. Since w is 7/3-power-free, by Theorem 1.11 we can write

$$w = u\mu(y)v, \tag{3.1}$$

where y is 7/3-power-free and $|y| \geq 24$. Similarly, we can write

$$y = u'\mu(y')v', \tag{3.2}$$

where y' is 7/3-power-free and $|y'| \geq 10$. Again, we can write

$$y' = u''\mu(y'')v'', \tag{3.3}$$

where y'' is 7/3-power-free and $|y''| \geq 3$. From (1)–(3), we get

$$\begin{aligned} w &= u\mu(u'\mu(u''\mu(y'')v'')v')v \\ &= u\mu(u')\mu^2(u'')\mu^3(y'')\mu^2(v'')\mu(v')v, \end{aligned}$$

where $u, u', u'', v, v', v'' \in \{\epsilon, 0, 1, 00, 11\}$. Since y'' is 7/3-power-free and $|y''| \geq 3$, y'' contains both 0 and 1, and so $\mu^3(y'')$, and consequently w, contains both $\mu^3(0) = 01101001$ and $\mu^3(1) = 10010110$ as subwords as required. □

Lemma 3.2. *Let w' be a subword of $w \in \{0,1\}^*$, where w' is either of the form $abb\mu(w'')$ or $\mu(w'')bba$ for some $a, b \in \{0,1\}$ and $w'' \in \{0,1\}^*$. Suppose also that $a \neq b$ and $|w''| \geq 2$. Then w contains a 7/3-power.*

Proof. Suppose $ab = 10$ and $w' = 100\mu(w'')$ (the other cases follow similarly). The word $\mu(w'')$ may not begin with a 0 as that would create the cube 000. Hence we have $w' = 10010\mu(w''')$ for some $w''' \in \{0,1\}^*$. If $\mu(w''')$ begins with 01, then w' contains the 7/3-power 1001001. If $\mu(w''')$ begins with 10, then w' contains the 5/2-power 01010. Hence, w contains a 7/3-power. □

Lemma 3.3. *For $i, j \in \mathbb{N}$, let w be a 7/3-power-free word over $\{0,1\}$ such that $|w| = (7+2j)2^i - 1$. Let a be an element of $\{0,1\}$. Then waw contains a 7/3-power x, where $|x| \leq 7 \cdot 2^i$.*

Proof. Suppose $a = 1$ (the case $a = 0$ follows similarly). The proof is by induction on i. For the base case we have $i = 0$. Hence, $|w| \geq 6$ and $|w|$ is even. If w either begins or ends with 11, then $w1w$ contains the cube 111, and the result follows. Suppose then that w neither begins nor ends with 11. By explicitly examining all 13 words of length six that avoid 7/3-powers and neither begin nor end with 11, we see that all such words of length at least six can be written in the form $pbbq$, where $p, q \in \{0,1\}^+$ and $b \in \{0,1\}$. Hence, $w1w$ must have at least one subword with prefix bb and suffix bb. Moreover, since $|w|$ is even, there must exist such a subword where the prefix bb and the suffix bb each begin at positions of different parity in $w1w$. Let x be a smallest such subword such that $w1w$ neither begins nor ends with x. Suppose $b = 0$ (the case $b = 1$ follows similarly). Then $x = 000$, $x = 00100$, or x contains a subword 01010 or 10101. Hence, $w1w$ contains one of the subwords 000, 01010, 10101, or 1001001 as required.

Let us now assume that the lemma holds for all i', where $0 < i' < i$. Since w avoids 7/3-powers, and since $|w| \geq 7$, by Theorem 1.11 we can write $w = u\mu(w')v$, where $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and $w' \in \{0,1\}^*$ is 7/3-power-free. By applying a case analysis similar to that used in Cases (1)–(4) of the proof of Theorem 3.7 below, we can eliminate all but three cases: $(u, v) \in \{(\epsilon, \epsilon), (\epsilon, 0), (0, \epsilon)\}$.

1. $(u, v) = (\epsilon, \epsilon)$. In this case $w = \mu(w')$. This is clearly not possible, since for $i > 0$, $|w| = (7 + 2j)2^i - 1$ is odd.

2. $(u, v) = (\epsilon, 0)$. Then $w = \mu(w')0$ and $w1w = \mu(w')01\mu(w')0 = \mu(w'0w')0$. If $|w| = (7 + 2j)2^i - 1$, we see that $|w'| = (7 + 2j)2^{i-1} - 1$. Hence, if $i' = i - 1$, we may apply the inductive assumption to $w'0w'$. We thus obtain that $w'0w'$ contains a 7/3-power x', where $|x'| \leq 7 \cdot 2^{i-1}$, and so $w1w$ must contain a 7/3-power $x = \mu(x')$, where $|x| \leq 7 \cdot 2^i$.

3. $(u, v) = (0, \epsilon)$. This case is handled similarly to the previous case, and we omit the details.

By induction then, we have that waw contains a 7/3-power x, where $|x| \leq 7 \cdot 2^i$. □

Lemma 3.4. *For $i \in \mathbb{N}$, let w be a 7/3-power-free word over $\{0,1\}$ such that $|w| = 5 \cdot 2^i - 1$. Let a be an element of $\{0,1\}$. Then waw contains a 7/3-power x, where $|x| \leq 5 \cdot 2^i$.*

Proof. Suppose $a = 1$ (the case $a = 0$ follows similarly). The proof is by induction on i. For the base case we have $i = 0$ and $|w| = 4$. An easy computation suffices to verify that for all w with $|w| = 4$, $w1w$ contains a 7/3-power x, where $|x| \leq 5$ as required.

Let us now assume that the lemma holds for all i', where $0 < i' < i$. Since w avoids 7/3-powers, and since $|w| \geq 7$, by Theorem 1.11 we can write $w = u\mu(w')v$, where $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and $w' \in \{0,1\}^*$ is 7/3-power-free. By applying a case analysis similar to that used in Cases (1)–(4) of the proof of Theorem 3.7 below, we can eliminate all but three cases: $(u, v) \in \{(\epsilon, \epsilon), (\epsilon, 0), (0, \epsilon)\}$.

1. $(u, v) = (\epsilon, \epsilon)$. In this case $w = \mu(w')$. This is clearly not possible, since for $i > 0$, $|w| = 5 \cdot 2^i - 1$ is odd.

2. $(u, v) = (\epsilon, 0)$. Then $w = \mu(w')0$ and $w1w = \mu(w')01\mu(w')0 = \mu(w'0w')0$. If $|w| = 5 \cdot 2^i - 1$, we see that $|w'| = 5 \cdot 2^{i-1} - 1$. Hence, if $i' = i - 1$, we may apply the inductive assumption to $w'0w'$. We thus obtain that $w'0w'$ contains a 7/3-power x', where $|x'| \leq 5 \cdot 2^{i-1}$, and so $w1w$ must contain a 7/3-power $x = \mu(x')$, where $|x| \leq 5 \cdot 2^i$.

3. $(u, v) = (0, \epsilon)$. This case is handled similarly to the previous case, and we omit the details.

By induction then, we have that waw contains a 7/3-power x, where $|x| \leq 5 \cdot 2^i$. □

Lemma 3.5. *For $i, j \in \mathbb{Z}^+$, let w and s be 7/3-power-free words over $\{0,1\}$ such that $|w| = 2^{i+1} - 1$ or $|w| = 3 \cdot 2^i - 1$, and $|s| = 2^{j+1} - 1$ or $|s| = 3 \cdot 2^j - 1$. Assume also that $|s| \geq |w|$. Let a be an element of $\{0,1\}$. Then $sawawas$ contains a 7/3-power.*

Proof. Suppose $a = 1$ (the case $a = 0$ follows similarly). The proof is by induction on i. For the base case we have $i = 1$ and either $|w| = 3$ or $|w| = 5$. An easy computation suffices to verify that for all w with $|w| = 3$ or $|w| = 5$, and all $a, b \in \{0,1\}^2$, $a1w1w1b$ contains a 7/3-power.

Let us now assume that the lemma holds for all i', where $1 < i' < i$. Since w avoids 7/3-powers, and since $|w| \geq 7$, by Theorem 1.11 we can write $w = u\mu(w')v$, where $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and $w' \in \{0,1\}^*$ is 7/3-power-free. Similarly, we can write $s = u'\mu(s')v'$, where $u', v' \in \{\epsilon, 0, 1, 00, 11\}$ and $s' \in \{0,1\}^*$ is 7/3-power-free. By applying a case analysis similar to that used in Cases (1)–(4) of the proof of Theorem 3.7 below, we can eliminate all but three cases: $(u, v, u', v') \in \{(\epsilon, \epsilon, \epsilon, \epsilon), (\epsilon, 0, 0, \epsilon), (0, \epsilon, \epsilon, 0)\}$.

1. $(u, v, u', v') = (\epsilon, \epsilon, \epsilon, \epsilon)$. In this case $w = \mu(w')$. This is clearly not possible, since for $i > 1$, both $|w| = 2^{i+1} - 1$ and $|w| = 3 \cdot 2^i - 1$ are odd.

2. $(u, v, u', v') = (\epsilon, 0, \epsilon, 0)$. Then $w = \mu(w')0$, $s = \mu(s')0$, and
$$s1w1w1s = \mu(s')01\mu(w')01\mu(w')01\mu(s')0 = \mu(s'0w'0w'0s')0.$$
If $|w| = 2^{i+1} - 1$ or $|w| = 3 \cdot 2^i - 1$, we see that $|w'| = 2^i - 1$ or $|w| = 3 \cdot 2^{i-1} - 1$. Similarly, if $|s| = 2^{j+1} - 1$ or $|s| = 3 \cdot 2^j - 1$, we see that $|s'| = 2^j - 1$ or $|s| = 3 \cdot 2^{j-1} - 1$. Hence, if $i' = i - 1$, we may apply the inductive assumption to $s'0w'0w'0s'$. We thus obtain that $s'0w'0w'0s'$ contains a 7/3-power x', and so $s1w1w1s$ must contain a 7/3-power $x = \mu(x')$.

3. $(u, v, u', v') = (0, \epsilon, 0, \epsilon)$. This case is handled similarly to the previous case, and we omit the details.

By induction then, we have that $sawawas$ contains a 7/3-power. □

Lemma 3.6. *Let n be a positive integer. Then n can be written in the form $2^i - 1$, $3 \cdot 2^i - 1$, $5 \cdot 2^i - 1$, or $(7 + 2j)2^i - 1$ for some $i, j \in \mathbb{N}$.*

Proof. If $n = 1$ then $n = 2^1 - 1$ as required. Suppose then that $n > 1$. Then we may write $n - 1 = m2^i$, where m is odd and $i \in \mathbb{N}$. But for any odd positive integer m, either $m \in \{1, 3, 5\}$, or m is of the form $7 + 2j$ for some $j \in \mathbb{N}$, and the result follows. □

3.3 Main theorem

Let $h : \Sigma^* \to \Sigma^*$ be a morphism. We say that h is *non-erasing* if, for all $a \in \Sigma$, $h(a) \neq \epsilon$. Let E be the morphism defined by $E(0) = 1$ and $E(1) = 0$. The following theorem is analogous to a result regarding overlap-free words due to Berstel and Séébold [43].

Theorem 3.7. *Let $h : \{0,1\}^* \to \{0,1\}^*$ be a non-erasing morphism. If $h(01101001)$ is 7/3-power-free, then there exists an integer $k \geq 0$ such that either $h = \mu^k$ or $h = E \circ \mu^k$.*

Proof. Let $h(0) = x$ and $h(1) = x'$ with $|x|, |x'| \geq 1$. The proof is by induction on $|x| + |x'|$. If $|x| < 7$ and $|x'| < 7$, then a quick computation suffices to verify that if $h(01101001)$ is 7/3-power-free, then either $h = \mu^k$ or $h = E \circ \mu^k$, where $k \in \{0, 1, 2\}$. Let us assume then, without loss of generality, that $|x| \geq |x'|$ and $|x| \geq 7$. The word x must avoid 7/3-powers, and so, by Theorem 1.11, we can write $x = u\mu(y)v$, where $u, v \in \{\epsilon, 0, 1, 00, 11\}$ and $y \in \{0,1\}^*$. We shall consider all 25 choices for (u, v).

1. $(u, v) \in \{(0, 00), (00, 0), (00, 00), (1, 11), (11, 1), (11, 11)\}$. Suppose $(u, v) = (0, 00)$. Then $h(00) = 0\mu(y)000\mu(y)00$ contains the cube 000, contrary to the assumptions of the theorem. The argument for the other choices for (u, v) follows similarly.

2. $(u, v) \in \{(0, 11), (00, 1), (00, 11), (1, 00), (11, 0), (11, 00)\}$. For any of these choices for (u, v), $h(00) = u\mu(y)vu\mu(y)v$ contains a subword of the form $abb\mu(y)$ or $\mu(y)bba$ for some $a, b \in \{0, 1\}$, where $a \neq b$. Since $|x| \geq 7$, $|y| \geq 2$, and so by Lemma 3.2 we have that $h(00)$ contains a 7/3-power, contrary to the assumptions of the theorem.

3. $(u, v) \in \{(\epsilon, 0), (0, \epsilon), (\epsilon, 1), (1, \epsilon)\}$. Suppose $(u, v) = (0, \epsilon)$. Then $h(00) = 0\mu(y)0\mu(y)$. We have two subcases.

   3a: $\mu(y)$ begins with 01 or ends with 10. Then by Lemma 3.2, $h(00)$ contains a 7/3-power, contrary to the assumptions of the theorem.

   3b: $\mu(y)$ begins with 10 and ends with 01. Then $h(00) = 0\mu(y')01010\mu(y'')$ contains the 5/2-power 01010, contrary to the assumptions of the theorem.

   The argument for the other choices for (u, v) follows similarly.

4. $(u, v) \in \{(\epsilon, 00), (0, 0), (00, \epsilon), (\epsilon, 11), (1, 1), (11, \epsilon)\}$. Suppose $(u, v) = (00, \epsilon)$. Then $h(00) = 00\mu(y)00\mu(y)$. The word $\mu(y)$ may not begin with a 0 as that would create the cube 000. We have then that $h(00) = 00\mu(y)0010\mu(y')$ for some $y' \in \{0,1\}^*$. By Lemma 3.2, $h(00)$ contains a 7/3-power, contrary to the assumptions of the theorem. The argument for the other choices for (u, v) follows similarly.

5. $(u, v) \in \{(0,1),(1,0)\}$. Suppose $(u, v) = (0,1)$. By Lemma 3.6, the following three subcases suffice to cover all possibilities for $|y|$.

   5a: $|y| = (7+2j)2^i - 1$ for some $i, j \in \mathbb{N}$. We have $h(00) = 0\mu(y)10\mu(y)1 = 0\mu(y1y)1$. By Lemma 3.3, $y1y$ contains a 7/3-power. The word $h(00)$ must then contain a 7/3-power, contrary to the assumptions of the theorem.

   5b: $|y| = 5 \cdot 2^i - 1$ for some $i \in \mathbb{N}$. Again we have $h(00) = 0\mu(y)10\mu(y)1 = 0\mu(y1y)1$. By Lemma 3.4, $y1y$ contains a 7/3-power. The word $h(00)$ must then contain a 7/3-power, contrary to the assumptions of the theorem.

   5c: $|y| = 2^i - 1$ or $|y| = 3 \cdot 2^i - 1$ for some $i \in \mathbb{N}$. We have two subcases.

   5c.i: $|x'| < 7$. We have $h(0110) = 0\mu(y)1x'x'0\mu(y)1$. The only $x' \in \{0,1\}^*$ where $|x'| < 7$ and $1x'x'0$ does not contain a 7/3-power is

$$x' \in \{10, 0110, 1001, 011010, 100110, 101001\}.$$

   However, each of these words either begins or ends with 10, and so we have that $h(0110)$ contains a subword of the form $100\mu(y)$ or $\mu(y)110$. Hence, by Lemma 3.2 we have that $h(0110)$ contains a 7/3-power, contrary to the assumptions of the theorem.

   5c.ii: $|x'| \geq 7$. By Theorem 1.11, we can write $x' = u'\mu(z)v'$, where $u', v' \in \{\epsilon, 0, 1, 00, 11\}$ and $z \in \{0,1\}^*$ is 7/3-power-free. Applying the preceding case analysis to x' allows us to eliminate all but three subcases.

   5c.ii.A: $(u', v') = (0,1)$. We have

$$h(0110) = 0\mu(y)10\mu(z)10\mu(z)10\mu(y)1 = 0\mu(y1z1z1y)1.$$

   Moreover, by the same reasoning used in Case 5a and Case 5b, we have $|z| = 2^j - 1$ or $|z| = 3 \cdot 2^j - 1$ for some $j \in \mathbb{N}$, and so by Lemma 3.5, $y1z1z1y$ contains a 7/3-power. The word $h(0110)$ must then contain a 7/3-power, contrary to the assumptions of the theorem.

   5c.ii.B: $(u', v') = (1,0)$. Then $h(01) = 0\mu(y)11\mu(z)0$. The word $\mu(z)$ may not begin with a 1 as that would create the cube 111. We have then that $h(01) = 0\mu(y)1101\mu(z')0$ for some $z' \in \{0,1\}^*$. By Lemma 3.2, $h(01)$ contains a 7/3-power, contrary to the assumptions of the theorem.

   5c.ii.C: $(u', v') = (\epsilon, \epsilon)$. Then $h(01) = 0\mu(y)1\mu(z)$. We have two subcases.

   - $\mu(z)$ begins with 01. Then $h(01) = 0\mu(y)101\mu(z')$ for some $z' \in \{0,1\}^*$. The word $\mu(y)$ may not end in 10 as that would create the 5/2-power 10101. Hence $h(01) = 0\mu(y')01101\mu(z')$

for some $y' \in \{0,1\}^*$. If $\mu(z')$ begins with 10, then $h(01)$ contains the 7/3-power 0110110. If $\mu(z')$ begins with 01, then $h(01)$ contains the 5/2-power 10101. Either situation contradicts the assumptions of the theorem.

- $\mu(z)$ begins with 10. Then $h(01) = 0\mu(y)110\mu(z')$ for some $z' \in \{0,1\}^*$. By Lemma 3.2, $h(01)$ contains a 7/3-power, contrary to the assumptions of the theorem.

The argument for the other choice for (u, v) follows similarly.

6. $(u, v) = (\epsilon, \epsilon)$. In this case we have $x = \mu(y)$.

All cases except $x = \mu(y)$ lead to a contradiction. The same reasoning applied to x' gives $x' = \mu(y')$ for some $y' \in \{0,1\}^*$. Let the morphism h' be defined by $h'(0) = y$ and $h'(1) = y'$. Then $h = \mu \circ h'$, and by Theorem 1.6, $h'(01101001)$ is 7/3-power-free. Moreover, $|y| < |x|$ and $|y'| < |x'|$. Also note that for the preceding case analysis it sufficed to consider the following words only: $h(00)$, $h(01)$, $h(10)$, $h(11)$, $h(0110)$, $h(1001)$, and $h(01101001)$. However, 00, 01, 10, 11, 0110, and 1001 are all subwords of 01101001. Hence, the induction hypothesis can be applied, and we have that either $h' = \mu^k$ or $h' = E \circ \mu^k$. Since $E \circ \mu = \mu \circ E$, the result follows. □

We now establish the following corollary.

Corollary 3.8. *Let $h : \{0,1\}^* \to \{0,1\}^*$ be a morphism such that $h(01) \neq \epsilon$. Then the following statements are equivalent.*

(a) The morphism h is non-erasing, and $h(01101001)$ is 7/3-power-free.

(b) There exists $k \geq 0$ such that $h = \mu^k$ or $h = E \circ \mu^k$.

(c) The morphism h maps any infinite 7/3-power-free word to an infinite 7/3-power-free word.

(d) There exists an infinite 7/3-power-free word whose image under h is 7/3-power-free.

Proof.

(a) $\Longrightarrow$ (b) was proved in Theorem 3.7.

(b) $\Longrightarrow$ (c) follows from Theorem 1.6 via König's Infinity Lemma (Theorem 1.1).

(c) $\Longrightarrow$ (d): We need only exhibit an infinite 7/3-power-free word: the Thue–Morse word, $\mathbf{t}$, is overlap-free and so is 7/3-power-free.

(d) $\Longrightarrow$ (a): Let $\mathbf{w}$ be an infinite 7/3-power-free word whose image under h is 7/3-power-free. By Lemma 3.1, $\mathbf{w}$ must contain 01101001, and so $h(01101001)$ is 7/3-power-free.

To see that h is non-erasing, note that if $h(0) = \epsilon$, then since $h(01) \neq \epsilon$, $h(1) \neq \epsilon$. But then $h(01101001) = h(1)^4$ is not 7/3-power-free, contrary to what we have just shown. Similarly, $h(1) \neq \epsilon$, and so h is non-erasing. □

Let $h : \{0,1\}^* \to \{0,1\}^*$ be a morphism. We say that h is the *identity morphism* if $h(0) = 0$ and $h(1) = 1$. The following corollary gives the main result.

Corollary 3.9. *An infinite* 7/3*-power-free binary word is a fixed point of a non-identity morphism if and only if it is equal to the Thue–Morse word,* $\mathbf{t}$*, or its complement,* $\overline{\mathbf{t}}$*.*

Proof. Let $h : \{0,1\}^* \to \{0,1\}^*$ be a non-identity morphism, and let us assume that h has a fixed point that avoids 7/3-powers. Then h maps an infinite 7/3-power-free word to an infinite 7/3-power-free word, and so, by Corollary 3.8, h is of the form μ^k or $E \circ \mu^k$ for some $k \geq 0$. Since h has a fixed point, it is not of the form $E \circ \mu^k$, and since h is not the identity morphism, $h = \mu^k$ for some $k \geq 1$. But the only fixed points of μ^k are $\mathbf{t}$ and $\overline{\mathbf{t}}$, and the result follows. □

3.4 The constant 7/3 is best possible

It remains to show that the constant 7/3 given in Corollary 3.9 is best possible; i.e., Corollary 3.9 would fail to be true if 7/3 were replaced by any larger rational number. To show this, it suffices to exhibit an infinite binary word $\mathbf{w}$ that avoids $(7/3)^+$-powers, such that $\mathbf{w}$ is the fixed point of a morphism $h : \{0,1\}^* \to \{0,1\}^*$, where h is not of the form μ^k for any $k \geq 0$. Kolpakov *et al.* [145] have already given an example of such a word. Their example was the fixed point of a 21-uniform morphism; we shall give a similar solution using a 19-uniform morphism.

Let $h : \{0,1\}^* \to \{0,1\}^*$ be the morphism defined by

$$\begin{aligned} h(0) &= 0110100110110010110 \\ h(1) &= 1001011001001101001. \end{aligned}$$

Since $|h(0)| = |h(1)| = 19$, h is not of the form μ^k for any $k \geq 0$. We show that the fixed point $h^\omega(0)$ avoids $(7/3)^+$-powers by using a technique similar to that given by Karhumäki and Shallit [137]. We first state the following lemma, which may be easily verified computationally.

Lemma 3.10. *(a) Suppose $h(ab) = th(c)u$ for some letters $a, b, c \in \{0,1\}$ and words $t, u \in \{0,1\}^*$. Then this inclusion is trivial (that is, $t = \epsilon$ or $u = \epsilon$).*

(b) Suppose there exist letters $a, b, c \in \{0,1\}$ and words $s, t, u, v \in \{0,1\}^*$ such that $h(a) = st$, $h(b) = uv$, and $h(c) = sv$. Then either $a = c$ or $b = c$.

Theorem 3.11. *The fixed point $h^\omega(0)$ avoids $(7/3)^+$-powers.*

Proof. The proof is by contradiction. Let $w \in \{0,1\}^*$ avoid $(7/3)^+$-powers, and suppose that $h(w)$ contains a $(7/3)^+$-power. Then we may write $h(w) = xyyy'z$ for some $x, z \in \{0,1\}^*$ and $y, y' \in \{0,1\}^+$, where y' is a prefix of y, and $|y'|/|y| > 1/3$. Let us assume further that w is a shortest such string, so that $0 \le |x|, |z| < 19$. We shall consider two cases.

Case 1: $|y| \le 38$. In this case we have $|w| \le 6$. Checking all 20 words $w \in \{0,1\}^*$ of length 6 that avoid $(7/3)^+$-powers, we see that, contrary to our assumption, $h(w)$ avoids $(7/3)^+$-powers in every case.

Case 2: $|y| > 38$. Noting that if $h(w)$ contains a $(7/3)^+$-power, it must contain a square, we may apply a standard argument (see Karhumäki and Shallit [137] for an example) to show that Lemma 3.10 implies that $h(w)$ can be written in the following form:

$$h(w) = A_1A_2 \ldots A_jA_{j+1}A_{j+2} \ldots A_{2j}A_{2j+1}A_{2j+2} \ldots A_{n-1}A'_nA''_n,$$

for some j, where

$$\begin{aligned}
A_i &= h(a_i) \quad \text{for} \quad i = 1, 2, \ldots, n \quad \text{and} \quad a_i \in \{0,1\} \\
A_n &= A'_nA''_n \\
y &= A_1A_2 \ldots A_j \\
&= A_{j+1}A_{j+2} \ldots A_{2j} \\
y' &= A_{2j+1}A_{2j+2} \ldots A_{n-1}A'_n \\
z &= A''_n.
\end{aligned}$$

Since y' is a prefix of y, and since $|y'|/|y| > 1/3$, A'_n must be a prefix of A_k, where $k = \lfloor j/3 \rfloor + 1$. However, noting that for any $a \in \{0,1\}$, any prefix of $h(a)$ suffices to uniquely determine a, we may conclude that $A_k = A_n$. Hence, we may write

$$h(w) = A_1A_2 \ldots A_{k-1}A_k \ldots A_jA_{j+1}A_{j+2} \ldots A_{j+k-1}A_{j+k} \ldots A_{2j} \\ A_{2j+1}A_{2j+2} \ldots A_{n-1}A_n,$$

where

$$\begin{aligned}
y &= A_1A_2 \ldots A_{k-1}A_k \ldots A_j \\
&= A_{j+1}A_{j+2} \ldots A_{j+k-1}A_{j+k} \ldots A_{2j} \\
y'z &= A_{2j+1}A_{2j+2} \ldots A_{n-1}A_n \\
&= A_1A_2 \ldots A_{k-1}A_k.
\end{aligned}$$

We thus have

$$w = (a_1a_2 \ldots a_j)^2 a_1a_2 \ldots a_k,$$

where $k = \lfloor j/3 \rfloor + 1$. Hence, w is a $(7/3)^+$-power, contrary to our assumption. The result now follows. □

Theorem 3.11 thus implies that the constant 7/3 given in Corollary 3.9 is best possible. We note that Theorem 3.11 may also be proved using techniques recently developed by Krieger [149, 150].

We have thus demonstrated that the only infinite 7/3-power-free binary words that can be generated by iteration of a morphism are the Thue–Morse word and its complement. In the next chapter we continue our study of overlap-free words and 7/3-power-free words.

Chapter 4

Binary Words Containing Infinitely Many Overlaps

4.1 Introduction

The reader will have noticed by now that many of the properties of binary overlap-free words are shared by the binary 7/3-power-free words. In this chapter we first give an example of another such property by showing that the set of binary 7/3-power-free squares is identical to the set of binary overlap-free squares. We then contrast the 7/3-power-free words with the overlap-free words by showing that there exist infinite 7/3-power-free binary words containing infinitely many overlaps[1]. More generally, we show that for any real number $\alpha > 2$ there exists a real number β arbitrarily close to α such that there exists an infinite β^+-power-free binary word containing infinitely many β-powers.

4.2 Properties of the Thue-Morse morphism

In this section we present some useful properties of the *Thue-Morse morphism*—*i.e.*, the morphism μ defined by $\mu(0) = 01$ and $\mu(1) = 10$—which we shall use in later sections. We shall need the following sharper version of one direction of Theorem 1.6 (implicit in Karhumäki and Shallit [137]).

Theorem 4.1. *Suppose that for $w \in \{0,1\}^*$, $\mu(w)$ contains a subword u of period p, with $|u|/p > 2$. Then w contains a subword v of length $\lceil |u|/2 \rceil$ and period $p/2$.*

Recall from Chapter 2.7 that the words $\mu^n(0)$ and $\mu^n(1)$, $n \geq 0$, are known as *Morse blocks*. Note that the reverse of a Morse block is a Morse block. We now prove an analogue, for 7/3-power-free words, of a well-known "progression lemma" for overlap-free words [39, 96, 139, 141, 206, 226].

[1] Most of the results in this chapter can be found in Currie, Rampersad, and Shallit [72].

Lemma 4.2. *Let $w = uvxy$ be a binary 7/3-power-free word with $|u| = |v| = |x| = |y| = 2^n$. If u and v are Morse blocks, then x is a Morse block.*

Proof. The proof is by induction on n. Clearly, the result holds for $n = 0$. We have either (1) $w = u'u''u'u''pqrs$ or (2) $w = u'u''u''u'pqrs$, where u' and u'' are distinct Morse blocks of length 2^{n-1} and $|p| = |q| = |r| = |s| = 2^{n-1}$. By induction, p, q, and r are also Morse blocks. We must show that $p \neq q$. In case (1), $pq = u'u'$ creates the 5/2-power $u'u''u'u''u'$, and $pq = u''u''$ creates the cube $u''u''u''$. In case (2), $pq = u'u'$ creates the cube $u'u'u'$; and if $pq = u''u''$, then $r = u'$ creates the 7/3-power $u'u''u''u'u''u''u'$, and $r = u''$ creates the cube $u''u''u''$. Thus, $p \neq q$, as required. □

4.3 Overlap-free squares

Dekking [76] showed that any infinite overlap-free binary word must contain arbitrarily large squares. This is also an easy consequence of Theorem 1.10. Recall the sets

$$A = \{00, 11, 010010, 101101\}$$

and

$$\mathcal{A} = \bigcup_{k \geq 0} \mu^k(A)$$

defined in Section 2.4. The set $\mathcal{A}$ is the set of squares appearing in the Thue–Morse word; however, $\mathcal{A}$ does not contain all possible overlap-free squares. Shelton and Soni [226] characterized the overlap-free squares (the result is also attributed to Thue by Berstel [38]), but it is not hard to show that there are some overlap-free squares, such as 00110011, that cannot occur in an infinite overlap-free binary word. In this section, we characterize those overlap-free squares that do occur in infinite overlap-free binary words.

Theorem 4.3 (Shelton and Soni)**.** *The overlap-free binary squares are the conjugates of the words in $\mathcal{A}$.*

Some overlap-free squares cannot occur in any infinite overlap-free binary word, as the following lemma shows.

Lemma 4.4. *Let $x = \mu^k(z)$ for some $k \geq 0$ and $z \in \{011011, 100100\}$. Then xa contains an overlap for all $a \in \{0, 1\}$.*

Proof. It is easy to see that $x = uvvuvv$ for some $u, v \in \{0,1\}^*$, where u and v begin with different letters. Thus one of $uvvuvva$ or vva is an overlap. □

Let

$$B = \{001001, 110110\}$$

and let

$$\mathcal{B} = \bigcup_{k \geq 0} \mu^k(B).$$

Theorem 4.5. *The set of squares that can occur in an infinite overlap-free binary word is $\mathcal{A} \cup \mathcal{B}$. Furthermore, if $\mathbf{w}$ is an infinite overlap-free binary word containing a subword $x \in \mathcal{B}$, then $\mathbf{w}$ begins with x and there are no other occurrences of x in $\mathbf{w}$.*

Proof. Let $\mathbf{w}$ be an infinite overlap-free binary word beginning with a square $yy \notin \mathcal{A} \cup \mathcal{B}$. Suppose further that yy is a smallest such square that can be extended to an infinite overlap-free word. If $|y| \leq 3$, then $yy \notin \mathcal{A} \cup \mathcal{B}$ is one of 011011 or 100100, neither of which can be extended to an infinite overlap-free word by Lemma 4.4.

We assume then that $|y| > 3$. Since, by Theorem 4.3, yy is a conjugate of a word in $\mathcal{A}$, we have two cases.

Case 1: $yy = \mu(zz)$ for some $z \in \{0,1\}^*$. By Theorem 1.11, $\mathbf{w} = \mu(zz\mathbf{w}')$ for some infinite $\mathbf{w}'$, where $zz\mathbf{w}'$ is overlap-free. Thus zz is a smaller square not in $\mathcal{A} \cup \mathcal{B}$ that can be extended to an infinite overlap-free word, contrary to our assumption.

Case 2: $yy = a\mu(zz')\overline{a}$ for some $a \in \{0,1\}$ and $z, z' \in \{0,1\}^*$. By Theorem 1.11, yy is followed by a in $\mathbf{w}$, and so yya is an overlap, contrary to our assumption.

Since both cases lead to a contradiction, our assumption that $yy \notin \mathcal{A} \cup \mathcal{B}$ must be false.

To see that each word in $\mathcal{A} \cup \mathcal{B}$ does occur in some infinite overlap-free binary word, note that Allouche, Currie, and Shallit [19] have shown that the word $\mathbf{s} = 001001\overline{\mathbf{t}}$ is overlap-free. Now consider the words $\mu^k(\mathbf{s})$ and $\mu^k(\overline{\mathbf{s}})$, which are overlap-free for all $k \geq 0$.

Finally, to see that any occurrence of $x \in \mathcal{B}$ in $\mathbf{w}$ must occur at the beginning of $\mathbf{w}$, we note that by an argument similar to that used in Lemma 4.4, ax contains an overlap for all $a \in \{0,1\}$, and so x occurs at the beginning of $\mathbf{w}$. □

4.4 7/3-power-free squares

In this section we show that the characterization of the overlap-free binary squares given by Theorem 4.3 is precisely the characterization of the 7/3-power-free binary squares[2]. The proof is somewhat more difficult than Shelton and Soni's proof of Theorem 4.3 [226]. We begin with some lemmas.

Lemma 4.6. *Let $xx \in \{0,1\}^*$ be 7/3-power-free. If $xx = \mu(y)$, then $|y|$ is even. Consequently, y is a square.*

[2] The results in this section do not appear in Currie, Rampersad, and Shallit [72] and are presented here for the first time.

Proof. Suppose to the contary that $|y| = |x|$ is odd. By an exhaustive enumeration one verifies that $|x| \geq 5$. But then xx contains two occurrences of 00 (or 11) in positions of different parities, which is impossible. □

The next lemma is a version of Theorem 1.11 specifically applicable to squares.

Lemma 4.7. *Let $xx \in \{0,1\}^*$ be 7/3-power-free. If $|xx| > 8$, then either*

(a) $xx = \mu(y)$, where $y \in \{0,1\}^*$; or

(b) $xx = \overline{a}\mu(y)a$, where $a \in \{0,1\}$ and $y \in \{0,1\}^*$.

Proof. Applying Theorem 1.11, we write $xx = u\mu(y)v$. We first show that $|u| = |v| \leq 1$. Suppose that $u = 00$. Then xx begins with one of the words 000, 00100, or 001010. The first and third words contain a 7/3-power, a contradiction. The second word, 00100, cannot occur later in xx, as that would also imply the existence of a 7/3-power. We conclude $u \neq 00$, and similarly, $u \neq 11$. A similar argument also holds for v. Since $|xx|$ is even, we must therefore have $|u| = |v|$, as required.

If $|u| = |v| = 0$, then we have established (a). If $|u| = |v| = 1$, it remains to show that $u = \overline{v}$. We leave it to the reader to verify that if $u = v$, then xx contains either the cube uuu, the 5/2-power $u\overline{u}u\overline{u}u$, or the 7/3-power $\overline{u}uu\overline{u}uu\overline{u}$, which is a contradiction. □

We are now ready to prove the main result of this section.

Theorem 4.8. *The 7/3-power-free binary squares are the conjugates of the words in $\mathcal{A}$.*

Proof. Let xx be a minimal 7/3-power-free square that is not a conjugate of a word in $\mathcal{A}$. That $|xx| > 8$ is easily verified computationally. Applying Lemma 4.7 leads to two cases.

Case 1: $xx = \mu(y)$. By Lemma 4.6, y is a square. Furthermore, y is not a conjugate of a word in $\mathcal{A}$, contradicting the minimality of xx.

Case 2: $xx = \overline{a}\mu(y)a$. Then $a\overline{a}\mu(y) = \mu(ay)$ is also a square zz. We show that zz is 7/3-power-free, and consequently, by Lemma 4.6, that ay is a 7/3-power-free square, contradicting the minimality of xx.

Suppose to the contrary that zz contains a 7/3-power $s = rrr'$, where r' is a prefix of r and $|r'|/|r| \geq 1/3$. The word s must occur at the beginning of zz and we must have $|s| > |x|$; otherwise, xx would contain an occurrence of s, contradicting the assumption that xx is 7/3-power-free. We have four cases, depending on the relative sizes of $|r|$ and $|z|$, as illustrated in Figures 4.1–4.4.

By analyzing the overlaps between zz and rrr', denoted X in the figures, we derive a contradiction in each case. Figures 4.1 and 4.3 show that xx contains the cube XXX, a contradiction. In Figure 4.4, X' denotes a prefix of X, and we see

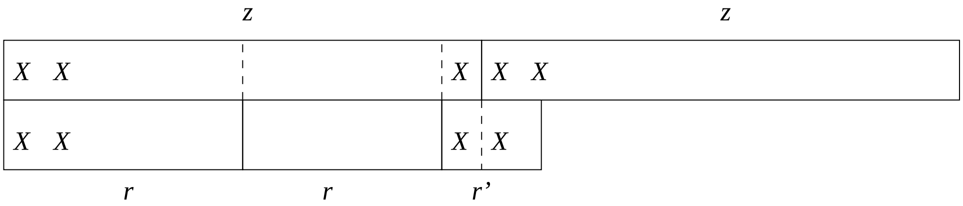

Figure 4.1: The case where $7/6 \cdot |r| \leq |z|$

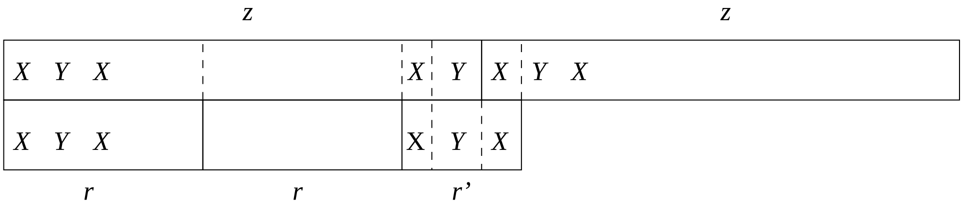

Figure 4.2: The case where $7/6 \cdot |r| > |z|$

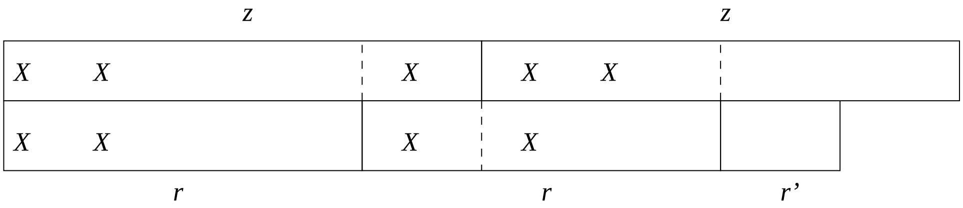

Figure 4.3: The case where $3/2 \cdot |r| \leq |z|$

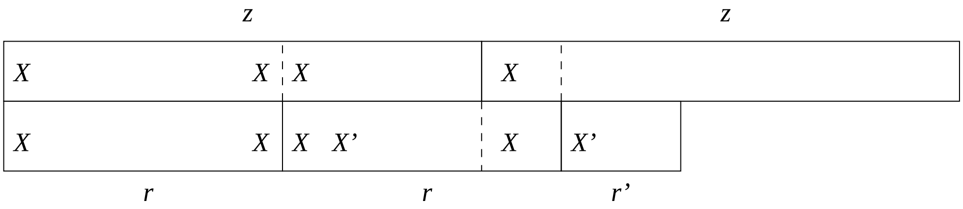

Figure 4.4: The case where $3/2 \cdot |r| > |z|$

that xx contains a subword XXX' that is at least a 7/3-power. This is again a contradiction. The case illustrated by Figure 4.2 requires some additional work.

We see from Figure 4.2 the existence of a 7/3-power-free square $XYXY$. By the assumed minimality of xx, $XYXY$ is a conjugate of a word in $\mathcal{A}$. If $XYXY = \mu^k(w)$, where w is a conjugate of a word in A (not $\mathcal{A}$!), then $XYXY$ can be written either as A_1A_1 or as $A_1A_2A_3A_1A_2A_3$, where the A_i's are all Morse blocks of the same length. Since $XYXY$ is followed by X in zz, by Lemma 4.2, $XYXY$ is followed by the Morse block A_1, creating either a cube or a 7/3-power in xx, contrary to our assumption.

If $XYXY \neq \mu^k(w)$, where w is a conjugate of a word in A, then we may write $XYXY$ as one of $uBABv$, $uBAABAv$, $uBBABBv$, or $uABAABv$, where A and B are Morse blocks, $B = \overline{A}$, and $vu = A$.

If $XYXY = uBABv$, then, since u is a suffix of A and v is a prefix of A, applying Lemma 4.2, we see that BAB is preceded and followed by A. Thus xx contains the 5/2-power $ABABA$, contrary to our assumption. Similarly, if $XYXY = uBAABAv$, then $BAABA$ is preceded and followed by A, creating the 7/3-power $ABAABAA$. The other possibilities for $XYXY$ lead to the existence of a 7/3-power in xx by a similar argument.

Since in all cases we have derived a contradiction by showing that xx contains a 7/3-power, we conclude that our assumption that zz contains a 7/3-power is false. Recalling that $zz = \mu(ay)$ and that ay is necessarily a square, we conclude that ay is a 7/3-power-free square, contradicting the minimality of xx. We conclude that there exists no 7/3-power-free square xx that is not a conjugate of a word in $\mathcal{A}$. □

We thus have a precise characterization of, not only the overlap-free binary squares, but the 7/3-power-free binary squares as well. Furthermore, the constant 7/3 in Theorem 4.8 is best possible. To see this, note that the word

$$011010011011001011001101001101100101 10$$

is a $(7/3)^+$-power-free square, but is not a conjugate of a word in $\mathcal{A}$.

This concludes our examination of the overlap-free and 7/3-power-free binary squares. In the next section we begin to examine some of the differences between the 7/3-power-free words and the overlap-free words.

4.5 Words containing infinitely many overlaps

In this section we construct various infinite α-power-free binary words containing infinitely many overlaps. We begin by considering the infinite 7/3-power-free binary words.

Proposition 4.9. *For all $p \geq 1$, an infinite 7/3-power-free word contains only finitely many occurrences of overlaps with period p.*

Proof. Let $\mathbf{x}$ be an infinite 7/3-power-free word containing infinitely many overlaps with period p. Let $k \geq 0$ be the smallest integer satisfying $p \leq 3 \cdot 2^k$. Suppose $\mathbf{x}$ contains an overlap w with period p starting in a position $\geq 2^{k+1}$. Then by Theorem 1.11, we can write

$$\mathbf{x} = u_1 \mu(u_2) \cdots \mu^{k-1}(u_k) \mu^k(\mathbf{y}),$$

where each $u_i \in \{\epsilon, 0, 1, 00, 11\}$. The overlap w occurs as a subword of $\mu^k(\mathbf{y})$. By Lemma 4.1, $\mathbf{y}$ contains an overlap with period $p/2^k \leq 3$. But any overlap with period ≤ 3 contains a 7/3-power. Thus, $\mathbf{x}$ contains a 7/3-power, a contradiction. □

The following theorem provides a striking contrast to Shur's result [227] that the bi-infinite 7/3-power-free words are overlap-free.

Theorem 4.10. *There exists a 7/3-power-free binary word containing infinitely many overlaps.*

Proof. We define the following sequence of words: $A_0 = 00$ and $A_{n+1} = 0\mu^2(A_n)$, $n \geq 0$. The first few terms in this sequence are

$$\begin{array}{rcl} A_0 & = & 00 \\ A_1 & = & 001100110 \\ A_2 & = & 0011001101001100101100110100110010110 \\ & \vdots & \end{array}$$

We first show that in the limit as $n \to \infty$, this sequence converges to an infinite word $\mathbf{a}$. It suffices to show that for all n, A_n is a prefix of A_{n+1}. We proceed by induction on n. Certainly, $A_0 = 00$ is a prefix of $A_1 = 0\mu^2(00) = 001100110$. Now $A_n = 0\mu^2(A_{n-1})$, $A_{n+1} = 0\mu^2(A_n)$, and by induction, A_{n-1} is a prefix of A_n. Applying Lemma 2.9, we see that A_n is a prefix of A_{n+1}, as required.

Note that for all n, A_{n+1} contains $\mu^{2n}(A_1)$ as a subword. Since A_1 is an overlap with period 4, $\mu^{2n}(A_1)$ contains 2^{2n} overlaps with period 2^{2n+2}. Thus, $\mathbf{a}$ contains infinitely many overlaps.

We must show that $\mathbf{a}$ does not contain a 7/3-power. It suffices to show that A_n does not contain a 7/3-power for all $n \geq 0$. Again, we proceed by induction on n. Clearly, $A_0 = 00$ does not contain a 7/3-power. Consider $A_{n+1} = 0\mu^2(A_n)$. By induction, A_n is 7/3-power-free, and by Theorem 1.6, so is $\mu^2(A_n)$. Thus, if A_{n+1} contains a 7/3-power, such a 7/3-power must occur as a prefix of A_{n+1}. Note that A_{n+1} begins with 00110011. The word 00110011 cannot occur anywhere else in A_{n+1}, as that would imply that A_{n+1} contained a cube 000 or 111, or the 5/2-power 1001100110. If A_{n+1} were to begin with a 7/3-power with period ≥ 8, it would contain two occurrences of 00110011, contradicting our earlier observation. We conclude that the period of any such 7/3-power is less than 8. Checking that no such 7/3-power exists is now a finite check and is left to the reader. □

Alternatively, we may define $\mathbf{a} = g(h^{\omega}(0))$, where h and g are the morphisms defined by

$$\begin{array}{rcl} h(0) & = & 0134 \\ h(1) & = & 2134 \\ h(2) & = & 3234 \\ h(3) & = & 2321 \\ h(4) & = & 3421 \end{array} \quad \text{and} \quad \begin{array}{rcl} g(0) & = & 0 \\ g(1) & = & 0 \\ g(2) & = & 0 \\ g(3) & = & 1 \\ g(4) & = & 1. \end{array}$$

The sequence $\mathbf{a}$ is thus an *automatic sequence* (in the sense of Allouche and Shallit [23]).

It is possible to prove the following stronger statement.

Theorem 4.11. *There exist uncountably many $7/3$-power-free binary words containing infinitely many overlaps.*

For the proof see Currie, Rampersad, and Shallit [72]. The result of Theorem 4.10 can be strengthened even further.

Theorem 4.12. *For every real number $\alpha > 2$ there exists a real number β arbitrarily close to α, such that there is an infinite β^+-power-free binary word containing infinitely many β-powers.*

Proof. Let $s \geq 3$ be a positive integer, and let $r = \lfloor \alpha + 1 \rfloor$. Let t be the largest positive integer such that $r - t/2^s > \alpha$, and such that the word obtained by removing the prefix of length t from $\mu^s(0)$ begins with 00. Let $\beta = r - t/2^s$. Since $\alpha \geq r - 1$, we have $t < 2^s$. Also, $\mu^3(0) = 01101001$ and $\mu^3(1) = 10010110$ are of length 8, and both contain 00 as a subword; it follows that $|\alpha - \beta| \leq 8/2^s$, so that by choosing s large enough, β can be made arbitrarily close to α.

We construct sequences of words A_n, B_n and C_n. Define $C_0 = 00$. For each $n \geq 0$:

1. Let $A_n = 0^{r-2}C_n$.
2. Let $B_n = \mu^s(A_n)$.
3. Remove the first t letters from B_n to obtain a new word C_{n+1} beginning with 00.

Since each A_n begins with the r-power 0^r, each $B_n = \mu^s(A_n)$ begins with an r-power of period 2^s. Removing the first t letters ensures that C_{n+1} commences with an $(r2^s - t)/2^s$-power, i.e., a β-power. The limit of the C_n's gives the desired infinite word. We first prove that this limit exists.

Let w be the word consisting of the first t letters of $\mu^s(0)$. Since all the A_n's begin with 0, all the B_n's begin with $\mu^s(0)$, and hence with w. Thus $B_n = wC_{n+1}$ for all $n \geq 0$.

We show by induction on n that for all $n \geq 0$, A_n is a prefix of A_{n+1}. Certainly A_0 is a prefix of A_1. Assume that A_{n-1} is a prefix of A_n. Since $A_n = 0^{r-2}C_n$ and $A_{n+1} = 0^{r-2}C_{n+1}$, A_n is a prefix of A_{n+1} if C_n is a prefix of C_{n+1}. Since $B_{n-1} = wC_n$ and $B_n = wC_{n+1}$, C_n is a prefix of C_{n+1} if B_{n-1} is a prefix of B_n. By Lemma 2.9, B_{n-1} is a prefix of B_n if A_{n-1} is a prefix of A_n, which is our inductive assumption. We conclude that A_n is a prefix of A_{n+1}.

It follows that C_n is a prefix of C_{n+1} for $n \geq 0$, so that the limit of the C_n's exists. To prove our theorem it thus suffices to prove the following claim.

Claim: For all $n \geq 0$, A_n, B_n and C_n satisfy the following:

1. The word C_n contains no β^+-powers.
2. The only β^+-power in A_n is 0^r.
3. Any β^+-powers in B_n appear only in the prefix $\mu^s(0^r)$.

Note that part 3 implies part 1, so it suffices to prove parts 2 and 3. We begin by proving part 3.

Certainly C_0 contains no β^+-powers, and since $\beta > r-1$, the only β^+-power in A_0 is 0^r. Suppose that the claim holds for A_n and C_n. We show that part 3 holds for B_n.

Suppose that $B_n = \mu^s(0^{r-2})\mu^s(C_n)$ contains a β^+-power u with period p. Since C_n contains no β^+-powers, Theorem 1.6 ensures that $\mu^s(C_n)$ contains no β^+-powers. We can therefore write $B_n = xuy$ where $|x| < |\mu^s(0^{r-2})|$. In other words, u overlaps $\mu^s(0^{r-2})$ from the right. By Theorem 4.1, the pre-image of B_n under μ, i.e., $\mu^{s-1}(A_n)$, contains a β^+-power of length at least $|u|/2$ and period $p/2$. Iterating this argument, A_n contains a β^+-power of length at least $|u|/2^s$ and period $p/2^s$. The only β^+-power in A_n is 0^r, with period 1, so $p/2^s = 1$. It follows that $p = 2^s$ and $|u| \leq r2^s$.

Recall that B_n has a prefix $\mu^s(0^r)$, which also has period 2^s, and that this prefix is overlapped by u. It follows that xu is a β^+-power with period $p = 2^s$. This implies that $|xu| \leq r2^s = |\mu^s(0^r)|$, so that u is contained in $\mu^s(0^r)$. Thus part 3 of our claim holds for B_n.

We now show that part 2 holds for A_{n+1}, which completes the proof of the claim and the theorem. Suppose that A_{n+1} contains a β^+-power u. Recall that $A_{n+1} = 0^{r-2}C_{n+1}$, and C_{n+1} begins with 00 but contains no β^+-powers. It follows that u is not a subword of C_{n+1}. Therefore, 000 must be a prefix of u. If $u = 0^q$ for some integer q, then $q \leq r$ by the construction of A_{n+1}, and

$$r \geq q > \beta > \alpha > r-1.$$

This implies that $q = r$ and $u = 0^r$, as claimed. If we cannot write $u = 0^q$, then, since u is a 2^+-power, 000 must appear twice in u with a 1 occurring somewhere between the two appearances. This implies that 000 is a subword of C_{n+1} and hence of $B_n = \mu^s(A_n)$. However, no word of the form $\mu(w)$ contains 000. This is a contradiction. We conclude that part 2 holds for A_{n+1}. This completes the proof. □

This concludes our study of the properties of overlap-free and 7/3-power-free words. In the next chapter, we shall still be concerned with overlaps and overlap-freeness, but we shall instead focus our attention on properties of *languages*, i.e., sets of words.

Chapter 5

Applications to the Theory of Context-free Languages

In this chapter, we examine some results in formal language theory that were inspired by concepts from combinatorics on words[1]. The main results of this chapter are that over a three-letter alphabet, the set of words containing overlaps is not context-free, and over a two-letter alphabet, the set of words containing overlaps is not unambiguously context-free. We also show that the set of words that do not occur as subwords of the Thue–Morse word is not unambiguously context-free.

5.1 Definition of the context-free languages

Most of the results in this chapter concern the context-freeness of certain sets of words. In this section we therefore review the standard definition of a context-free language. For further information on context-free languages or formal language theory in general the reader may consult Hopcroft and Ullman [124].

A *context-free grammar* G is a 4-tuple $G = (V, T, P, S)$, where

- V is a finite set of *variables*;
- T is a finite set of *terminals* (V and T are disjoint);
- P is a finite set of *productions*, where a production is of the form $A \to \alpha$, $A \in V$ and $\alpha \in (V \cup T)^*$; and,
- $S \in V$ is the *start variable.*

For any $\alpha, \gamma \in (V \cup T)^*$, if $A \to \beta$ is a production in P, we write $\alpha A \gamma \Longrightarrow \alpha\beta\gamma$, and we denote by $\overset{*}{\Longrightarrow}$ the reflexive and transitive closure of $\Longrightarrow$. The language $L(G)$

[1] The results in this chapter can be found in Rampersad [203].

generated by G is the set of all $w \in T^*$ such that $S \overset{*}{\Longrightarrow} w$. If a language $L = L(G)$ for some G, we say that L is a *context-free language*. Recall that a language is context-free if and only if it is accepted by a pushdown automaton (see Hopcroft and Ullman [124]).

A derivation

$$S \Longrightarrow \alpha_1 \Longrightarrow \alpha_2 \Longrightarrow \cdots \Longrightarrow \alpha_n \Longrightarrow w$$

of a word w is a *leftmost derivation* if for every $i = 1, 2, \ldots, n-1$ we obtain α_{i+1} from α_i by applying a production to the leftmost variable of α_i. A grammar G is *ambiguous* if there exists a word w that has at least two distinct leftmost derivations in G. A context-free language L is *inherently ambiguous* if every context-free grammar G for which $L = L(G)$ is ambiguous; otherwise, L is *unambiguously context-free.*

5.2 Historical background

Autebert, Beauquier, Boasson, and Nivat [27] presented a series of open problems and conjectures regarding context-free languages. One conjecture was that the language of all words (over a sufficiently large alphabet) containing a square was not context-free.

Ross and Winklmann [209] used the following theorem, usually refered to as the *interchange lemma*, to prove this conjecture (see Ogden, Ross, and Winklmann [190]).

Theorem 5.1 (Ogden, Ross, and Winklmann)**.** *Let $L \subseteq \Sigma^*$ be a context-free language. There exists a constant c, depending only on L, such that for all $n \geq 2$, all subsets $R \subseteq L \cap \Sigma^n$, and all m, $2 \leq m \leq n$, there exists a subset $Z \subseteq R$, $Z = \{z_1, z_2, \ldots, z_k\}$, such that*

(a) $k \geq \frac{|R|}{c(n+1)^2}$;

(b) $z_i = w_i x_i y_i$, $1 \leq i \leq k$;

(c) $|w_1| = |w_2| = \cdots = |w_k|$;

(d) $|y_1| = |y_2| = \cdots = |y_k|$;

(e) $m/2 \leq |x_1| = |x_2| = \cdots = |x_k| \leq m$;

(f) $w_i x_j y_i \in L$, $1 \leq i, j \leq k$.

Theorem 5.2 (Ross and Winklmann; Ehrenfeucht and Rozenberg)**.** *The language consisting of all ternary words containing a square is not context-free.*

This result was also independently obtained by Ehrenfeucht and Rozenberg [84] using different methods. Note that in order for this result to be non-trivial, there must exist infinitely many squarefree ternary words. By Thue's construction of an infinite squarefree ternary word, we know this to be the case.

Berstel [36] asked if a similar result held for the language of words containing overlaps. Gabarró [104] used the interchange lemma to show that over a 4-letter alphabet, the language of all words containing an overlap is not context-free. We shall show below how to improve this result to a 3-letter alphabet.

Gabarró also proved that over a 4-letter alphabet, the language of all words containing a cube is not context-free. Applying Theorem 1.5, together with the well-known fact that the class of context-free languages is closed under inverse homomorphism, one can improve Gabarró's result to the following.

Theorem 5.3. *The language consisting of all binary words containing a cube is not context-free.*

Main [167] adapted the argument of Ross and Winklmann to show that over a 16-letter alphabet, the languages of all words containing an abelian square is not context-free. Gabarró improved this to a 7-letter alphabet. This result can be further improved to a 6-letter alphabet by a slight modification of Gabarró's argument. The cases of a 4 or 5-letter alphabet remain open.

Theorem 5.4. *The language consisting of all ternary words containing an overlap is not context-free.*

Proof. Let $n = 2^{2k+1} + 1$ for some $k \geq 0$. Let $x = \mu^{2k}(0)$ and let $w = 0xx$. Then w is an overlap, but no proper subword of w is an overlap. To see this, note that xx is a subword of the Thue-Morse word and is therefore overlap-free. Any overlap contained in w must therefore begin from the first position of w. If w begins with two distinct overlaps, then xx begins with two distinct squares, contradicting the result of Proposition 2.10.

Suppose that the set of words over $\{0, 1, 2\}$ that contain an overlap is context-free. Let ψ be the morphism defined by $\psi(0) = 0$ and $\psi(1) = \psi(2) = 1$. Define

$$R = \{0yy : y \in \psi^{-1}(x)\}.$$

Note that $|R| = 2^{(n-1)/4}$. Applying the interchange lemma, we see that there exists $Z \subseteq R$ with

$$|Z| \geq \frac{2^{(n-1)/4}}{c(n+1)^2}. \tag{5.1}$$

Choosing $m = (n - 1)/2$, and recalling that if $z_i = w_i x_i y_i \in Z$, then $m/2 \leq |x_i| \leq m$, we see that $w_i x_j y_i \in L$ only if $x_i = x_j$. Fixing x_i, we easily verify that there are at most $2^{(n-1)/8}$ words $w_j x_j y_j$ with $x_i = x_j$, so that $|Z| \leq 2^{(n-1)/8}$, contradicting (5.1) for n sufficiently large. This concludes the proof. □

5.3 Languages derived from infinite words

Let $\mathbf{w}$ be an infinite word. Perhaps the most natural way to derive a language from $\mathbf{w}$ is to consider the language consisting of all subwords of $\mathbf{w}$. Let $\mathrm{Sub}(\mathbf{w})$ denote this language, and let $\mathrm{Pref}(\mathbf{w})$ denote the language consisting of all prefixes of $\mathbf{w}$. Let $\mathrm{Cosub}(\mathbf{w})$ and $\mathrm{Copref}(\mathbf{w})$ denote the complements of $\mathrm{Sub}(\mathbf{w})$ and $\mathrm{Pref}(\mathbf{w})$ respectively. It is clear that if $\mathbf{w}$ is ultimately periodic, then $\mathrm{Sub}(\mathbf{w})$ and $\mathrm{Pref}(\mathbf{w})$ are both regular languages.

For any infinite word $\mathbf{w}$ it follows from the pumping lemma that $\mathrm{Pref}(\mathbf{w})$ is context-free if and only if $\mathbf{w}$ is ultimately periodic. Berstel [37] proved the following theorem regarding $\mathrm{Copref}(\mathbf{w})$.

Theorem 5.5 (Berstel)**.** *Let* $\mathbf{w}$ *be the infinite fixed point of a morphism. Then Copref(*$\mathbf{w}$*) is a context-free language.*

Autebert, Flajolet, and Gabarró [28] used the Chomsky–Schützenberger Theorem (see Section 5.6 below) along with the Pólya–Carlson Theorem (see Section 5.7 below) to show that when $\mathbf{w}$ is an aperiodic fixed point of a morphism, $\mathrm{Copref}(\mathbf{w})$ is an inherently ambiguous context-free language (for an alternative proof, see Honkala [122]).

Grazon [109] showed that if

$$\mathbf{w} = a^{1!}ba^{2!}ba^{3!}b\cdots,$$

then $\mathrm{Copref}(\mathbf{w})$ is not context-free. The proof is surprisingly difficult.

Autebert, Beauquier, Boasson, and Nivat [27] asked if it were possible to have a context-free language whose complement is infinite and consists only of squarefree words. Main [168] answered this question in the affirmative. One can deduce Main's result from Theorem 5.5 in the following way. Let $\mathbf{w}$ be any infinite squarefree word generated by a morphism. Then $\mathrm{Copref}(\mathbf{w})$ has the desired property by Theorem 5.5. Main, Bucher, and Haussler [169] disproved several conjectures of Autebert, Beauquier, Boasson, and Nivat by similar arguments.

Marcus and Păun [170] presented some open problems regarding languages derived from infinite words. In particular, they asked the following questions (see Section 1.4 for the definition of *uniformly recurrent*):

> Is there a uniformly recurrent infinite word such that the set of its subwords is context-free but not regular?

> Is there an infinite word such that the set of its subwords that appear infinitely often is context-free but not regular?

The first question was answered negatively by Istrate [131] (see also Thomsen [232]). The second question was answered affirmatively by Ilie [127].

By an easy application of the pumping lemma, one deduces that if $\mathbf{w}$ is k-power-free for some integer $k \geq 2$, then Sub($\mathbf{w}$) is not context-free. By completely characterizing the binary morphisms that generate non-repetitive words, Kobayashi, Otto, and Séébold [143] (see also Kobayashi and Otto [142]) proved the following result regarding binary words.

Theorem 5.6 (Kobayashi, Otto, and Séébold)**.** *Let* $\mathbf{w}$ *be the infinite fixed point of a binary morphism. If* $\mathbf{w}$ *is aperiodic, then Sub*($\mathbf{w}$) *is not context-free.*

Frid [103] generalized this result in a very strong way. We shall discuss Frid's result in Section 5.4, after we have introduced the family of *bounded* context-free languages.

Recall from Section 1.6 that a non-empty word is *primitive* if it is not a k-power for any $k \geq 2$. Petersen [196] (see also Allouche [15]) proved that the set of primitive binary words is not unambiguously context-free. Proving that this language is not context-free has remained a long standing open problem in the area of formal languages.

5.4 Sparse context-free languages

Let $L \subseteq \Sigma^*$ be a language. If there exist words $w_1, w_2, \ldots, w_k \in \Sigma^*$ such that $L \subseteq w_1^* w_2^* \cdots w_k^*$, then L is called a *bounded language*. Ginsburg [105] presented many results regarding the structure of bounded context-free languages.

If there exists a polynomial $p(n)$ such that $|L \cap \Sigma^n| \leq p(n)$, then L is said to be *sparse* or *poly-slender*. Trofimov [237] gave the following characterization of sparse context-free languages. (This result was independently rediscovered by Latteux and Thierrin [154], Ibarra and Ravikumar [126], Raz [205], Incitti [129], and Bridson and Gilman [49].)

Theorem 5.7 (Trofimov)**.** *A context-free language is sparse if and only if it is bounded.*

Frid [103] proved the following result.

Theorem 5.8 (Frid)**.** *Let* $\mathbf{w}$ *be an infinite word with polynomially bounded subword complexity. If* $\mathbf{w}$ *is aperiodic, then Sub*($\mathbf{w}$) *is not context-free.*

Note that this result implies that of Theorem 5.6. Moreover, Frid's proof is much simpler that that of Kobayashi, Otto, and Séébold. Recalling the decidability result of Theorem 1.4, we have the following corollary.

Corollary 5.9. *Let* $\mathbf{w}$ *be the fixed point of a morphism. It is decidable whether Sub*($\mathbf{w}$) *is context-free.*

5.5 The language Cosub(w)

We begin with some simple results concerning the language Cosub($\mathbf{w}$). In this section we use the following notation from formal language theory. If $L \subseteq \Sigma^*$, then $\overline{L}$ denotes $\Sigma^* \setminus L$. We also define $L^* = \{w^k : w \in L, k \geq 0\}$. If $L = \{w\}$, then we write w^* for $\{w\}^*$. If $L_1, L_2 \subseteq \Sigma^*$, then $L_1 + L_2$ denotes $L_1 \cup L_2$.

Proposition 5.10. *There exists a word* $\mathbf{x}$ *generated by a morphism for which Cosub*($\mathbf{x}$) *is not context-free.*

Proof. Let h be the morphism that maps $0 \to 012$, $1 \to 11$, $2 \to 2$, and let $\mathbf{x} = h^\omega(0) = 0121121^4 21^8 2\cdots$. Then $\text{Cosub}(\mathbf{x}) \cap 21^*2 = \{21^n 2 : n \neq 2^k, k \geq 1\}$, which is clearly not context-free. □

Proposition 5.11. *There exists an aperiodic word* $\mathbf{y}$ *generated by a morphism for which Cosub*($\mathbf{y}$) *is context-free.*

Proof. Let h be the morphism that maps $0 \to 012$, $1 \to 1$, $2 \to 12$, and let $\mathbf{y} = h^\omega(0) = 0121121^3 21^4 2\cdots$. Define three sets A_1, A_2, and A_3 as follows. Let

$$A_1 = \overline{(1^*2)^*1^* + 0(1^*2)^*1^*}.$$

Let A_2 consist of all words of the form

$$1^{i_1}21^{i_2}2\cdots 21^{i_{k-1}}21^{i_k},$$

where either (a) $i_1 \geq i_2$; or, (b) $i_k > i_{k-1} + 1$; or, (c) there exists j, $2 \leq j \leq k-2$, such that $i_{j+1} \neq i_j + 1$. Similarly, let A_3 consist of all words of the form

$$01^{i_1}21^{i_2}2\cdots 21^{i_{k-1}}21^{i_k},$$

where either $i_k > i_{k-1} + 1$ or there exists j, $1 \leq j \leq k-2$, such that $i_{j+1} \neq i_j + 1$. Then $\text{Cosub}(\mathbf{y}) = A_1 \cup A_2 \cup A_3$. The set A_1 is regular and hence context-free. The set A_2 is easily seen to be context-free: to accept A_2 a pushdown automaton first non-deterministically guesses which of the conditions (a), (b), or (c) holds. The automaton can then use its stack to verify that (a) or (b) holds. To check that (c) holds, the pushdown automaton guesses an index j and uses its stack to verify that $i_{j+1} \neq i_j + 1$. One shows that A_3 is context-free in the same manner. Thus, Cosub($\mathbf{y}$) is context-free. □

The reader may have noted a certain similarity between the language Cosub($\mathbf{y}$) of Proposition 5.11 and the classical "Goldstine language"

$$\{a^{i_1}ba^{i_1}b\cdots a^{i_k}b : \text{ there exists } j,\ 1 \leq j \leq k, \text{ such that } i_j \neq j\},$$

a language well-known to be an inherently ambiguous context-free language (see Flajolet [97]).

Having established these simple results, we now present the following open problems.

Problem 5.12. *Let* $\mathbf{t}$ *be the Thue–Morse word. Is Cosub*$(\mathbf{t})$ *context-free?*

In Theorem 5.20 we shall prove that if Cosub$(\mathbf{t})$ is context-free, it must be inherently ambiguous.

Problem 5.13. *Let* $\mathbf{f}$ *be the Fibonacci word: i.e., the word*

$$0100101001001010010100\cdots$$

generated by iterating the morphism $0 \to 01$, $1 \to 0$. *Is Cosub*$(\mathbf{f})$ *context-free?*

Recall from Section 2.5 that we have a characterization of the set $M_{\mathbf{t}}$ of minimal forbidden subwords of $\mathbf{t}$. We thus have Cosub$(\mathbf{t}) = \{0,1\}^* M_{\mathbf{t}} \{0,1\}^*$, where $M_{\mathbf{t}}$ is a well-understood language. Mignosi, Restivo, and Sciortino [175] gave a characterization of the set $M_{\mathbf{f}}$ of minimal forbidden subwords of the Fibonacci word:

$$\begin{aligned} M_{\mathbf{f}} &= \{w : w = 0v0 \text{ and } v \text{ is the } 2k\text{-th palindromic prefix of } \mathbf{f}\} \cup \\ &\quad \{w : w = 1v1 \text{ and } v \text{ is the } (2k+1)\text{-th palindromic prefix of } \mathbf{f}\}, \end{aligned}$$

so that Cosub$(\mathbf{f}) = \{0,1\}^* M_{\mathbf{f}} \{0,1\}^*$.

5.6 The Chomsky–Schützenberger Theorem

We now review a deep result of Chomsky and Schützenberger [63] connecting unambiguous context-free languages to algebraic power series. We shall use this result to prove that certain languages cannot be unambiguously context-free.

If $L \subseteq \Sigma^*$ is a language, then the *generating series* of L is the formal power series

$$F(X) = \sum_{n \geq 0} |L \cap \Sigma^n| \cdot X^n;$$

i.e., the coefficient of X^n is the number of words in L of length n. The Chomsky–Schützenberger Theorem is the following.

Theorem 5.14 (Chomsky and Schützenberger). *If L is an unambiguous context-free language, then the generating series of L is algebraic over $\mathbb{Q}(X)$.*

For a proof of this theorem, see Kuich and Salomaa [152, Chap. 16], Panholzer [191], or Haiman [113].

Flajolet [97] and Allouche [15], among others, have given numerous examples of context-free languages that can be proved to be inherently ambiguous by means of the Chomsky–Schützenberger Theorem. Note that by viewing the generating series as not merely a formal power series, but as a complex power series with some well-defined radius of convergence, the techniques of classical complex analysis can be applied to prove transcendence over $\mathbb{Q}(X)$. Such techniques are the particular focus of Flajolet's paper.

D'Alessandro, Intrigila, and Varricchio [12] proved the following result regarding the generating series of sparse context-free languages.

Theorem 5.15 (D'Alessandro, Intrigila, and Varricchio)**.** *The generating series of a sparse context-free language is a rational function.*

5.7 Techniques from analysis

In this section we review some concepts and results from complex analysis that we shall need later in the chapter.

Consider a power series in the complex variable z

$$\sum_{n=0}^{\infty} a_n z^n,$$

with complex coefficents a_n. Then there exists a number R, $0 \leq R \leq \infty$, called the *radius of convergence*, such that the series converges for $|z| < R$ and diverges for $|z| > R$. If $R < \infty$, we call the set $\{z : |z| = R\}$ the *circle of convergence*. Additionally, if $R < \infty$, it follows that there exists some z_0, $|z_0| = R$, such that the series diverges for $z = z_0$. Such points z_0 are called *singularities*.

Recall that a power series defines a complex analytic function within its circle of convergence. An analytic function $f(z)$ is *algebraic* if $y = f(z)$ is a solution to some polynomial equation

$$p_0(z)y^n + p_1(z)y^{n-1} + \cdots + p_n(z) = 0,$$

where $p_i \in \mathbb{Q}[z]$, $0 \leq i \leq n$. An algebraic function $f(z)$ is *rational* if it can written as quotient of two polynomials; i.e, $f(z) = p(z)/q(z)$ for some polynomials $p, q \in \mathbb{Q}[z]$. An analytic function is *transcendental* if it is not algebraic.

The following is a well-known fact (see, for example, Hille [121, Theorem 12.2.1]).

An algebraic function has only a finite number of singularities.

Consequently, any analytic function with infinitely many singularities is transcendental. If there exists a region D of the complex plane for which every point on the boundary of D is a singularity of $f(z)$, then that boundary is said to be a *natural boundary* of $f(z)$. It follows that $f(z)$ does not have an analytic continuation to any region of the complex plane that properly contains D. For our purposes, it suffices to observe that any function with a natural boundary is transcendental.

A standard example given in most textbooks of a power series with a natural boundary is the series

$$\sum_{n=0}^{\infty} z^{2^n}.$$

Theorem 5.17 below gives a general result regarding series of this form.

The two analytic results most readily applicable to the languages we shall study are the Pólya–Carlson Theorem [54] and Fabry's gap theorem (see Hille [121, Theorem 11.7.2]).

Theorem 5.16 (Pólya and Carlson)**.** *A complex power series $\sum_{n\geq 0} a_n z^n$ with integer coefficients and radius of convergence 1 is either rational or admits the unit circle as a natural boundary (and consequently is transcendental over $\mathbb{Q}(z)$).*

(For a weaker but more accessible result, see Fatou [94] or Pólya and Szegő [198, Part VIII, Chap. 3, No. 167].)

Theorem 5.17 (Fabry)**.** *A complex power series $\sum_{n\geq 0} a_n z^{b_n}$, where $a_n \neq 0$ and*

$$\lim_{n\to\infty} \frac{n}{b_n} = 0,$$

admits its circle of convergence as a natural boundary (and consequently is transcendental over $\mathbb{Q}(z)$).

(Again, for a weaker but more accessible gap theorem due to Hadamard, see Hille [121, Theorem 11.7.1] or Rudin [214, Theorem 16.6].)

Power series satisfying the conditions of Fabry's gap theorem are called *gap series* or *lacunary series*.

5.8 Binary words containing overlaps

In this section we use the Chomsky–Schützenberger Theorem, the Pólya–Carlson Theorem, and the estimates (1.2) of Lepistö [158] to prove the following theorem.

Theorem 5.18. *The language consisting of all binary words containing an overlap is not unambiguously context-free.*

Proof. Let

$$F(X) = \sum_{n\geq 0} a_n X^n$$

be the generating series of the overlap-free words. That is, a_n is the number of overlap-free words of length n over a binary alphabet. By the Chomsky–Schützenberger Theorem, if the set of binary words that contain an overlap is unambiguously context-free, then $F(X)$ is algebraic over $\mathbb{Q}(X)$. To prove the theorem it suffices then to show that $F(X)$ is transcendental.

From (1.2) we have $a_n = O(n^{1.369})$, so $F(X)$, as a complex power series, has radius of convergence 1. By the Pólya-Carlson Theorem $F(X)$ is either rational or transcendental over $\mathbb{Q}(X)$. To complete the proof we must show that $F(X)$ is not rational. If $F(X)$ were rational, then the coefficients a_n could be written in the form

$$a_n = \sum_{i=1}^{m} A_i(n)\alpha_i^n,$$

for some m, where α_i is a characteristic root of multiplicity n_i of the linear recurrence satisfied by $(a_n)_{n\geq 0}$, and $A_i(X)$ is a polynomial of degree at most $n_i - 1$ (see Everest, van der Poorten, Shparlinski, and Ward [93, Section 1.1.6]). But from (1.2) we see that this is not possible, so $F(X)$ is not rational and the proof is complete. □

We now consider the language of all binary words w such that some *conjugate* of w contains an overlap as a subword. Recall the set $\mathcal{C}$ defined in Section 1.12. Harju [117] proved (Theorem 1.7) that the set of circular overlap-free binary words is the set of conjugates of $\mathcal{C}$. Let us denote the set of these conjugates by $\widetilde{\mathcal{C}}$. We are thus interested in the complement of $\widetilde{\mathcal{C}}$.

Theorem 5.19. *The complement of* $\widetilde{\mathcal{C}}$ *is not unambiguously context-free.*

Proof. Harju's characterization of the circular overlap-free binary words implies that these words have lengths of the form 2^n or $3 \cdot 2^n$, $n \geq 0$. The generating series of $\widetilde{\mathcal{C}}$ is a thus a gap series. By Fabry's gap theorem it admits its circle of convergence as a natural boundary and hence is transcendental. Applying the Chomsky–Schützenberger Theorem, we conclude that the complement of $\widetilde{\mathcal{C}}$ is not unambiguously context-free. □

Next we consider the set of words that do not occur as subwords of the Thue–Morse word. Recall the formula (2.1) for the subword complexity $p_{\mathbf{t}}(n)$ of the Thue–Morse word. We shall use this formula to derive the following result.

Theorem 5.20. *Let* $\mathbf{t}$ *be the Thue–Morse word. The language Cosub*$(\mathbf{t})$ *is not unambiguously context-free.*

Proof. Let

$$F(X) = \sum_{n\geq 1} p_{\mathbf{t}}(n) X^n$$

be the generating series of the subwords of the Thue–Morse word. We show that $F(X)$ is transcendental over $\mathbb{Q}(X)$. Suppose to the contrary that $F(X)$ is algebraic. Then the series

$$G(X) = \sum_{n\geq 1} (p_{\mathbf{t}}(n+1) - p_{\mathbf{t}}(n)) X^n,$$

whose coefficients form the sequence of *first differences* of $p_{\mathbf{t}}(n)$, is also algebraic. Note that for all $n \geq 1$,

$$p_{\mathbf{t}}(n+1) - p_{\mathbf{t}}(n) \leq 4,$$

so that the coefficients of $G(X)$ are bounded. Applying the Pólya-Carlson Theorem to $G(X)$, we see that $G(X)$ is either rational or transcendental. By assumption, $G(X)$ is algebraic, so it must be rational. But then the sequence

$$\Delta = (p_{\mathbf{t}}(n+1) - p_{\mathbf{t}}(n))_{n\geq 1}$$

is ultimately periodic. We easily verify from (2.1) that this is not the case: for instance, Δ contains arbitrarily large "runs" of 4's. This contradiction implies the transcendence of $F(X)$.

Alternatively, one may note that the series $H(X)$, whose coefficients form the sequence of *second differences* of $p_{\mathbf{t}}(n)$, is a gap series, and one may therefore apply Fabry's gap theorem to $H(X)$.

The desired result follows by applying the Chomsky–Schützenberger Theorem. □

The use of analytic techniques in the proof of Theorem 5.20 may be avoided by applying instead the theorems of Christol [64, 65] and Cobham [66]. See the paper of Allouche [15] for some examples of this approach.

Next we consider generalizations of the Thue–Morse word. Recall the formula (2.2) for the subword complexity $p_{\mathbf{t}_{2,m}}(n)$ of the generalized Thue–Morse words $\mathbf{t}_{2,m}$. One therefore proves the following result in a manner entirely analogous to that of Theorem 5.20.

Theorem 5.21. *For $m \geq 2$, the language Cosub($\mathbf{t}_{2,m}$) is not unambiguously context-free.*

To complete the work discussed in this section, it remains to determine whether or not the languages considered in Theorems 5.18, 5.19, 5.20, and 5.21 are context-free.

Mossé [181] and Frid [99, 100, 101] have written several papers showing that a large class of words generated by iterating morphisms have subword complexity functions that behave similarly to that of the Thue–Morse word; i.e., they are piecewise linear on exponentially growing intervals. The first difference sequence of such subword complexity functions is therefore either constant or aperiodic. If it were possible to characterize those words for which the latter situation occurs, one might generalize the argument of Theorem 5.20 to a larger class of words.

We have thus described several problems and results concerning languages that arise naturally from concepts or constructions from the theory of combinatorics on words. In the next chapter we return to our study of infinite words and seek to construct words avoiding repetitions in a much stronger sense than we have considered so far.

Chapter 6

Words Avoiding Repetitions in Arithmetic Progressions

6.1 Introduction

In this chapter we give a method for constructing infinite words containing no squares in any subsequence indexed by an arithmetic progression of odd difference[1]. Similarly, we show how to construct infinite words avoiding overlaps in all arithmetic progressions of odd difference. These constructions provide examples of words avoiding repetitions in a much stronger sense than we have considered previously.

Recall from Section 1.10 that Entringer, Jackson, and Schatz [87] constructed an infinite binary word containing no squares xx, where $|x| \geq 3$. In related work, Prodinger and Urbanek [199] gave an example of an infinite binary word whose only squares xx satisfy $|x| \in \{1, 3, 5\}$. The particular word studied by Prodinger and Urbanek is the *paperfolding word*

$$0010011000110110 \cdots .$$

In later sections we shall give additional properties of this word and the reasons for its nomenclature.

The paperfolding word seems to have first been discovered by J. E. Heighway in the 60's. For surveys on properties of the paperfolding words—in particular on connections to the so-called "dragon curve"—see Davis and Knuth [74] or Dekking, Mendès France, and van der Poorten [79]. We shall rely mostly on various combinatorial results of Allouche and Bousquet-Mélou [13, 14, 16, 17].

6.2 Van der Waerden and Carpi

Taking Thue's problem in another direction, Carpi [55], as a preliminary step in constructing non-repetitive labelings of the integer lattice, considered the question

[1] The results in this chapter can be found in Kao, Rampersad, Shallit, and Silva [135].

of the existence of infinite words that avoid squares in all subsequences indexed by arithmetic progressions. Of course, by the classical theorem of van der Waerden [239], no such words exist. We state van der Waerden's theorem below in a form commonly encountered in combinatorics texts; we leave it to the reader to translate it into a result on words. The original paper of van der Waerden is hard to find; for a proof of the theorem the reader may consult the classic textbook of Graham, Rothschild, and Spencer [108].

Theorem 6.1 (van der Waerden)**.** *Let the natural numbers be partitioned into finitely many disjoint sets $C_1, C_2, \ldots, C_k$. Then some C_i contains arbitrarily long arithmetic progressions.*

By contrast, Carpi showed that for any prime p, there exists an infinite word over a finite alphabet that avoids squares in arithmetic progressions of all differences, except those differences that are a multiple of p. For example, taking $p = 2$, there exists an infinite word over a 4-letter alphabet that contains no squares in any arithmetic progression of odd difference. As we shall see later, Carpi's construction has a surprising connection to the paperfolding words.

6.3 Definitions and notation

In this section we define the new concepts introduced in this chapter.

A *subsequence* of $\mathbf{w}$ is word of the form

$$w_{i_0} w_{i_1} \cdots ,$$

where $0 \leq i_0 < i_1 < \cdots$. An *arithmetic subsequence of difference j* of $\mathbf{w}$ is a word of the form

$$w_i w_{i+j} w_{i+2j} \cdots ,$$

where $i \geq 0$ and $j \geq 1$. We also define finite subsequences in the obvious way.

If a word $\mathbf{w}$ has the property that no arithmetic subsequence of difference j contains a square (resp. cube, r-power, r^+-power), we say that $\mathbf{w}$ *contains no squares (resp. cubes, r-powers, r^+-powers) in arithmetic progressions of difference j.*

6.4 Paperfolding words

Consider the following procedure.

1. Take an ordinary 8.5×11 piece of paper and fold it in half.

2. Now unfold the paper and record the pattern of hills and valleys created, writing 0 for a hill and 1 for a valley.

$$0$$

3. Now fold the paper twice, unfold, and record the pattern of hills and valleys.

$$0 \quad 0 \quad 1$$

4. Now fold three times, unfold, and record the pattern.

$$0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1$$

5. Now fold infinitely (!) many times. After unfolding, one obtains the following infinite sequence.

$$0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad \cdots$$

This infinite sequence over $\{0,1\}$ is called a paperfolding word. Formally, a *paperfolding word* $\mathbf{f} = f_0 f_1 f_2 \cdots$ over the alphabet $\{0,1\}$ satisfies the following recursive definition: there exists $a \in \{0,1\}$ such that

$$\begin{array}{rcl} f_{4n} & = & a, \quad n \geq 0 \\ f_{4n+2} & = & \overline{a}, \quad n \geq 0 \\ (f_{2n+1})_{n\geq 0} & & \text{is a paperfolding word.} \end{array}$$

The *ordinary paperfolding word*

$$0010011000110110\cdots$$

is the paperfolding word uniquely characterized by $f_{2^m-1} = 0$ for all $m \geq 0$.

6.5 Perturbed symmetry

One may also define the paperfolding words by means of the *perturbed symmetry* of Mendès France [45, 171] in the following way. For $i \geq 0$, let $c_i \in \{0,1\}$ and define the sequence of words

$$\begin{array}{rcl} F_0 & = & c_0 \\ F_1 & = & F_0\, c_1\, \overline{F_0}^R \\ F_2 & = & F_1\, c_2\, \overline{F_1}^R \\ & \vdots & \end{array}$$

Then

$$\mathbf{f} = \lim_{i\to\infty} F_i$$

is a paperfolding word. For example, taking $c_i = 0$ for all $i \geq 0$, one obtains the sequence

$$\begin{aligned} F_0 &= 0 \\ F_1 &= 0\,0\,1 \\ F_2 &= 001\,0\,011 \\ &\vdots \end{aligned}$$

which converges, in the limit, to the ordinary paperfolding word.

6.6 The Toeplitz construction

There is yet another way to define the paperfolding words, namely, the so-called Toeplitz construction. This construction is due to Jacobs and Keane [132] and is so named for its similarity to a construction of Toeplitz [235]. The Toeplitz construction is as follows.

1. Start with an infinite sequence of *gaps*, denoted by ?.

$$? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad ? \quad \cdots$$

2. Fill every other gap with alternating 0's and 1's.

$$0 \quad ? \quad 1 \quad ? \quad 0 \quad ? \quad 1 \quad ? \quad 0 \quad ? \quad 1 \quad ? \quad 0 \quad ? \quad 1 \quad \cdots$$

3. Goto step 2.

$$\begin{array}{cccccccccccccccc} 0 & 0 & 1 & ? & 0 & 1 & 1 & ? & 0 & 0 & 1 & ? & 0 & 1 & 1 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & ? & 0 & 0 & 1 & 1 & 0 & 1 & 1 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & \cdots \end{array}$$

In the limit, one again obtains the ordinary paperfolding word

$$0010011000110110\cdots$$

At each step, one may choose to fill in the gaps by either

$$0101010101\cdots$$

or

$$1010101010\cdots.$$

Different choices at each step results in the construction of different paperfolding words. In general, words constructed by such a process are called *Toeplitz words*. However, we shall only be concerned here with the specific case of the paperfolding words.

6.7 Repetitions in the paperfolding words

The following properties of paperfolding words were proved by Allouche and Bousquet-Mélou [13, 16] (the particular case of the ordinary paperfolding word was studied by Prodinger and Urbanek [199]).

Theorem 6.2 (Allouche and Bousquet-Mélou)**.** *For any paperfolding word* $\mathbf{f}$*, if* xx *is a non-empty subword of* $\mathbf{f}$*, then* $|x| \in \{1, 3, 5\}$*.*

Corollary 6.3 (Allouche and Bousquet-Mélou)**.** *For any paperfolding word* $\mathbf{f}$*,* $\mathbf{f}$ *contains no fourth powers and no cubes except* 000 *and* 111*. In particular,* $\mathbf{f}$ *contains no* 3^+*-power.*

Unfortunately, the proof of Theorem 6.2 given in [16] contains an error. For completeness we therefore provide a proof below. We first prove the following corrected version of [16, Proposition 5.1].

Proposition 6.4. *If a paperfolding word* $\mathbf{f}$ *contains a subword* wcw*, where* w *is a non-empty word and* c *is a single letter, then either* $|w| \in \{2, 4\}$ *or* $|w| = 2^k - 1$ *for some* $k \geq 1$*.*

We shall need the following result due to Allouche [14].

Lemma 6.5 (Allouche)**.** *Let* u *and* v *be subwords of a paperfolding word* $\mathbf{f}$*, with* $|u| = |v| \geq 7$*. If* u *and* v *occur at positions of different parity in* $\mathbf{f}$*, then* $u \neq v$*.*

Proof of Proposition 6.4. Suppose to the contrary that

$$wcw = f_i f_{i+1} \cdots f_{i+t} f_{i+t+1} \cdots f_{i+2t}$$

is a subword of $\mathbf{f}$, where $|w| = t$, $t \notin \{2, 4\}$, $t \neq 2^k - 1$ for all $k \geq 1$. Suppose further that $\mathbf{f}$ is chosen so as to minimize t. We consider four cases.

Case 1: $t = 6$. Because the letters in successive even positions of $\mathbf{f}$ alternate between 0 and 1, any subword of $\mathbf{f}$ of length 13 starting at an even position must be of the form

$$0 * 1 * 0 * 1 * 0 * 1 * 0 \quad \text{or} \quad 1 * 0 * 1 * 0 * 1 * 0 * 1,$$

where the $*$ denotes an arbitrary symbol from $\{0, 1\}$. Consequently, if such a subword is of the form wcw, it must be one of the words

$$0011001001100 \quad \text{or} \quad 1100110110011.$$

Similarly, if wcw begins at an odd position, it must be one of the words

$$011001c011001 \quad \text{or} \quad 100110c100110.$$

Taking the odd indexed positions of wcw, we see that if i is even, then either 010010 or 101101 is a subword of a paperfolding word, which is impossible, since neither word obeys the required alternation of 0's and 1's in even indexed positions. Similarly, if i is odd, then either $010c101$ or $101c010$ is a subword of a paperfolding word, which again is impossible for any choice of c.

Case 2: t even, $t \geq 8$. Then w occurs at positions of two different parities in $\mathbf{f}$, contradicting Lemma 6.5.

Case 3: $t \equiv 1 \pmod 4$, $t \geq 5$. Let $\ell \in \{i, i+1\}$ such that ℓ is even. Then $f_\ell \neq f_{\ell+t+1}$, since ℓ and $\ell + t + 1$ are even but $\ell \not\equiv \ell + t + 1 \pmod 4$.

Case 4: $t \equiv 3 \pmod 4$, $t \geq 11$. Let $t = 4m + 3$, where $m \geq 2$ and $m + 1$ is not a power of 2. Let $\ell \in \{i, i+1\}$ such that ℓ is odd. Then

$$w'c'w' = f_\ell f_{\ell+2} \cdots f_{\ell+t-1} f_{\ell+t+1} \cdots f_{\ell+2t-2}$$

is a subword of a paperfolding word, where $|w'| = t' = (t-1)/2 = 2m + 1$. By the argument of Case 3, $t' \not\equiv 1 \pmod 4$. Let us write $t' = 4m' + 3$, where $m' = (m-1)/2$. Since $m + 1$ is not a power of 2, $m' + 1$ is not a power of 2. Thus $11 \leq t' < t$, contradicting the minimality of t. □

The following result is not needed for the proof of Theorem 6.2 but will be useful in the next section.

Proposition 6.6. *Let $\mathbf{f}$ be a paperfolding word. For all $k \geq 1$, $\mathbf{f}$ contains a subword wcw, where w is a non-empty word, c is a single letter, and $|w| = 2^k - 1$.*

Proof. By the perturbed symmetry construction, $\mathbf{f}$ begins with a prefix $zc_0\overline{z}^R$, where $|z| = 2^{k-1} - 1$ and $c_0 \in \{0, 1\}$. Applying the perturbed symmetry map twice to $zc_0\overline{z}^R$, we see that $\mathbf{f}$ begins with a prefix

$$z\ c_0\ \overline{z}^R\ c_1\ z\ \overline{c_0}\ \overline{z}^R\ c_2\ z\ c_0\ \overline{z}^R\ \overline{c_1}\ z\ \overline{c_0}\ \overline{z}^R,$$

where $c_1, c_2 \in \{0, 1\}$. If $c_1 = c_2$, then

$$wcw = \overline{z}^R\ c_1\ z\ \overline{c_0}\ \overline{z}^R\ c_2\ z$$

is the desired subword. If $c_1 \neq c_2$, then

$$wcw = \overline{z}^R\ c_2\ z\ c_0\ \overline{z}^R\ \overline{c_1}\ z$$

is the desired subword. □

We shall also need the following lemma.

Lemma 6.7. *For all $k \geq 1$, no paperfolding word $\mathbf{f}$ contains a subword xx with $|x| = 2^k$.*

Proof. The proof is by induction on k. If $k = 1$, then let $f_i f_{i+1} f_{i+2} f_{i+3}$ be a subword of $\mathbf{f}$. If i is even (resp. odd), then $f_i \neq f_{i+2}$ (resp. $f_{i+1} \neq f_{i+3}$).

Now suppose

$$xx = f_i f_{i+1} \cdots f_{i+2^{k+2}-1}$$

is a subword of $\mathbf{f}$. Let $\ell \in \{i, i+1\}$ such that ℓ is odd. Then

$$x'x' = f_\ell f_{\ell+2} \cdots f_{\ell+2^{k+2}-2}$$

is a subword of a paperfolding word with $|x'| = 2^k$. The result follows by induction. □

We are now ready to prove Theorem 6.2.

Proof of Theorem 6.2. If $\mathbf{f}$ contains a square xx, then writing $x = wc$, where c is a single letter, we see that $\mathbf{f}$ contains the subword wcw. By Proposition 6.4, either $|x| \in \{1, 3, 5\}$, or $|x| = 2^k$ for some $k \geq 1$. But we have seen in Lemma 6.7 that the latter is impossible. □

6.8 An application to β-expansions

In this section we prove a result concerning the lexicographical ordering of the paperfolding words. We then give an application to the theory of expansions of real numbers in irrational bases.

Let σ denote the *shift map*; that is, if $\mathbf{w} = w_0 w_1 w_2 \cdots$ is an infinite word, then $\sigma(\mathbf{w}) = w_1 w_2 w_3 \cdots$. Let $\mathrm{Orb}(\mathbf{w}) = \{\sigma^k(\mathbf{w}) : k \geq 0\}$ denote the *orbit* of the shift map on $\mathbf{w}$. Finally, we define the *orbit closure* of $\mathbf{w}$ as the closure of $\mathrm{Orb}(\mathbf{w})$ with respect to the usual topology on infinite words.

Proposition 6.8. *Let* $\mathbf{f}$ *be the ordinary paperfolding word over* $\{0, 1\}$. *Then* $0\mathbf{f}$ *is the lexicographically least word in the orbit closure of any paperfolding word.*

Proof. Taking the subsequence of $\mathbf{f}$ indexed by the odd positions yields the word $\mathbf{f}$ again, so taking the subsequence of $0\mathbf{f}$ indexed by the even positions yields the word $0\mathbf{f}$.

Let $\mathbf{w} = w_0 w_1 w_2 \cdots$ be the lexicographically least word in the orbit closure of any paperfolding word. Let us assume that $\mathbf{w}$ begins with 0001, since it cannot begin with anything lexicographically smaller. Since $w_0 = w_2$, the following is forced: $w_1 w_3 w_5 w_7 \cdots = 0101 \cdots$.

We shall prove by induction on n that the prefixes of $\mathbf{w}$ of length $2n$ are the prefixes of $0\mathbf{f}$. We have already established the base case, so let us suppose $n \geq 2$ and $w_0 w_1 w_2 \cdots w_{2n-1} = 0 f_0 f_1 f_2 \cdots f_{2n-2}$. Since $w_1 w_3 w_5 w_7 \cdots = 0101 \cdots$, we see that $w_{2n+1} = f_{2n}$. Note that $w_0 w_2 w_4 \cdots w_{2n} = 0 f_1 f_3 f_5 \cdots f_{2n-1}$ is a prefix of a word in the orbit closure of a paperfolding word. By our inductive assumption, $w_0w_1w_2\cdots w_{n-1}w_n$ is the lexicographically least such prefix. Choosing $w_{2n} = w_n = f_{n-1} = f_{2n-1}$ thus ensures that $w_0w_1w_2\cdots w_{2n}w_{2n+1}$ is lexicographically minimal. We have thus established that $\mathbf{w}$ and $0\mathbf{f}$ agree on the first $2(n+1)$ positions, as required. □

We now give an application of Proposition 6.8 to the theory of expansions in non-integer bases (β-expansions)[2]. Let

$$\mathbf{s} = s_1s_2s_3\cdots = \overline{0\mathbf{f}} = 1110110011100100\cdots.$$

For all $k \geq 1$ the strict inequality

$$\overline{\mathbf{s}} <_{\text{lex}} \sigma^k(\mathbf{s}) <_{\text{lex}} \mathbf{s}$$

holds, where $<_{\text{lex}}$ denotes the lexicographical order. It follows that there exists a unique real number β, $1 < \beta < 2$, such that

$$1 = \sum_{n\geq 1} s_n\beta^{-n}, \tag{6.1}$$

and further, by a result of Erdős, Joó, and Komornik [89] (see Allouche, Frougny, and Hare [20, Theorem 2.3]), this expansion of 1 in base β is unique. We wish to show that β is transcendental (see Allouche and Cosnard [18] for a similar result involving the Thue–Morse word). In fact, the transcendence of β follows from a much more general result of Mendès France and van der Poorten [172] that relies on a deep generalization of Mahler's method. We sketch a somewhat simpler proof for the particular case considered here.

First, note that by the Pólya–Carlson Theorem, the power series

$$F(z) = \sum_{n\geq 1} s_nz^n$$

defines a transcendental function with radius of convergence 1. Second, observe that $F(z)$ satisfies the functional equation

$$F(z^2) = zF(z) - \frac{z^3}{1-z^4}$$

(see Dekking, Mendès France, and van der Poorten [79] for an nice explanation of how to derive such an identity). From these two observations it follows by a result of Mahler [166] (see Nishioka [183, Theorem 1.2]) that $F(z)$ takes transcendental values at every algebraic point inside the unit disc (except zero). Thus, if β were algebraic, then $F(1/\beta)$ would be transcendental, contradicting (6.1). We conclude that β is transcendental.

[2] The result of Proposition 6.8 appears in Kao, Rampersad, Shallit, and Silva [135], but the application to β-expansions given here is new.

The number β is an explicit example of a transcendental number in Blanchard's "Class 3" [44], i.e., a real number β for which the β-expansion of 1 is aperiodic but does not contain arbitrarily large runs of 0's. Previously known examples are the Komornik–Loreti constant [146] (proved to be transcendental by Allouche and Cosnard [18]) and the self-Sturmian numbers of Chi and Kwon [61].

6.9 A language-theoretic result

Here we recall a result of Lehr [157, 17] concerning the set of subwords of the paperfolding words. The following theorem is an immediate consequence of Theorem 6.2 and the pumping lemma for context-free languages.

Theorem 6.9 (Lehr; Allouche and Bousquet-Mélou). *The set of subwords of the paperfolding words is not context-free.*

We may also derive this result by considering the generating series of the paperfolding words. Allouche and Bousquet-Mélou [17] showed that the number $f(n)$ of subwords of length n of the paperfolding words is given by

$$f(n) = \begin{cases} 2^n, & \text{if } 1 \leq n \leq 3; \\ 2^{a+1}n - 5 \cdot 4^{a-1}, & \text{if } 2^a \leq n < 2^a + 2^{a-1}, \text{ where } a \geq 2; \\ 3 \cdot 2^a n - 11 \cdot 4^{a-1}, & \text{if } 2^a + 2^{a-1} \leq n < 2^a + 2^{a-1} + 2^{a-2}, \text{ where } a \geq 2; \\ 2^{a+1}n - 4^a, & \text{if } 2^a + 2^{a-1} + 2^{a-2} \leq n < 2^{a+1}, \text{ where } a \geq 2; \end{cases}$$

and they showed that the corresponding generating function

$$F(X) = \sum_{n \geq 1} f(n) X^n$$

is transcendental. We may also deduce the transcendence of $F(X)$ by considering the series $H(X)$, whose coefficients form the second difference sequence of $f(n)$. Noting that $H(X)$ is a gap series, we may apply Fabry's gap theorem along with Theorem 5.15 to derive the desired result.

6.10 Avoiding repetitions in arithmetic progressions

In this section we use the paperfolding words to construct infinite words avoiding squares (resp. overlaps) in all arithmetic progressions of odd difference.

The following result is implicit in the work of Avgustinovich, Fon-Der-Flaass, and Frid (see the proof of [30, Theorem 3] as well as [30, Example 2]).

Theorem 6.10 (Avgustinovich, Fon-Der-Flaass, and Frid)**.** *If w is a finite arithmetic subsequence of odd difference of a paperfolding word, then w is a subword of a paperfolding word.*

Corollary 6.11. *There exists an infinite word over a binary alphabet that contains no 3^+-powers in arithmetic progressions of odd difference.*

Proof. It follows from Corollary 6.3 and Theorem 6.10 that all paperfolding words have this property. □

We note further that the 3^+ of the preceding corollary may not be replaced by 3. A computer search suffices to verify that all sufficiently long binary words contain a cube in an arithmetic progression of odd difference. The longest binary words that do not contain a cube in an arithmetic progression of odd difference are the following words of length 13:

$$\begin{array}{ll} 0010011001100 & 0101100110011 \\ 1010011001100 & 1101100110011. \end{array}$$

The problem of avoiding repetitions in arithmetic progressions seems to have first been studied by Carpi [55] and subsequently by Currie and Simpson [73]. Downarowicz [81] studied a related problem.

Theorem 6.12 (Carpi)**.** *There exists an infinite word over a 4-letter alphabet that contains no squares in arithmetic progressions of odd difference.*

The word

$$\mathbf{c} = 515351735153717351 53\cdots$$

constructed by Carpi satisfying the conditions of this theorem is over the alphabet $\{1, 3, 5, 7\}$ and is generated by iterating the morphism $1 \to 53$, $3 \to 73$, $5 \to 51$, $7 \to 71$, starting with the symbol 5. It can also be derived from a paperfolding sequence, as we shall see below. The alphabet size of 4 in Theorem 6.12 is optimal, since the longest words over the alphabet $\{0, 1, 2\}$ that avoid squares in all odd difference arithmetic progressions are the words

$$010212021 \quad 012010201$$

of length 9, along with the words obtained from these by permuting the alphabet symbols in all possible ways.

Let $\mathbf{f} = f_0 f_1 f_2 \cdots$ be any paperfolding word over $\{1, 4\}$. Define $\mathbf{v} = v_0 v_1 v_2 \cdots$ by

$$\begin{array}{rcl} v_{4n} & = & 2 \\ v_{4n+2} & = & 3 \\ v_{2n+1} & = & f_{2n+1}, \end{array}$$

for all $n \geq 0$. In other words, we have recoded the periodic subsequence formed by taking the even positions of $\mathbf{f}$ by mapping $1 \rightarrow 2$ and $4 \rightarrow 3$ (or vice-versa). For example, if

$$\mathbf{f} = 1141144111441441 \cdots$$

is the ordinary paperfolding word over $\{1, 4\}$, then

$$\mathbf{v} = 2131243121342431 \cdots .$$

Theorem 6.13. *Let* $\mathbf{v}$ *be any word obtained from a paperfolding word* $\mathbf{f}$ *by the construction described above. Then the word* $\mathbf{v}$ *contains no squares in arithmetic progressions of odd difference but does not avoid* r*-powers for any real* $r < 2$.

Proof. By the construction of $\mathbf{v}$, any arithmetic subsequence

$$w = v_{i_0} v_{i_1} \cdots v_{i_k}$$

of odd difference of $\mathbf{v}$ can be obtained from the corresponding subsequence

$$x = f_{i_0} f_{i_1} \cdots f_{i_k}$$

of $\mathbf{f}$ by recoding the symbols in either the even positions of x or the odd positions of x by mapping $1 \rightarrow 2$ and $4 \rightarrow 3$ (or vice-versa). Note that this recoding cannot create any new squares. Now suppose that $\mathbf{v}$ contains a square ww in an arithmetic progression of odd difference. Let xx be the corresponding subsequence of $\mathbf{f}$. By Theorems 6.2 and 6.10, $|x| \in \{1, 3, 5\}$ and hence $|w| \in \{1, 3, 5\}$. Clearly, $|w| = 1$ is impossible. If $|w| = 3$, then ww has one of the forms $(\,{*}\,2\,{*}\,)(\,3\,{*}\,2\,)$, $(\,{*}\,3\,{*}\,)(\,2\,{*}\,3\,)$, $(\,2{*}3\,)({*}2{*})$, or $(\,3{*}2\,)({*}3{*})$, where the $*$ denotes an arbitrary symbol from $\{1, 4\}$. Clearly, none of these can be squares. A similar argument applies for $|w| = 5$.

That $\mathbf{v}$ does not avoid r-powers for any $r < 2$ follows easily from Proposition 6.6. □

The word $\mathbf{c}$ constructed by Carpi, after relabeling the alphabet symbols by the map $1 \rightarrow 2$, $3 \rightarrow 3$, $5 \rightarrow 1$, $7 \rightarrow 4$, is the word $1\mathbf{v}$, where $\mathbf{v}$ is constructed from the ordinary paperfolding word as described above. Note that since there are uncountably many paperfolding words $\mathbf{f}$, there are uncountably many words $\mathbf{v}$ over a 4-letter alphabet that contain no squares in arithmetic progressions of odd difference. We offer the following conjectures regarding such words.

Conjecture 6.14. *For all real numbers* $r < 2$, r*-powers are not avoidable in arithmetic progressions of odd difference over a 4-letter alphabet.*

A computer search confirms that Conjecture 6.14 holds for all $r \leq 7/4$.

Conjecture 6.15. *Any infinite word over a 4-letter alphabet that avoids squares in arithmetic progressions of odd difference is in the orbit closure of one of the words* $\mathbf{v}$ *constructed above.*

Next we consider words over a ternary alphabet.

Theorem 6.16. *There exists an infinite word over a ternary alphabet that contains no 2^+-powers (overlaps) and no squares xx, $|x| \geq 2$, in arithmetic progressions of odd difference.*

Proof. Let $\mathbf{v} = v_0v_1v_2\cdots$ be any word obtained from a paperfolding word by the construction described above. Let h be the morphism that sends $1 \to 00$, $2 \to 11$, $3 \to 12$, $4 \to 02$. Then $\mathbf{w} = w_0w_1w_2\cdots = h(\mathbf{v})$ has the desired properties.

Suppose to the contrary that there exists $i \geq 0$, j odd, and $t \geq 2$ such that for $s \in \{0, \ldots, t-1\}$, $w_{i+sj} = w_{i+(s+t)j}$. Note that there exists $a \in \{0, 1\}$ such that for $\ell \equiv 0 \pmod 4$, $w_\ell = a$ and for $\ell \equiv 2 \pmod 4$, $w_\ell = \overline{a}$. We consider four cases.

Case 1: $t = 3$. Because the letters in successive even positions of $\mathbf{w}$ alternate between 0 and 1, $w_iw_{i+j}\cdots w_{i+5j}$ is one of the words 001001, 011011, 100100, or 110110. Thus there exists $s \in \{0, 1, 2\}$ such that $w_{i+sj} \neq w_{i+(s+4)j}$. Now consider the morphism h. The symbol 0 only occurs in the images of 1 and 4, and the symbol 1 only occurs in the images of 2 and 3. Let $i' = \lfloor (i+sj)/2 \rfloor$. Since $w_{i+sj} \neq w_{i+(s+4)j}$, we have that either $v_{i'} \in \{1, 4\}$ and $v_{i'+2j} \in \{2, 3\}$, or vice versa. Either case is impossible, since the symbols 1 and 4 only occur in positions of odd parity in $\mathbf{v}$, and the symbols 2 and 3 only occur in positions of even parity in $\mathbf{v}$, but i' and $i' + 2j$ both have the same parity.

Case 2: t odd, $t \geq 5$. Since j is odd, $\{i \pmod 8, i+j \pmod 8, \ldots, i+(2t-1)j \pmod 8\}$ is a complete set of residues $\pmod 8$. Since $\mathbf{v}$ contains a 3 in every position congruent to $2 \pmod 4$, $\mathbf{w}$ contains a 2 in every position congruent to $5 \pmod 8$. Thus there exists $s \in \{0, \ldots, 2t-1\}$ such that $w_{i+sj} = 2$. If $s < t$, then since t is odd, $s \not\equiv s + t \pmod 2$, and consequently, $i + sj \not\equiv i + (s+t)j \pmod 2$. But $\mathbf{w}$ only contains 2's in positions of even parity, so $w_{i+sj} \neq w_{i+(s+t)j}$, contrary to our assumption. Similarly, if $s \geq t$, we have $w_{i+(s-t)j} \neq w_{i+sj}$.

Case 3: $t \equiv 2 \pmod 4$. Then either $w_i \neq w_{i+tj}$ or $w_{i+j} \neq w_{i+(t+1)j}$, accordingly as i is even or odd, contrary to our assumption.

Case 4: $t \equiv 0 \pmod 4$. Let $k \in \{i, i+j\}$ such that k is odd. Let $k' = \lfloor k/2 \rfloor$. It follows from the definition of h that for $s \in \{0, \ldots, t-1\}$, $v_{k'+sj}$ is uniquely determined by the value of w_{k+2sj} and the congruence class of $k + 2sj \pmod 4$:

- if $w_{k+2sj} = 0$, then $v_{k'+sj} = 1$;
- if $w_{k+2sj} = 1$, then $v_{k'+sj} = 2$; and
- if $w_{k+2sj} = 2$, then $v_{k'+sj}$ is either 3 or 4, accordingly as $k + 2sj \equiv 1$ or $3 \pmod 4$.

From this observation, combined with the fact that $k+2sj \equiv k+(2s+t)j \pmod 4$, we see that since $w_kw_{k+2j}\cdots w_{k+2(t-1)j}$ is a square, $v_{k'}v_{k'+j}\cdots v_{k'+(t-1)j}$ is also a square in an arithmetic progression of odd difference j in $\mathbf{v}$, a contradiction.

These four cases cover all possibilities. It remains to consider the existence of the cubes 000, 111, and 222. Suppose there exists $w_i w_{i+j} w_{i+2j} \in \{000, 111, 222\}$ for some $i \geq 0$ and j odd. Since $\mathbf{w}$ only contains 2's in positions of even parity, we may suppose $w_i w_{i+j} w_{i+2j} \in \{000, 111\}$. If i is even, then $i + 2j$ is even and $i \not\equiv i + 2j \pmod 4$, so $w_i \neq w_{i+2j}$. If i is odd, then by the same reasoning as in Case 4 above, $v_{\lfloor i/2 \rfloor} v_{\lfloor i/2 \rfloor + j}$ is a square in an arithmetic progression of odd difference in $\mathbf{v}$, a contradiction. □

The alphabet size of 3 in Theorem 6.16 is optimal, since the longest words over the alphabet $\{0, 1\}$ that avoid overlaps in all odd difference arithmetic progressions are the words

$$0010011001 \quad 0101100110 \quad 0110100101$$

of length 10, along with their complements.

The next result improves upon the result of Entringer, Jackson, and Schatz [87] noted previously. For the proof, see Kao, Rampersad, Shallit, and Silva [135].

Theorem 6.17. *There exists an infinite word over a binary alphabet that contains no squares xx with $|x| \geq 3$ in any arithmetic progression of odd difference.*

6.11 Avoiding repetitions in higher dimensions

In this section we apply our previous constructions to define non-repetitive labelings of the integer lattice.

An infinite word $\mathbf{w}$ over a finite alphabet A is a map from $\mathbb{N}$ to A, where we write w_n for $\mathbf{w}(n)$. Now consider a map $\mathbf{w}$ from $\mathbb{N}^2$ to A, where we write $w_{m,n}$ for $\mathbf{w}(m, n)$. We call such a $\mathbf{w}$ a *2-dimensional word.* A word $\mathbf{x}$ is a *line* of $\mathbf{w}$ if there exists i_1, i_2, j_1, j_2 such that $\gcd(j_1, j_2) = 1$, and for $t \geq 0$,

$$x_t = w_{i_1 + j_1 t, i_2 + j_2 t}.$$

Carpi [55] proved the following surprising result.

Theorem 6.18 (Carpi)**.** *There exists a 2-dimensional word $\mathbf{w}$ over a 16-letter alphabet such that every line of $\mathbf{w}$ is squarefree.*

Proof. Let $\mathbf{u} = u_0 u_1 u_2 \cdots$ and $\mathbf{v} = v_0 v_1 v_2 \cdots$ be any infinite words over the alphabet $A = \{1, 2, 3, 4\}$ that avoid squares in all arithmetic progressions of odd difference. We define $\mathbf{w}$ over the alphabet $A \times A$ by

$$w_{m,n} = (u_m, v_n).$$

Consider an arbitrary line

$$\begin{aligned} \mathbf{x} &= (w_{i_1 + j_1 t, i_2 + j_2 t})_{t \geq 0}, \\ &= (u_{i_1 + j_1 t}, v_{i_2 + j_2 t})_{t \geq 0}, \end{aligned}$$

for some $i_1, i_2,\ j_1, j_2$, with $\gcd(j_1, j_2) = 1$. Without loss of generality, we may assume j_1 is odd. Then the word $(u_{i_1+j_1t})_{t\geq 0}$ is an arithmetic subsequence of odd difference of $\mathbf{u}$ and hence is squarefree. The line $\mathbf{x}$ is therefore also squarefree. □

A computer search shows that there are no 2-dimensional words $\mathbf{w}$ over a 7-letter alphabet, such that every line of $\mathbf{w}$ is squarefree. It remains an open problem to determine if the alphabet size of 16 in Theorem 6.18 is best possible.

Using the results of Theorems 6.11, 6.16, and 6.17 respectively, one proves the following theorems in a manner analogous to that of Theorem 6.18.

Theorem 6.19. *There exists a 2-dimensional word* $\mathbf{w}$ *over a 4-letter alphabet such that every line of* $\mathbf{w}$ *is* 3^+*-power-free.*

Theorem 6.20. *There exists a 2-dimensional word* $\mathbf{w}$ *over a 9-letter alphabet such that every line of* $\mathbf{w}$ *is* 2^+*-power-free (overlapfree).*

Theorem 6.21. *There exists a 2-dimensional word* $\mathbf{w}$ *over a 4-letter alphabet such that every line of* $\mathbf{w}$ *avoids squares* xx*, where* $|x| \geq 3$*.*

The reader will easily see how to generalize these results to higher dimensions. The reader may also wish to contrast the results of this section with that of the classical Hales–Jewett Theorem [114], a powerful multi-dimensional generalization of van der Waerden's theorem.

Dumitrescu and Radoičić [82] proved that every 2-dimensional word over a 2-letter alphabet must contain a line containing a cube. Grytczuk [112] presented the problem of determining the *Thue threshold* of $\mathbb{N}^2$, namely, the smallest integer t such that there exists an integer $k \geq 2$ and a 2-dimensional word $\mathbf{w}$ over a t-letter alphabet such that every line of $\mathbf{w}$ is k-power-free. Carpi's result showed that $t \leq 16$; Theorem 6.19 shows that $t \leq 4$.

In this chapter we generalized the concept of avoiding repetitions in words by considering the avoidance of repetitions in arithmetic progressions. In the next chapter we consider another generalization: we avoid approximate repetitions, rather than exact repetitions.

Chapter 7

Avoiding Approximate Squares

7.1 Introduction

In this chapter we generalize the notion of square to that of *approximate square* and construct words avoiding approximate squares[1]. We consider two different definitions of approximate square.

For words x, x' of the same length, the *Hamming distance* $d(x, x')$ is the number of positions in which x and x' differ. For example, $d(\texttt{mattino}, \texttt{cammino}) = 3$. A word xx' with $|x| = |x'|$ is a *c-approximate square* if $d(x, x') \leq c$. In particular, a 0-approximate square is simply a square in the ordinary sense. We note that a c-approximate square is typically refered to as a c-approximate tandem repeat in the biological sequence analysis literature.

A word z *avoids c-approximate squares* if for all its subwords xx' where $|x| = |x'|$ we have $d(x, x') \geq \min(c + 1, |x|)$. This is the "additive" definition of approximate square; we also define a "multiplicative" version. Given two words x, x' of the same length, we define their *similarity* $s(x, x')$ as the fraction of the number of positions in which x and x' agree. Formally,

$$s(x, x') := \frac{|x| - d(x, x')}{|x|}.$$

For example, $s(\texttt{mattino}, \texttt{cammino}) = 4/7$. The *similarity* of a finite word z is defined to be

$$\alpha = \max_{\substack{xx' \text{ a subword of } z \\ |x| = |x'|}} s(x, x');$$

we say such a word is α-similar.

[1] The results in this chapter can be found in Krieger, Ochem, Rampersad, and Shallit [151]. The morphisms used in the proofs were discovered by computer searches performed by J. Shallit and P. Ochem.

An infinite word $\mathbf{z}$ is *α-similar* if

$$\alpha = \sup_{\substack{xx' \text{ a subword of } \mathbf{z} \\ |x|=|x'|}} s(x,x')$$

and there exists at least one subword xx' with $|x| = |x'|$ and $s(x,x') = \alpha$. Otherwise, if

$$\alpha = \sup_{\substack{xx' \text{ a subword of } \mathbf{z} \\ |x|=|x'|}} s(x,x'),$$

but α is not attained by any subword xx' of $\mathbf{z}$, then $\mathbf{z}$ is *α^--similar*.

For example, recall the squarefree word

$$\mathbf{x} = 210201210120210\cdots$$

that we derived from the Thue–Morse word in Section 2.2. Since the Thue–Morse word contains arbitrarily large squares, it is easy to see that $\mathbf{x}$ contains arbitrarily large subwords of the form $wawb$, where $w \in \{0,1,2\}^*$ and $a, b \in \{0,1,2\}$, $a \neq b$ (see the proof of Theorem 2.23). Thus $\mathbf{x}$ contains arbitrarily large subwords that match in all but one position, so $\mathbf{x}$ is 1^--similar.

7.2 Words of low similarity

In this section we consider the multiplicative definition of approximate square. Given an alphabet Σ of size k, we wish to determine the smallest similarity possible over all infinite words over Σ. We call this the *similarity coefficient* of k.

In order to get an idea of what the similarity coefficient should be for a given alphabet, we begin by performing some computations. These computations are done using a standard backtracking algorithm. The backtracking algorithm works as follows. Imagine an infinite rooted k-ary tree, where each of the k outgoing edges at a given vertex is labeled by a letter of the alphabet. For any given vertex v, the label of v is obtained by concatenating the labels of the edges encountered along the path from the root to v. The backtracking search amounts to a depth-first search through this tree. At each vertex v, the label of v is checked to see if it has the desired property (e.g., "is 3/4-similar"). If not, the subtree rooted at v is "pruned" (i.e., is not searched any further). In either case, the depth-first search continues. If during the course of the search we prune so much of the original tree that we are left with only a finite tree, we end the search and conclude that there is no infinite word with the desired property.

Table 7.1 reports the results of computations performed by J. Shallit. The entries in the column "Similarity Coefficient α" represent lower bounds on the similarity coefficent for alphabet size k. The other columns give statistics concerning the finite tree obtained at the end of the search. For $k = 7$ we were not able to obtain a good estimate of the similarity coefficient by computer calculations.

Alphabet Size k	Similarity Coefficient α	Height of Tree	Number of Leaves	Number of Maximal Words
2	1	3	4	1
3	3/4	41	2475	36
4	1/2	9	382	6
5	2/5	75	3902869	48
6	1/3	17	342356	480
7	?	?	?	?
8	1/4	71	—	—

Table 7.1: Similarity bounds

The similarity coefficient of $k = 2$ is clearly 1, since we cannot avoid squares on a binary alphabet. Next we determine the similarity coefficients for $k = 3$ and $k = 4$.

Theorem 7.1. *There exists an infinite* $3/4$*-similar word* $\mathbf{w}$ *over* $\{0, 1, 2\}$.

Before proceeding with the proof of this result, let us first recall the result of Dejean [75] mentioned in Section 1.7. Dejean proved that over a 3-letter alphabet there exists an infinite $(7/4)^+$-power-free word and the 7/4 is best possible. Note that any 3/4-similar word must be $(7/4)^+$-power-free. Of course, being 3/4-similar is a much stronger property, so Theorem 7.1 is a nice improvement of Dejean's result. We now proceed with the proof.

Proof of Theorem 7.1. Let h be the 24-uniform morphism defined by

$$\begin{aligned} 0 &\rightarrow 012021201021012102120210 \\ 1 &\rightarrow 120102012102120210201021 \\ 2 &\rightarrow 201210120210201021012102. \end{aligned}$$

The following lemma may be verified computationally.

Lemma 7.2. *Let* $a, b, c \in \{0, 1, 2\}$*,* $a \neq b$*. Let* w *be any subword of length* 24 *of* $h(ab)$*. If* w *is neither a prefix nor a suffix of* $h(ab)$*, then* $h(c)$ *and* w *mismatch in at least* 9 *positions.*

Let $\mathbf{w} = h^\omega(0)$. We shall show that $\mathbf{w}$ has the desired property. We argue by contradiction. Suppose that $\mathbf{w}$ contains a subword yy' with $|y| = |y'|$ such that y and y' match in more than $3/4 \cdot |y|$ positions. Let us suppose further that $|y|$ is minimal.

We may verify computationally that $\mathbf{w}$ contains no such subword yy' where $|y| \leq 72$. We therefore assume from now on that $|y| > 72$.

Let $w = a_1a_2 \cdots a_n$ be a word of minimal length such that $h(w) = xyy'z$ for some $x, z \in \{0, 1, 2\}^*$. By the minimality of w, we have $0 \le |x|, |z| < 24$.

For $i = 1, 2, \ldots, n$, define $A_i = h(a_i)$. Then if $h(w) = xyy'z$, we can write

$$h(w) = A_1A_2 \cdots A_n = A_1'A_1''A_2 \cdots A_{j-1}A_j'A_j''A_{j+1} \cdots A_{n-1}A_n'A_n''$$

where

$$\begin{aligned}
A_1 &= A_1'A_1'' \\
A_j &= A_j'A_j'' \\
A_n &= A_n'A_n'' \\
x &= A_1' \\
y &= A_1''A_2 \cdots A_{j-1}A_j' \\
y' &= A_j''A_{j+1} \cdots A_{n-1}A_n' \\
z &= A_n'',
\end{aligned}$$

and $|A_1''|, |A_j''| > 0$. See Figure 7.1.

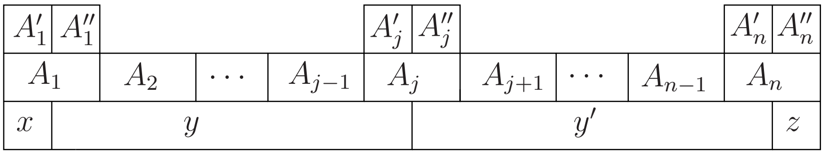

Figure 7.1: The string $xyy'z$ within $h(w)$

If $|A_1''| > |A_j''|$, then, writing y and y' atop one another, as illustrated in Figure 7.2, one observes that for $t = j + 1, j + 2, \ldots, n - 1$, each A_t "lines up" with a subword, say B_t, of $A_{t-j}A_{t-j+1}$. We now apply Lemma 7.2 to conclude that each A_t mismatches with B_t in at least 9 of 24 positions. Consequently, y and y' mismatch in at least $9(j - 2)$ positions. Since $j \ge |y|/24 + 1$, we have that $9(j - 2) \ge 9(|y|/24 - 1)$. However, $9(|y|/24 - 1) > |y|/4$ for $|y| > 72$, so that y and y' mismatch in more than $1/4 \cdot |y|$ positions, contrary to our assumption.

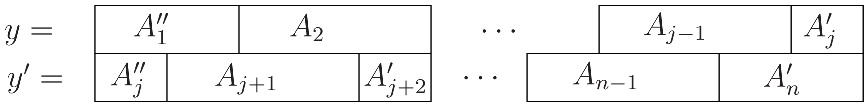

Figure 7.2: The case $|A_1''| > |A_j''|$

If $|A_1''| < |A_j''|$, as illustrated in Figure 7.3, then a similar argument shows that y and y' mismatch in more than $1/4 \cdot |y|$ positions, contrary to our assumption.

Therefore $|A_1''| = |A_j''|$. We first observe that any pair of words taken from $\{h(0), h(1), h(2)\}$ mismatch at every position. We now consider several cases.

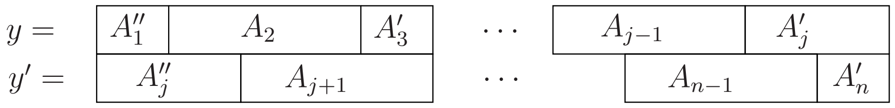

Figure 7.3: The case $|A''_1| < |A''_j|$

Case 1: $A_1 = A_j = A_n$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

Case 2: $A_1 = A_j \neq A_n$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in at least as many positions as y and y'.

Case 3: $A_1 \neq A_j = A_n$. Then letting $u = A_2A_3 \cdots A_j$ and $u' = A_{j+1}A_{j+2} \cdots A_n$, we see that u and u' match in at least as many positions as y and y'.

Case 4: $A_1 = A_n \neq A_j$. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

Case 5: A_1, A_j, and A_n are all distinct. Then letting $u = A_1A_2 \cdots A_{j-1}$ and $u' = A_jA_{j+1} \cdots A_{n-1}$, we see that u and u' match in exactly the same number of positions as y and y'.

We finish the argument by considering the word uu'. First observe that either

$$uu' = h(a_1a_2 \cdots a_{j-1})h(a_ja_{j+1} \cdots a_{n-1})$$

or

$$uu' = h(a_2a_3 \cdots a_j)h(a_{j+1}a_{j+2} \cdots a_n).$$

Without loss of generality, let us assume that the first case holds.

Recall our previous observation that the words $h(0)$, $h(1)$, and $h(2)$ have distinct letters at every position. Suppose then that there is a mismatch between u and u' occuring within blocks A_t and A_{t+j} for some t, $1 \leq t \leq j$. Then A_t and A_{t+j} mismatch at every position. Moreover, we have $a_j \neq a_{j+t}$. Conversely, if A_t and A_{t+j} match at any single position, then they match at every position, and we have $a_t = a_{t+j}$.

Let $v = a_1a_2 \cdots a_{j-1}$ and $v' = a_ja_{j+1} \cdots a_{n-1}$. Let m be the number of matches between u and u'. From our previous observations we deduce that the number of matches m' between v and v' is $m/24$, but since $|v| = |u|/24$, $m'/|v| = m/|u|$. Thus, if $m/|u| > 3/4$, as we have assumed, then $m'/|v| > 3/4$. But the set $\{h(0), h(1), h(2)\}$ is a code, so that vv' is the unique pre-image of uu'. The word vv' is thus a subword of $\mathbf{w}$, contradicting the assumed minimality of yy'. We conclude that no such yy' occurs in $\mathbf{w}$, and this completes the argument that $\mathbf{w}$ is 3/4-similar. □

Next, we consider the case $k = 4$.

Theorem 7.3. *There exists an infinite* $1/2$*-similar word* $\mathbf{x}$ *over* $\{0, 1, 2, 3\}$.

Proof. Let g be the 36-uniform morphism defined by

$$\begin{aligned}
0 &\rightarrow 012132303202321020123021203020121310\\
1 &\rightarrow 123203010313032131230132310131232021\\
2 &\rightarrow 230310121020103202301203021202303132\\
3 &\rightarrow 301021232131210313012310132313010203.
\end{aligned}$$

Then $\mathbf{x} = g^{\omega}(0)$ has the desired property. The proof is entirely analogous to that of Theorem 7.1, with the following lemma used in place of Lemma 7.2.

Lemma 7.4. *Let $a, b, c \in \{0, 1, 2, 3\}$, $a \neq b$. Let w be any subword of length 36 of $g(ab)$. If w is neither a prefix nor a suffix of $g(ab)$, then $g(c)$ and w mismatch in at least 21 positions.*

□

Over larger alphabets we no longer have constructive proofs. We turn instead to a probabilistic lemma first proved by Erdős and Lovász [90]. The following form of the Lovász Local Lemma was given by Spencer [228].

Lemma 7.5 (Lovász Local Lemma; asymmetric version)**.** *Let I be a finite set, and let $\{A_i\}_{i\in I}$ be events in a probability space. Let E be a set of pairs $(i, j) \in I \times I$ such that A_i is mutually independent of all the events $\{A_j : (i, j) \notin E\}$. Suppose there exist real numbers $\{x_i\}_{i\in I}$, $0 \leq x_i < 1$, such that for all $i \in I$,*

$$\text{Prob}(A_i) \leq x_i \prod_{(i,j)\in E} (1 - x_j).$$

Then

$$\text{Prob}\left(\bigcap_{i\in I} \overline{A_i}\right) \geq \prod_{i\in I}(1 - x_i) > 0.$$

For examples of how the local lemma has been successfully applied to problems in combinatorics on words, see Beck [32], Currie [71], or Grytczuk [111].

Theorem 7.6. *Let $c > 1$ be an integer. There exists an infinite $1/c$-similar word.*

Proof. Let k and N be positive integers, and let $w = w_1 w_2 \cdots w_N$ be a random word of length N over a k-letter alphabet Σ. Here each letter of w is chosen uniformly and independently at random from Σ.

Let

$$I = \{(t, r) : 0 \leq t < N, 1 \leq r \leq \lfloor (N - t)/2 \rfloor\}.$$

For $i = (t, r) \in I$, write $y = w_t \cdots w_{t+r-1}$ and $y' = w_{t+r} \cdots w_{t+2r-1}$. Let A_i denote the event $s(y, y') > 1/c$. A crude overestimate of $\operatorname{Prob}(A_i)$ is

$$\begin{aligned}
\operatorname{Prob}(A_i) &\leq \frac{\binom{r}{\lfloor r/c \rfloor + 1} k^{\lfloor r/c \rfloor + 1} k^{2r - 2(\lfloor r/c \rfloor + 1)}}{k^{2r}} \\
&\leq \binom{r}{\lfloor r/2 \rfloor} k^{-r/c} \\
&\leq 2^r k^{-r/c}.
\end{aligned}$$

For all positive integers r, define $\xi_r = 2^{-2r}$. For any real number $\alpha \leq 1/2$, we have $(1-\alpha) \geq e^{-2\alpha}$. Hence, $(1 - \xi_r) \geq e^{-2\xi_r}$. For $i = (t, r) \in I$, define $x_i = \xi_r$. Let E be as in the local lemma. Note that a subword of length $2r$ of w overlaps with at most $2r + 2s - 1$ subwords of length $2s$. Thus, for all $i = (t, r) \in I$, we have

$$\begin{aligned}
x_i \prod_{(i,j) \in E} (1 - x_j) &\geq \xi_r \prod_{s=1}^{\lfloor N/2 \rfloor} (1 - \xi_s)^{2r+2s-1} \\
&\geq \xi_r \prod_{s=1}^{\infty} (1 - \xi_s)^{2r+2s-1} \\
&\geq \xi_r \prod_{s=1}^{\infty} e^{-2\xi_s(2r+2s-1)} \\
&\geq 2^{-2r} \prod_{s=1}^{\infty} e^{-2(2^{-2s})(2r+2s-1)} \\
&\geq 2^{-2r} \exp\left[-2\left(2r \sum_{s=1}^{\infty} \frac{1}{2^{2s}} + \sum_{s=1}^{\infty} \frac{2s-1}{2^{2s}}\right)\right] \\
&\geq 2^{-2r} \exp\left[-2\left(2r\left(\frac{1}{3}\right) + \frac{5}{9}\right)\right] \\
&\geq 2^{-2r} \exp\left(-\frac{4}{3}r - \frac{10}{9}\right).
\end{aligned}$$

The hypotheses of the local lemma are met if

$$2^r k^{-r/c} \leq 2^{-2r} \exp\left(-\frac{4}{3}r - \frac{10}{9}\right).$$

Taking logarithms, we require

$$r \log 2 - \frac{r}{c} \log k \leq -2r \log 2 - \frac{4}{3}r - \frac{10}{9}.$$

Rearranging terms, we require

$$c\left(3 \log 2 + \frac{4}{3} + \frac{10}{9r}\right) \leq \log k.$$

The left side of this inequality is largest when $r = 1$, so we define

$$d_1 = 3\log 2 + \frac{4}{3} + \frac{10}{9},$$

and insist that $c \cdot d_1 \leq \log k$. Hence, for $k \geq e^{c \cdot d_1}$, we may apply the local lemma to conclude that with positive probability, w is $1/c$-similar. Since $N = |w|$ is arbitrary, we conclude that there are arbitrarily large such w. By König's Infinity Lemma (Theorem 1.1), there exists an infinite $1/c$-similar word, as required. □

7.3 Words avoiding c-approximate squares

In this section we consider the "additive" version of the problem. The table below reflects the results of computations performed by J. Shallit using a backtracking algorithm: there is no infinite word over a k-letter alphabet that avoids c-approximate squares for the k and c given below.

Alphabet Size k	c	Height of Tree	Number of Leaves	Number of Maximal Words
2	0	4	3	1
3	1	5	23	2
4	2	7	184	6
5	2	11	3253	24
6	3	11	35756	960
7	4	13	573019	6480
8	5	15	-	-

Table 7.2: Lower bounds on avoiding c-approximate squares

Theorem 7.7. *There is an infinite word over a* 3*-letter alphabet that avoids* 0*-approximate squares, and the* 0 *is best possible.*

Proof. Any ternary word avoiding squares, such the word constructed in Section 2.2, satisfies the conditions of the theorem. The result is best possible, from Table 7.2. □

Theorem 7.8. *There is an infinite word over a* 4*-letter alphabet that avoids* 1*-approximate squares, and the* 1 *is best possible.*

Proof. Let $\mathbf{c}$ be any squarefree word over $\{0, 1, 2\}$, and consider the image under the morphism α defined by

$$\begin{aligned} 0 &\rightarrow 012031023120321031201321032013021320123013203123 \\ 1 &\rightarrow 012031023120321023103213021032013210312013203123 \\ 2 &\rightarrow 012031023012310213023103210231203210312013203123 \end{aligned}$$

The resulting word $\mathbf{d} = \alpha(\mathbf{c})$ avoids 1-approximate squares. The result is best possible, from Table 7.2.

The proof is similar to that of Theorem 7.1. Suppose to the contrary that $\mathbf{d}$ contains a 1-approximate square yy', $|y| = |y'|$. We may verify computationally that $\mathbf{d}$ contains no such subword yy' where $|y| \leq 96$. We therefore assume from now on that $|y| > 96$.

Let $w = a_1a_2 \cdots a_n$ be a word of minimal length such that $\alpha(w) = xyy'z$ for some $x, z \in \{0, 1, 2, 3\}^*$. By the minimality of w, we have $0 \leq |x|, |z| < 48$.

For $i = 1, 2, \ldots, n$, define $A_i = \alpha(a_i)$. Just as in the proof of Theorem 7.1, we write

$$\alpha(w) = A_1A_2 \cdots A_n = A_1'A_1''A_2 \cdots A_{j-1}A_j'A_j''A_{j+1} \cdots A_{n-1}A_n'A_n'',$$

so that the situation illustrated in Figure 4.1 applies to $xyy'z$ within $\alpha(w)$. We now make the following observations regarding the morphism α:

1. Let $a, b, c \in \{0, 1, 2\}$, $a \neq b$. Let u be any subword of length 48 of $\alpha(ab)$. If u is neither a prefix nor a suffix of $\alpha(ab)$, then $\alpha(c)$ and u mismatch in at least 18 positions.

2. Let $a, b \in \{0, 1, 2\}$, $a \neq b$. Then $\alpha(a)$ and $\alpha(b)$ mismatch in at least 18 positions.

3. Let u, u', v, v' be words satisfying the following:

   - $|u| = |u'|$, $|v| = |v'|$, and $|uv| = |u'v'| = 48$;
   - each of u and u' is a suffix of a word in $\{\alpha(0), \alpha(1), \alpha(2)\}$; and
   - each of v and v' is a prefix of a word in $\{\alpha(0), \alpha(1), \alpha(2)\}$.

   Then either $uv = u'v'$ or uv and $u'v'$ mismatch in at least 18 positions.

4. Let $a \in \{0, 1, 2\}$. Then $\alpha(a)$ is uniquely determined by either its prefix of length 17 or its suffix of length 17.

From the first observation, we deduce, as in the proof of Theorem 7.1, that the cases illustrated by Figures 4.2 and 4.3 cannot occur. In particular, we have that $|A_1''| = |A_j''|$ and $|A_j'| = |A_n'|$.

From the second observation, we deduce that for $i = 2, 3, \ldots, j-1$, $A_i = A_{i+j-1}$, and consequently, $a_i = a_{i+j-1}$.

From the third observation, we deduce that $A_1'' = A_j''$ and $A_j' = A_n'$.

From the fourth observation, we deduce that either $A_1 = A_j$ or $A_j = A_n$. If $A_1 = A_j$, then $a_1 = a_j$; if $A_j = A_n$, then $a_j = a_n$. In the first case, $a_1a_2 \cdots a_{j-1}a_ja_{j+1} \cdots a_{n-1}$ is a square in $\mathbf{c}$, contrary to our assumption. In the second case, $a_2a_3 \cdots a_ja_{j+1}a_{j+2} \cdots a_n$ is a square in $\mathbf{c}$, contrary to our assumption.

We conclude that $\mathbf{d}$ contains no 1-approximate square yy', as required. □

Theorem 7.9. *There is an infinite word over a* 6*-letter alphabet that avoids* 2*-approximate squares, and the* 2 *is best possible.*

Proof. Let $\mathbf{c}$ be any squarefree word over $\{0, 1, 2\}$, and consider the image under the morphism β defined by

$$\begin{array}{rcl} 0 & \rightarrow & 012345 \\ 1 & \rightarrow & 012453 \\ 2 & \rightarrow & 012345 \end{array}$$

The resulting word avoids 2-approximate squares. The result is best possible, from Table 7.2.

The proof is similar to that of Theorem 7.8, so we only note the properties of the morphism β needed to derived the result:

1. Let $a, b, c \in \{0, 1, 2\}$, $a \neq b$. Let u be any subword of length 6 of $\beta(ab)$. If u is neither a prefix nor a suffix of $\beta(ab)$, then $\beta(c)$ and u mismatch in at least 3 positions.

2. Let $a, b \in \{0, 1, 2\}$, $a \neq b$. Then $\beta(a)$ and $\beta(b)$ mismatch in at least 3 positions.

3. Let u, u', v, v' be words satisfying the following:

   - $|u| = |u'|$, $|v| = |v'|$, and $|uv| = |u'v'| = 6$;
   - each of u and u' is a suffix of a word in $\{\beta(0), \beta(1), \beta(2)\}$; and
   - each of v and v' is a prefix of a word in $\{\beta(0), \beta(1), \beta(2)\}$.

   Then either $uv = u'v'$ or uv and $u'v'$ mismatch in at least 3 positions.

4. Let $a \in \{0, 1, 2\}$. Then $\beta(a)$ is uniquely determined by either its prefix of length 4 or its suffix of length 1.

□

Results over larger alphabets are summarized below.

Theorem 7.10. *For every k, n, and d given in Table 7.3, there exists an infinite word over a k-letter alphabet that avoids n-approximate squares. In each case such a word can be obtained by applying the d-uniform morphism given in the table to any infinite squarefree word over* $\{0, 1, 2\}$.

In each case the proof is similar to that of Theorem 7.9 and is omitted.

k	n	d	Morphism
7	3	14	$0 \rightarrow 01234056132465$ $1 \rightarrow 01234065214356$ $2 \rightarrow 01234510624356$
8	4	16	$0 \rightarrow 0123456071326547$ $1 \rightarrow 0123456072154367$ $2 \rightarrow 0123456710324765$
9	5	36	$0 \rightarrow 012345607821345062718345670281346578$ $1 \rightarrow 012345607182346750812347685102346578$ $2 \rightarrow 012345607182346510872345681702346578$
11	6	20	$0 \rightarrow 012345670A812954768A$ $1 \rightarrow 0123456709A1843576A9$ $2 \rightarrow 01234567089A24365798$
12	7	24	$0 \rightarrow 012345678091AB2354687A9B$ $1 \rightarrow 012345678091A3B4257689AB$ $2 \rightarrow 012345678091A2B3465798AB$
13	8	26	$0 \rightarrow 01234567890A1BC24635798BAC$ $1 \rightarrow 01234567890A1B3C4257689ABC$ $2 \rightarrow 01234567890A1B2C354687A9BC$
14	9	28	$0 \rightarrow 0123456789A0B1DC32465798BDAC$ $1 \rightarrow 0123456789A0B1DC243576A98DBC$ $2 \rightarrow 0123456789A0B1CD325468A79CBD$
15	10	30	$0 \rightarrow 0123456789AB0D1CE3246579B8ACDE$ $1 \rightarrow 0123456789AB0D1CE2435768A9DCBE$ $2 \rightarrow 0123456789AB0CED32154687BA9DEC$

Table 7.3: Avoiding n-approximate squares over a k-letter alphabet

It is possible to give a general construction as follows.

Theorem 7.11. *For all integers $n \geq 3$, there is an infinite word over an alphabet of $2n$ letters that avoid $(n-1)$-approximate squares.*

Proof. We define the $2n$-uniform morphism $h : \{0,1,2\}^* \rightarrow \{0,1,\ldots,2n-1\}^*$ as follows:

$$\begin{array}{rcl} 0 & \rightarrow & 012\cdots(n-1)n\cdots(2n-1) \\ 1 & \rightarrow & 012\cdots(n-1)(n+1)(n+2)\cdots(2n-1)n \\ 2 & \rightarrow & 012\cdots(n-1)(n+2)(n+3)\cdots(2n-1)n(n+1). \end{array}$$

We claim that if $\mathbf{w}$ is any infinite squarefree word over $\{0,1,2\}$, then $h(\mathbf{w})$ has the desired properties. The proof is a straightforward generalization of Theorem 7.9. □

This concludes our work on avoiding approximate squares. In the next (and final) chapter we summarize the contents of the entire thesis and present or recall some open problems.

Chapter 8

Conclusion and Open Problems

We have studied a variety of generalizations of Thue's results concerning the avoidance of repetitions in infinite words. We have also presented some applications of this theory to problems in Diophantine approximation and formal language theory.

We began with an historical survey of the area of combinatorics on words in Chapter 1. In Chapter 2 we reviewed some basic results concerning the Thue–Morse word and overlap-free words in general. One open problem noted in Section 2.5 is:

Problem (pg. 20)**.** *Determine the minimal forbidden subwords of the generalized Thue–Morse words.*

In Chapter 2 we also gave a simplified exposition of Fife's characterization of the infinite overlap-free binary words. Additionally, we gave an application of the theory of overlap-free words to a problem in transcendental number theory. There are many related open problems in transcendental number theory; we present two such problems below.

Problem 8.1. *Prove that the binary expansion of any algebraic number contains arbitrarily large squares.*

Problem 8.2. *Prove that the binary expansion of any algebraic number contains arbitrarily large palindromes.*

With regards to the Fife theory for the infinite overlap-free words, it seems likely that a similar theory can be developed for the infinite 7/3-power-free words. We thus ask the following.

Problem (pg. 28)**.** *Does there exist a characterization similar to that of Theorem 2.14 of the infinite* $7/3$*-power-free binary words?*

In Chapters 3 and 4 we presented additional results concerning overlap-free and 7/3-power-free words. We generalized a result of Séébold by showing that the Thue–Morse word and its complement are the only infinite 7/3-power-free binary words that can be generated by iteration of a morphism. We characterized the squares that can appear as subwords of an infinite overlap-free binary word as well as the 7/3-power-free binary squares. We also constructed infinite 7/3-power-free binary words containing infinitely many overlaps.

In Chapter 5 we considered some problems in formal language theory that derive from the theory of combinatorics on words. We showed that set of ternary words containing overlaps is not context-free; we also showed that the set of binary words containing overlaps is not unambiguously context-free. The obvious open problem is the following.

Problem 8.3. *Is the language consisting of all binary words that contain an overlap context-free?*

We also mention the following well-known and long standing open problem.

Problem 8.4. *Prove or disprove that the set of* primitive words *(i.e., the set of binary words w such that w cannot be written as x^k for some $k > 1$) is not context-free.*

We may also consider various questions concerning the language of words that do not occur as subwords of a particular infinite word. Two such questions are presented below.

Problem (pg. 59)**.** *Let $\mathbf{t}$ be the Thue–Morse word. Is Cosub($\mathbf{t}$) context-free?*

Problem (pg. 59)**.** *Let $\mathbf{f}$ be the Fibonacci word: i.e., the word generated by iterating the morphism $0 \to 01$, $1 \to 0$. Is Cosub($\mathbf{f}$) context-free?*

In Chapter 6 we studied the problem of avoiding repetitions in arithmetic progressions. We presented some examples of words avoiding squares (cubes, overlaps, etc.) in all arithmetic progressions of odd difference. We discovered interesting connections to the well-studied paperfolding words as well as to certain classical results from Ramsey theory, such as van der Waerden's theorem. We list a couple of intriguing open problems below.

Problem 8.5. *Determine if the alphabet size of* 16 *in Theorem 6.18 is best possible.*

Problem 8.6. *Determine exactly the* Thue threshold *of $\mathbb{N}^2$, i.e., the smallest integer t such that there exists an integer $k \geq 2$ and a 2-dimensional word $\mathbf{w}$ over a t-letter alphabet such that every line of $\mathbf{w}$ is k-power-free.*

By the results of Section 6.11, the possible values for t in the previous problem are $t = 2$, 3, or 4.

In Chapter 7 we considered avoiding approximate repetitions rather than exact repetitions and constructed some interesting infinite words avoiding certain types of approximate repetition.

This final chapter concludes the work. Needless to say, there are many interesting directions for future research. Moreover, this work has only been able to touch on a relatively small number of topics in the area of combinatorics on words. There are many other deep and interesting concepts and results in this area that we have not been able to discuss here.

Appendix A

Prouhet's Memoir

In 1851, Prouhet submitted a memoir to the French Academy in which he described a rule for partitioning the first n^m positive integers into n groups of n^{m-1} integers such that for all positive integers $k < m$, the sums of the k-th powers of the elements of a group is the same for each group.

It seems that the full memoir was never published, but an "extract by the author" was published in the *Comptes Rendus* of the French Academy of Sciences [200]. This extract describes the construction of sequences that we now refer to as either *generalized Thue–Morse sequences* or as *Prouhet sequences*. These sequences encode the desired partion of the integers. However, the short extract of Prouhet does not contain a proof of the claimed result.

Prouhet's result went unnoticed for many years. Several weaker results were proved in the mean time, in particular by Tarry and Escott. Indeed this problem is generally referred to as the *Tarry–Escott problem*. Dickson gives a catalogue of results in this area in his *History of the Theory of Numbers* [80].

It was not until 1947 that Lehmer [156] independently rediscovered Prouhet's result (along with a complete proof). Wright [241] seems to have been the first person to draw attention to the earlier work of Prouhet; he also gave his own proof of the result.

The sequences of Prouhet can also be used to construct so-called *magic squares* (or in higher dimensions, *magic cubes*), as described by Adler and Li [11].

We have transcribed the extract of Prouhet's memoir below. A translation into English follows.

THÉORIE DES NOMBRES. — *Mémoire sur quelques relations entre les puissances des nombres; par* M. E. PROUHET. (Extrait par l'auteur.)

(Commissaires, MM. Sturm, Lamé, Binet.)

n et m étant deux nombres entiers quelconques, il existe une infinité de suites de n^m nombres, susceptibles de se partager en n groupes de n^{m-1} termes chacun et tels que la somme des puissances k des termes soit la mêmes pour tous les groupes, k étant un nombre entier inférieur à m.

n^m nombres en progression arithmétique jouissent de la propriété précédente. Pour opérer le partage de ces nombres en groupes, on écrira en cercle les indices $0, 1, 2, \ldots, n-1$; on lira ces indices en suivant le cercle et en ayant soin d'en passer un à chaque tour; deux, tous les n tours; trois, tous les n^2 tours, et ainsi de suite. Ces indices, écrits à mesure qu'on les lit sous termes de la progression, apprendront à quel groupe appartient chaque terme.

Si l'on applique la règle et le théorème précédents aux 27 premiers nombres de la suite naturelle, on arrive aux identités suivantes:

$$\begin{aligned} & 1+6+8+12+14+16+20+22+27 \\ = \; & 2+4+9+10+15+17+21+23+25 \\ = \; & 3+5+7+11+13+18+19+24+26 \end{aligned}$$

$$1^2+6^2+8^2+\cdots = 2^2+4^2+9^2+\cdots = 3^2+5^2+7^2+\cdots$$

Lorsque $n = 10$ et que la progression commence à 0, tous les nombres dont la somme des chiffres, divisée par 10, laisse le mème reste, appartiennent à la mème classe.

NUMBER THEORY. — *Memoir concerning certain relations among powers of numbers; by* M. E. PROUHET. (Extract by the author.)

(Commissioners, Messrs. Sturm, Lamé, Binet.)

Let n and m be any two whole numbers. Then there exists infinitely many sequences of n^m numbers that can be partitioned into n groups of n^{m-1} terms each such that the sum of the k-th powers of the terms is the same for all the groups, where k is a whole number less than m.

n^m numbers in arithmetic progression enjoy the preceeding property. To carry out the partition of these numbers into groups, we write on a circle the indices $0, 1, 2, \ldots, n-1$; we read these indices following the circle and taking care to skip one of them at each turn; two, every n turns; three, every n^2 turns, and so on. These indices, written as one reads them under terms of the progression, tells us to which group each term belongs.

If we apply the preceeding rule and theorem to the first 27 natural numbers, we obtain the following identities:

$$\begin{aligned} & 1+6+8+12+14+16+20+22+27 \\ = \;& 2+4+9+10+15+17+21+23+25 \\ = \;& 3+5+7+11+13+18+19+24+26 \end{aligned}$$

$$1^2+6^2+8^2+\cdots = 2^2+4^2+9^2+\cdots = 3^2+5^2+7^2+\cdots$$

When $n = 10$ and the progression begins with 0, all the numbers for which the sum of their digits, when divided by 10, leaves the same remainder, belong to the same class.

List of References

[1] A. Aberkane and J. Currie. Attainable lengths for circular binary words avoiding k powers. *Bull. Belgian Math. Soc.* **12** (2005), 525–534.

[2] A. Aberkane, J. Currie, and N. Rampersad. The number of ternary words avoiding abelian cubes grows exponentially. *J. Integer Sequences* **7** (2004), Article 04.2.7. Available electronically at `http://www.math.uwaterloo.ca/JIS/VOL7/Currie/currie18.html`.

[3] A. Aberkane, V. Linek, and S. J. Mor. On the powers in the Thue–Morse word. *Australasian J. Combinat.* **35** (2006), 41–49.

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

[5] B. Adamczewski. Symbolic discrepancy and self-similar dynamics. *Ann. Inst. Fourier (Grenoble)* **54** (2004), 2201–2234.

[6] B. Adamczewski and J.-P. Allouche. Reversals and palindromes in continued fractions. *Theoret. Comput. Sci.* **380** (2007), 220–237.

[7] B. Adamczewski and Y. Bugeaud. On the complexity of algebraic numbers I. Expansions in integer bases. *Ann. Math.* **165** (2007), 547–565.

[8] B. Adamczewski, Y. Bugeaud, and F. Luca. Sur la complexité des nombres algébriques. *C. R. Acad. Sci. Paris* **1339** (2004), 11–14.

[9] B. Adamczewski and N. Rampersad. On patterns occurring in binary algebraic numbers. To appear in *Proc. Amer. Math. Soc.*, 2007.

[10] S. I. Adjan. *The Burnside Problem and Identities in Groups.* Springer-Verlag, 1979.

[11] A. Adler and S.-Y. Li. Magic cubes and Prouhet sequences. *Amer. Math. Monthly* **84** (1977), 618–627.

[12] F. D'Alessandro, B. Intrigila, and S. Varricchio. On the structure of the counting function of sparse context-free languages. *Theoret. Comput. Sci.* **356** (2006), 104–117.

[13] J.-P. Allouche. Suites infinies a répétitions bornées. *Séminaire de Théorie des Nombres de Bordeaux* (1983–1984), Exposé 20.

[14] J.-P. Allouche. The number of factors in a paperfolding sequence. *Bull. Austral. Math. Soc.* **46** (1992), 23–32.

[15] J.-P. Allouche. Transcendence of formal power series with rational coefficients. *Theoret. Comput. Sci.* **218** (1999), 143–160.

[16] J.-P. Allouche and M. Bousquet-Mélou. Facteurs des suites de Rudin-Shapiro généralisées. *Bull. Belgian Math. Soc.* **1** (1994), 145–164.

[17] J.-P. Allouche and M. Bousquet-Mélou. Canonical positions for the factors in paperfolding sequences. *Theoret. Comput. Sci.* **129** (1994), 263–278.

[18] J.-P. Allouche and M. Cosnard. The Komornik–Loreti constant is transcendental. *Amer. Math. Monthly* **107** (2000), 448–449.

[19] J.-P. Allouche, J. Currie, and J. Shallit. Extremal infinite overlap-free binary words. *Electronic J. Combinatorics* **5** (1998), #R27. Available electronically at `http://www.combinatorics.org/Volume_5/Abstracts/v5i1r27.html`.

[20] J.-P. Allouche, C. Frougny, and K. G. Hare. On univoque Pisot numbers. *Math. Comp.* **259** (2007), 1639–1660.

[21] J.-P. Allouche and J. Shallit. The ubiquitous Prouhet–Thue–Morse sequence. In *Sequences and their Applications: Proceedings of SETA 98*, pp. 1–16. Springer-Verlag, 1999.

[22] J.-P. Allouche and J. Shallit. Sums of digits, overlaps, and palindromes. *Discrete Math. & Theoret. Comput. Sci.* **4** (2000), 001–010.

[23] J.-P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations.* Cambridge University Press, 2003.

[24] J.-P. Allouche and L. Zamboni. Algebraic irrational binary numbers cannot be fixed points of non-trivial constant length or primitive morphisms. *J. Number Theory* **69** (1998), 119–124.

[25] N. Alon, J. Grytczuk, M. Haluszczak, and O. Riordan. Nonrepetitive colorings of graphs. *Random Structures and Algorithms* **21** (2002), 336–346.

[26] S. E. Aršon. Proof of the existence of asymmetric infinite sequences (Russian). *Mat. Sbornik* **2** (1937), 769–779.

[27] J.-M. Autebert, J. Beauquier, L. Boasson, and M. Nivat. Quelques problèmes ouverts en théorie des langages algébriques. *RAIRO Inform. Théor. App.* **13** (1979), 363–378.

[28] J.-M. Autebert, P. Flajolet, and J. Gabarro. Prefixes of infinite words and ambiguous context-free languages. *Inform. Process. Lett.* **25** (1987), 211–216.

[29] S. V. Avgustinovich. The number of distinct subwords of fixed length in the Morse–Hedlund sequence. *Sibirsk. Zh. Issled. Oper.* **1** (1994), 3–7.

[30] S. V. Avgustinovich, D. G. Fon-Der-Flaass, and A. E. Frid. Arithmetical complexity of infinite words. In *Words, Languages & Combinatorics III*, pp. 51–62. World Scientific, 2003.

[31] D. Bean, A. Ehrenfeucht, and G. McNulty. Avoidable patterns in strings of symbols. *Pacific J. Math.* **85** (1979), 261–294.

[32] J. Beck. An application of Lovász local lemma: there exists an infinite 01-sequence containing no near identical intervals. In A. Hajnal et al., editors, *Finite and Infinite Sets, Vol. I, II (Eger, 1981)*, Vol. 37 of *Colloq. Math. Soc. János Bolyai*, pp. 103–107. North-Holland, 1984.

[33] P.-G. Becker. k-Regular power series and Mahler-type functional equations. *J. Number Theory* **49** (1994), 269–286.

[34] J. Berstel. Sur les mots sans carré définis par un morphisme. In *Proc. 6th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 71 of *Lecture Notes in Computer Science*, pp. 16–25. Springer-Verlag, 1979.

[35] J. Berstel. Sur la construction de mots sans carré. *Séminaire de Théorie des Nombres de Bordeaux* (1978–1979), Exposé 18.

[36] J. Berstel. Some recent results on square-free words. In *STACS 84, Proc. 1st Symp. Theoretical Aspects of Comp. Sci.*, Vol. 166 of *Lecture Notes in Computer Science*, pp. 14–25. Springer-Verlag, 1984.

[37] J. Berstel. Every iterated morphism yields a co-CFL. *Inform. Process. Lett.* **22** (1986), 7–9.

[38] J. Berstel. Axel Thue's work on repetitions in words. In P. Leroux and C. Reutenauer, editors, *Séries formelles et combinatoire algébrique*, Vol. 11 of *Publications du Laboratoire de Combinatoire et d'Informatique Mathématique*, pp. 65–80. Université du Québec à Montréal, 1992.

[39] J. Berstel. A rewriting of Fife's theorem about overlap-free words. In J. Karhumäki, H. Maurer, and G. Rozenberg, editors, *Results and Trends in Theoretical Computer Science*, Vol. 812 of *Lecture Notes in Computer Science*, pp. 19–29. Springer-Verlag, 1994.

[40] J. Berstel. Growth of repetition-free words—a review. *Theoret. Comput. Sci.* **340** (2005), 280–290.

[41] J. Berstel and J. Karhumäki. Combinatorics on words. TUCS Technical Report 530, Turku Centre for Computer Science, June 2003.

[42] J. Berstel and D. Perrin. The origins of combinatorics on words. *European J. Combinatorics* **28** (2007), 996–1022.

[43] J. Berstel and P. Séébold. A characterization of overlap-free morphisms. *Disc. Appl. Math.* **46** (1993), 275–281.

[44] F. Blanchard. β-expansions and symbolic dynamics. *Theoret. Comput. Sci.* **65** (1989), 131–141.

[45] A. Blanchard and M. Mendès France. Symétrie et transcendance. *Bull. Sci. Math.* **106** (1982), 325–335.

[46] A. Blondin-Massé, S. Brlek, A. Glen, and S. Labbé. On the critical exponent of generalized Thue–Morse sequences. To appear in *Discrete Math. & Theoret. Comput. Sci.*, 2007.

[47] É. Borel. Sur les chiffres décimaux de $\sqrt{2}$ et divers problèmes de probabilités en chaîne. *C. R. Acad. Sci. Paris* **230** (1950), 591–593.

[48] F.-J. Brandenburg. Uniformly growing kth power-free homomorphisms. *Theoret. Comput. Sci.* **23** (1983), 69–82.

[49] M. Bridson and R. Gilman. Context-free languages of sub-exponential growth. *J. Comput. System Sci.* **64** (2002), 308–310.

[50] J. Brinkhuis. Nonrepetitive sequences on three symbols. *Quart. J. Math. Oxford Ser. (2)* **34** (1983), 145–149.

[51] S. Brlek. Enumeration of factors in the Thue-Morse word. *Disc. Appl. Math.* **24** (1989), 83–96.

[52] S. Brown, N. Rampersad, J. Shallit, and T. Vasiga. Squares and overlaps in the Thue–Morse sequence and some variants. *RAIRO Inform. Théor. App.* **40** (2006), 473–484.

[53] W. Burnside. On an unsettled question in the theory of discontinuous groups. *Quart. J. Pure Appl. Math.* **33** (1902), 230–238.

[54] F. Carlson. Über Potenzreihen mit ganzzahligen Koeffizientem. *Math. Zeitschrift* **9** (1921), 1–13.

[55] A. Carpi. Multidimensional unrepetitive configurations. *Theoret. Comput. Sci.* **56** (1988), 233–241.

[56] A. Carpi. Overlap-free words and finite automata. *Theoret. Comput. Sci.* **115** (1993), 243–260.

[57] A. Carpi. On the number of abelian square-free words on four letters. *Disc. Appl. Math.* **81** (1998), 155–167.

[58] A. Carpi. On Dejean's conjecture over large alphabets. *Theoret. Comput. Sci.* **385** (2007), 137–151.

[59] J. Cassaigne. Unavoidable binary patterns. *Acta Informatica* **30** (1993), 385–395.

[60] J. Cassaigne. Counting overlap-free binary words. In *STACS 93, Proc. 10th Symp. Theoretical Aspects of Comp. Sci.*, Vol. 665 of *Lecture Notes in Computer Science*, pp. 216–225. Springer-Verlag, 1993.

[61] D. P. Chi and D. Y. Kwon. Sturmian words, β-shifts, and transcendence. *Theoret. Comput. Sci.* **321** (2004), 395–404.

[62] C. Choffrut and J. Karhumäki. Combinatorics of words. In *Handbook of formal languages*, Vol. 1, chapter 6, pp. 329–438. Springer-Verlag, 1997.

[63] N. Chomsky and M.-P. Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Hirschbert, editors, *Computer Programming and Formal Systems*, pp. 118–161. North-Holland, 1963.

[64] G. Christol. Ensembles presques périodiques k-reconnaissables. *Theoret. Comput. Sci.* **9** (1979), 141–145.

[65] G. Christol, T. Kamae, M. Mendès France, and G. Rauzy. Suites algébriques, automates et substitutions. *Bull. Soc. Math. France* **108** (1980), 410–419.

[66] A. Cobham. On the base-dependence of sets of numbers recognizable by finite automata. *Math. Systems Theory* **3** (1969), 186–192.

[67] A. Cobham. Uniform tag sequences. *Math. Systems Theory* **6** (1972), 164–192.

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

[69] J. Currie. There are ternary circular square-free words of length n for $n \geq 18$. *Electronic J. Combinatorics* **9** (2002), #N10. Available electronically at `http://www.combinatorics.org/Volume_9/Abstracts/v9i1n10.html`.

[70] J. Currie. The number of binary words avoiding abelian fourth powers grows exponentially. *Theoret. Comput. Sci.* **319** (2004), 441–446.

[71] J. Currie. Pattern avoidance: themes and variations. *Theoret. Comput. Sci.* **339** (2005), 7–18.

[72] J. Currie, N. Rampersad, and J. Shallit. Binary words containing infinitely many overlaps. *Electronic J. Combinatorics* **13** (2006), #R82. Available electronically at `http://www.combinatorics.org/Volume_13/Abstracts/v13i1r82.html`.

[73] J. Currie and J. Simpson. Non-repetitive tilings. *Electronic J. Combinatorics* **9** (2002), #R28. Available electronically at `http://www.combinatorics.org/Volume_9/Abstracts/v9i1r28.html`.

[74] C. Davis and D. E. Knuth. Number representations and dragon curves—I,II. *J. Recreational Math.* **3** (1970), 66–81, 133–149.

[75] F. Dejean. Sur un théorème de Thue. *J. Combin. Theory. Ser. A* **13** (1972), 90–99.

[76] F. M. Dekking. On repetitions of blocks in binary sequences. *J. Combin. Theory. Ser. A* **20** (1976), 292–299.

[77] F. M. Dekking. Transcendance du nombre du Thue–Morse. *C. R. Acad. Sci. Paris* **285** (1977), 157–160.

[78] F. M. Dekking. Strongly non-repetitive sequences and progression-free sets. *J. Combin. Theory. Ser. A* **27** (1979), 181–185.

[79] M. Dekking, M. Mendès France, and A. van der Poorten. FOLDS! *Math. Intelligencer* **4** (1982), 130–138; 173–181; 190–195.

[80] L. E. Dickson. *History of the Theory of Numbers*, Vol. 2. Washington, 1920.

[81] T. Downarowicz. Reading along arithmetic progressions. *Colloq. Math.* **80** (1999), 293–296.

[82] A. Dumitrescu and R. Radoičić. On a coloring problem for the integer grid. In *Towards a theory of geometric graphs*, Vol. 342 of *Contemp. Math.*, pp. 67–74. Amer. Math. Soc., 2004.

[83] A. Ehrenfeucht, K. P. Lee, and G. Rozenberg. Subword complexities of various classes of deterministic developmental languages without interaction. *Theoret. Comput. Sci.* **1** (1975), 59–75.

[84] A. Ehrenfeucht and G. Rozenberg. On the separating power of EOL systems. *RAIRO Inform. Théor. App.* **17** (1982), 13–22.

[85] A. Ehrenfeucht and G. Rozenberg. Repetitions of subwords in D0L languages. *Inform. Comput.* **59** (1983), 13–35.

[86] S. Ekhad and D. Zeilberger. There are more than $2^{n/17}$ n-letter ternary square-free words. *J. Integer Sequences* **1** (1998), Article 98.1.9. Available electronically at `http://www.math.uwaterloo.ca/JIS/zeil.html`.

[87] R. C. Entringer, D. E. Jackson, and J. A. Schatz. On nonrepetitive sequences. *J. Combin. Theory. Ser. A* **16** (1974), 159–164.

[88] P. Erdős. Some unsolved problems. *Magyar Tud. Akad. Kutató Int. Közl.* **6** (1961), 221–254.

[89] P. Erdős, I. Joó, and V. Komornik. Characterization of the unique expansions $1 = \sum_{i=1}^{\infty} q^{-n_i}$ and related problems. *Bull. Soc. Math. France* **118** (1990), 377–390.

[90] P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In A. Hajnal et al., editors, *Infinite and Finite Sets, Vol. II*, Vol. 10 of *Colloq. Math. Soc. János Bolyai*, pp. 609–628. North-Holland, 1975.

[91] M. Euwe. Mengentheoretische Betrachtungen über das Schachspiel. *Proc. Konin. Akad. Wetenschappen Amsterdam* **32** (1929), 633–642.

[92] A. A. Evdokimov. Strongly asymmetric sequences generated by a finite number of symbols. *Soviet Math. Dokl.* **9** (1968), 536–539.

[93] G. Everest, A. van der Poorten, I. Shparlinski, and T. Ward. *Recurrence sequences.* Amer. Math. Soc., 2003.

[94] P. Fatou. Séries trigonométriques et séries de Taylor. *Acta Math.* **30** (1906), 335–400.

[95] S. Ferenczi and C. Mauduit. Transcendence of numbers with a low complexity expansion. *J. Number Theory* **67** (1997), 146–161.

[96] E. Fife. Binary sequences which contain no *BBb*. *Trans. Amer. Math. Soc.* **261** (1980), 115–136.

[97] P. Flajolet. Analytic models and ambiguity of context-free languages. *Theoret. Comput. Sci.* **49** (1987), 283–309.

[98] A. S. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electronic J. Combinatorics* **2** (1995), #R2. Available electronically at `http://www.combinatorics.org/Volume_2/Abstracts/v2i1r2.html`.

[99] A. E. Frid. The subword complexity of fixed points of binary uniform morphisms. In *Fundamentals of Computation Theory: FCT '97*, Vol. 1279 of *Lecture Notes in Computer Science*, pp. 178–187. Springer-Verlag, 1997.

[100] A. E. Frid. On the combinatorial complexity of iteratively generated symbol sequences. *Diskretn. Anal. Issled. Oper. Ser. 1* **4** (1997), 53–59. English translation in *Discrete Appl. Math.* **114** (2001), 115–120.

[101] A. E. Frid. On uniform D0L words. In *STACS 98, Proc. 15th Symp. Theoretical Aspects of Comp. Sci.*, Vol. 1373 of *Lecture Notes in Computer Science*, pp. 544–554. Springer-Verlag, 1998.

[102] A. E. Frid. Overlap-free symmetric D0L words. *Discrete Math. & Theoret. Comput. Sci.* **4** (2001), 357–362.

[103] A. E. Frid, 2007. Personal communication.

[104] J. Gabarró. Some applications of the interchange lemma. *Bull. European Assoc. Theor. Comput. Sci.* **25** (1985), 19–21.

[105] S. Ginsburg. *The Mathematical Theory of Context-free Languages*. McGraw-Hill, 1966.

[106] W. Gottschalk and G. Hedlund. *Topological Dynamics*, Vol. 36. Am. Math. Soc. Colloq. Publ., 1955.

[107] W. Gottschalk and G. Hedlund. A characterization of the Morse minimal set. *Proc. Amer. Math. Soc.* **15** (1964), 70–74.

[108] R. Graham, B. Rothschild, and J. Spencer. *Ramsey Theory*. Wiley, second edition edition, 1990.

[109] A. Grazon. An infinite word language which is not co-CFL. *Inform. Process. Lett.* **24** (1987), 81–85.

[110] U. Grimm. Improved bounds on the number of ternary square-free words. *J. Integer Sequences* **4** (2001), Article 01.2.7. Available electronically at `http://www.math.uwaterloo.ca/JIS/VOL4/GRIMM/words.html`.

[111] J. Grytczuk. Thue-like sequences and rainbow arithmetic progressions. *Electronic J. Combinatorics* **9** (2002), #R44. Available electronically at `http://www.combinatorics.org/Volume_9/Abstracts/v9i1r44.html`.

[112] J. Grytczuk. Thue type problems for graphs, points, and numbers. Manuscript, 2006.

[113] M. Haiman. Non-commutative rational power series and algebraic generating functions. *European J. Combinatorics* **14** (1993), 335–339.

[114] A. W. Hales and R. I. Jewett. Regularity and positional games. *Trans. Amer. Math. Soc.* **106** (1963), 222–229.

[115] M. Hall. Generators and relations in groups—the Burnside problem. In T. L. Saaty, editor, *Lectures on Modern Mathematics*, Vol. 2, chapter 2, pp. 42–92. Wiley, 1964.

[116] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, fifth edition, 1979.

[117] T. Harju. On cyclically overlap-free words in binary alphabets. In *The Book of L*, pp. 125–130. Springer-Verlag, 1986.

[118] T. Harju and M. Linna. On the peridicity of morphisms on free monoids. *RAIRO Inform. Théor. App.* **20** (1986), 47–54.

[119] T. Harju and D. Nowotka. Border correlation of binary words. *J. Combin. Theory. Ser. A* **108** (2004), 331–341.

[120] T. Harju and D. Nowotka. Binary words with few squares. *Bull. European Assoc. Theor. Comput. Sci.* **89** (2006), 164–166.

[121] E. Hille. *Analytic Function Theory, Vol. II.* Ginn, 1962.

[122] J. Honkala. On Parikh slender languages and power series. *J. Comput. System Sci.* **52** (1996), 185–190.

[123] J. Honkala. Cancellation and periodicity properties of iterated morphisms. To appear in *Theoret. Comput. Sci.*, 2007.

[124] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.

[125] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[126] O. Ibarra and B. Ravikumar. On sparseness, ambiguity and other decision problems for acceptors and transducers. In *STACS 86, Proc. 3rd Symp. Theoretical Aspects of Comp. Sci.*, Vol. 210 of *Lecture Notes in Computer Science*, pp. 171–179. Springer-Verlag, 1986.

[127] L. Ilie. On subwords of infinite words. *Disc. Appl. Math.* **63** (1995), 277–279.

[128] L. Ilie, P. Ochem, and J. Shallit. A generalization of repetition threshold. *Theoret. Comput. Sci.* **345** (2005), 359–369.

[129] R. Incitti. The growth function of context-free languages. *Theoret. Comput. Sci.* **255** (2001), 601–605.

[130] S. Istrail. On irreducible languages and nonrational numbers. *Bull. Math. Soc. Sci. Math. R. S. Roumanie (N.S.)* **21** (1977), 301–308.

[131] G. Istrate. Some remarks on almost periodic sequences and languages. In G. Păun, editor, *Mathematical Linguistics and Related Topics*, pp. 191–194. Romanian Academy of Sciences, 1994.

[132] K. Jacobs and M. Keane. 0–1-sequences of Toeplitz type. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **13** (1969), 123–131.

[133] J. Justin. Characterization of the repetitive semigroups. *J. Algebra* **21** (1972), 87–90.

[134] J. Justin. Généralization du théorème de van der Waerden sur les semigroupes répétitifs. *J. Combin. Theory. Ser. A* **12** (1972), 357–367.

[135] J.-Y. Kao, N. Rampersad, J. Shallit, and M. Silva. Words avoiding repetitions in arithmetic progressions. To appear in *Theoret. Comput. Sci.*

[136] J. Karhumäki. On cube free ω-words generated by binary morphisms. *Disc. Appl. Math.* **5** (1983), 279–297.

[137] J. Karhumäki and J. Shallit. Polynomial versus exponential growth in repetition-free binary words. *J. Combin. Theory. Ser. A* **105** (2004), 335–347.

[138] V. Keränen. Abelian squares are avoidable on 4 letters. In *Proc. 19th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 623 of *Lecture Notes in Computer Science*, pp. 41–52. Springer-Verlag, 1992.

[139] A.-J. Kfoury. A linear-time algorithm to decided whether a binary word contains an overlap. *RAIRO Inform. Théor. App.* **22** (1988), 135–145.

[140] A. Khalyavin. The minimal density of a letter in an infinite ternary square-free sord is 883/3215. *J. Integer Sequences* **10** (2007), Article 07.6.5. Available electronically at `http://www.cs.uwaterloo.ca/journals/JIS/VOL10/Khalyavin/khalyavin13.html`.

[141] Y. Kobayashi. Enumeration of irreducible binary words. *Disc. Appl. Math.* **20** (1988), 221–232.

[142] Y. Kobayashi and F. Otto. Repetitiveness of languages generated by morphisms. *Theoret. Comput. Sci.* **240** (2000), 337–378.

[143] Y. Kobayashi, F. Otto, and P. Séébold. A complete characterization of repetitive morphisms over the two-letter alphabet. In *Proc. 3rd Ann. Int'l. Conf. on Computing and Combinatorics*, Vol. 1276 of *Lecture Notes in Computer Science*, pp. 393–402. Springer-Verlag, 1997.

[144] R. Kolpakov. Efficient lower bounds on the number of repetition-free words. *J. Integer Sequences* **10** (2007), Article 07.3.2. Available electronically at `http://www.cs.uwaterloo.ca/journals/JIS/VOL10/Kolpakov/kolpakov42.html`.

[145] R. Kolpakov, G. Kucherov, and Y. Tarannikov. On repetition-free binary words of minimal density. *Theoret. Comput. Sci.* **218** (1999), 161–175.

[146] V. Komornik and P. Loreti. Unique developments in non-integer bases. *Amer. Math. Monthly* **105** (1998), 636–639.

[147] D. König. Sur les correspondances multivoques des ensembles. *Fundamenta Math.* **8** (1926), 114–134.

[148] D. König. Über eine Schlußweise aus dem Endlichen ins Unendliche. *Acta Lit. Sci. (Szeged)* **3** (1927), 121–130.

[149] D. Krieger. On critical exponents in fixed points of binary k-uniform morphisms. In *STACS 2006, Proc. 23rd Symp. Theoretical Aspects of Comp. Sci.*, Vol. 3884 of *Lecture Notes in Computer Science*, pp. 104–114. Springer-Verlag, 2006.

[150] D. Krieger. On critical exponents in fixed points of non-erasing morphisms. *Theoret. Comput. Sci.* **376** (2007), 70–88.

[151] D. Krieger, P. Ochem, N. Rampersad, and J. Shallit. Avoiding approximate squares. In *Proc. 11th International Conference, Developments in Language Theory 2007*, Vol. 4588 of *Lecture Notes in Computer Science*, pp. 278–289. Springer-Verlag, 2007.

[152] W. Kuich and A. Salomaa. *Semirings, Automata, Languages.* Springer-Verlag, 1986.

[153] A. Kündgen and M. Pelsmajer. Nonrepetitive colorings of graphs of bounded treewidth. Manuscript, 2006.

[154] M. Latteux and G. Thierrin. On bounded context-free languages. *Elektronische Informationsverarbeitung und Kybernetik* **20** (1984), 3–8.

[155] J. Leech. A problem on strings of beads. *Math. Gazette* **41** (1957), 277–278.

[156] D. H. Lehmer. The Tarry–Escott problem. *Scripta Math.* **13** (1947), 37–41.

[157] S. Lehr. A result about languages concerning paperfolding sequences. *Math. Systems Theory* **25** (1992), 309–313.

[158] A. Lepistö. A characterization of 2^+-free words over a binary alphabet. Master's thesis, University of Turku, 1995.

[159] M. Lothaire. *Combinatorics on Words*, Vol. 17 of *Encyclopedia of Mathematics.* Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, 1997.

[160] M. Lothaire. *Algebraic Combinatorics on Words*, Vol. 90 of *Encyclopedia of Mathematics.* Cambridge University Press, 2002.

[161] M. Lothaire. *Applied Combinatorics on Words*, Vol. 105 of *Encyclopedia of Mathematics.* Cambridge University Press, 2005.

[162] J. Loxton and A. van der Poorten. Arithmetic properties of the solutions of a class of functional equations. *J. Reine Angew. Math.* **330** (1982), 159–172.

[163] J. Loxton and A. van der Poorten. Arithmetic properties of automata: regular sequences. *J. Reine Angew. Math.* **392** (1988), 57–69.

[164] A. de Luca and L. Mione. On bispecial factors of the Thue–Morse word. *Inform. Process. Lett.* **49** (1994), 179–183.

[165] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue–Morse sequence and a problem in semigroups. *Theoret. Comput. Sci.* **63** (1989), 333–348.

[166] K. Mahler. Arithmetische Eigenschaften der Lösungen einer Klasse von Funktionalgleichungen. *Math. Annalen* **101** (1929), 342–366. Corrigendum, **103** (1930), 532.

[167] M. Main. Permutations are not context-free: an application of the interchange lemma. *Inform. Process. Lett.* **15** (1982), 68–71.

[168] M. Main. An infinite square-free co-CFL. *Inform. Process. Lett.* **20** (1985), 105–107.

[169] M. Main, W. Bucher, and D. Haussler. Applications of an infinite square-free co-CFL. *Theoret. Comput. Sci.* **49** (1987), 113–119.

[170] S. Marcus and G. Păun. Infinite (almost periodic) words, formal languages and dynamical systems. *Bull. European Assoc. Theor. Comput. Sci.* **54** (1994), 224–231.

[171] M. Mendès France. Principe de la symétrie perturbée. *Séminaire de Théorie des Nombres de Paris* (1979–1980), 77–98.

[172] M. Mendès France and A. van der Poorten. Arithmetic and analytic properties of paper folding sequences. *Bull. Austral. Math. Soc.* **24** (1981), 123–131.

[173] P. Michel. *Sur les ensembles minimaux engendrés par les substitutions de longueur non constante.* PhD thesis, Université de Rennes, 1975.

[174] P. Michel. Stricte ergodicité d'ensembles minimaux de substitution. In J.-P. Conze and M. S. Keane, editors, *Théorie Ergodique: Actes des Journées Ergodiques, Rennes 1973/1974*, Vol. 532 of *Lecture Notes in Mathematics*, pp. 189–201. Springer-Verlag, 1976.

[175] F. Mignosi, A. Restivo, and M. Sciortino. Words and forbidden factors. *Theoret. Comput. Sci.* **273** (2002), 99–117.

[176] F. Mignosi and P. Séébold. If a D0L language is k-power free then it is circular. In *Proc. 20th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 700 of *Lecture Notes in Computer Science*, pp. 507–518. Springer-Verlag, 1993.

[177] M. Mohammad-Noori and J. D. Currie. Dejean's conjecture and Sturmian words. *European J. Combinatorics* **28** (2007), 876–890.

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

[179] M. Morse and G. A. Hedlund. Unending chess, symbolic dynamics and a problem in semigroups. *Duke Math. J.* **11** (1944), 1–7.

[180] P. Morton and W. J. Mourant. Digit patterns and transcendental numbers. *J. Austral. Math. Soc. Ser. A* **51** (1991), 216–236.

[181] B. Mossé. Reconnaissabilité des substitutions et complexité des suites automatiques. *Bull. Soc. Math. France* **124** (1996), 329–346.

[182] J. Moulin-Ollagnier. Proof of Dejean's conjecture for alphabets with $5, 6, 7, 8, 9, 10$ and 11 letters. *Theoret. Comput. Sci.* **95** (1992), 187–205.

[183] K. Nishioka. *Mahler Functions and Transcendence*, Vol. 1631 of *Lecture Notes in Mathematics*. Springer-Verlag, 1996.

[184] P. S. Novikov. On periodic groups. *Dokl. Akad. Nauk SSSR* **127** (1959), 749–752.

[185] P. S. Novikov and S. I. Adjan. Infinite periodic groups, I. *Izv. Akad. Nauk SSSR Ser. Mat.* **32** (1968), 212–244.

[186] P. S. Novikov and S. I. Adjan. Infinite periodic groups, II. *Izv. Akad. Nauk SSSR Ser. Mat.* **32** (1968), 251–524.

[187] P. S. Novikov and S. I. Adjan. Infinite periodic groups, III. *Izv. Akad. Nauk SSSR Ser. Mat.* **32** (1968), 709–731.

[188] P. Ochem. A generator of morphisms for infinite words. *RAIRO Inform. Théor. App.* **40** (2006), 427–441.

[189] P. Ochem. Letter frequency in infinite repetition-free words. *Theoret. Comput. Sci.* **380** (2007), 388–392.

[190] W. Ogden, R. Ross, and K. Winklmann. An "interchange lemma" for context-free languages. *SIAM J. Comput.* **14** (1985), 410–415.

[191] A. Panholzer. Gröbner bases and the defining polynomial of a context-free grammar generating function. *J. Automata, Languages, and Combinatorics* **10** (2005), 79–97.

[192] J.-J. Pansiot. The Morse sequence and iterated morphisms. *Inform. Process. Lett.* **12** (1981), 68–70.

[193] J.-J. Pansiot. Complexité des facteurs des mot infinis engendrés par morphismes itérés. In *Proc. 11th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 172 of *Lecture Notes in Computer Science*, pp. 380–389. Springer-Verlag, 1984.

[194] J.-J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots. *Disc. Appl. Math.* **7** (1984), 297–311.

[195] J.-J. Pansiot. Decidability of periodicity for infinite words. *RAIRO Inform. Théor. App.* **20** (1986), 43–46.

[196] H. Petersen. On the language of primitive words. *Theoret. Comput. Sci.* **161** (1996), 141–156.

[197] P. A. B. Pleasants. Non-repetitive sequences. *Proc. Cambridge Phil. Soc.* **68** (1970), 267–274.

[198] G. Pólya and G. Szegő. *Aufgaben und Lehrsätze aus der Analysis. Band II.* Springer-Verlag, 1971. English translation as *Problems and Theorems in Analysis II*, Springer-Verlag, 1998.

[199] H. Prodinger and F. J. Urbanek. Infinite 0–1-sequences without long adjacent identical blocks. *Discrete Math.* **28** (1979), 277–289.

[200] M. E. Prouhet. Mémoire sur quelques relations entre les puissances des nombres. *C. R. Acad. Sci. Paris* **33** (1851), 225.

[201] M. Queffélec. *Substitution Dynamical Systems—Spectral Analysis*, Vol. 1294 of *Lecture Notes in Mathematics*. Springer-Verlag, 1987.

[202] N. Rampersad. Words avoiding 7/3-powers and the Thue–Morse morphism. *Internat. J. Found. Comp. Sci.* **16** (2005), 755–766.

[203] N. Rampersad. On the context-freeness of the set of words containing overlaps. *Inform. Process. Lett.* **102** (2007), 74–78.

[204] N. Rampersad, J. Shallit, and M.-w. Wang. Avoiding large squares in infinite binary words. *Theoret. Comput. Sci.* **339** (2005), 19–34.

[205] D. Raz. Length considerations in context-free languages. *Theoret. Comput. Sci.* **183** (1997), 21–32.

[206] A. Restivo and S. Salemi. Overlap-free words on two symbols. In M. Nivat and D. Perrin, editors, *Automata on Infinite Words*, Vol. 192 of *Lecture Notes in Computer Science*, pp. 198–206. Springer-Verlag, 1984.

[207] J. Ridout. Rational approximations to algebraic numbers. *Mathematika* **4** (1957), 125–131.

[208] R. Rivest. Abelian square-free dithering for iterated hash functions. Manuscript, 2005.

[209] R. Ross and K. Winklmann. Repetitive strings are not context-free. *RAIRO Inform. Théor. App.* **16** (1982), 191–199.

[210] K. Roth. Rational approximations to algebraic numbers. *Mathematika* **2** (1955), 1–20. Corrigendum p. 168.

[211] P. Roth. l-occurrences of avoidable patterns. In *STACS 91, Proc. 8th Symp. Theoretical Aspects of Comp. Sci.*, Vol. 480 of *Lecture Notes in Computer Science*, pp. 42–29. Springer-Verlag, 1991.

[212] P. Roth. Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Informatica* **29** (1992), 95–107.

[213] G. Rozenberg and A. Salomaa. *The Mathematical Theory of L Systems*, Vol. 90 of *Pure and Applied Mathematics.* Academic Press, 1980.

[214] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1986.

[215] K. Saari. On the frequency of letters in pure binary morphic sequences. In *Proc. 9th International Conference, Developments in Language Theory 2005*, Vol. 3572 of *Lecture Notes in Computer Science*, pp. 397–408. Springer-Verlag, 2005.

[216] K. Saari. On the frequency of letters in morphic sequences. In *Proceedings of CSR 2006*, Vol. 3967 of *Lecture Notes in Computer Science*, pp. 334–345. Springer-Verlag, 2006.

[217] K. Saari. Everywhere α-repetitive sequences and Sturmian words. In *Proceedings of CSR 2007*, Vol. 4649 of *Lecture Notes in Computer Science*, pp. 362–372. Springer-Verlag, 2007.

[218] H. Schlickewei. The p-adic Thue–Siegel–Roth–Schmidt theorem. *Arch. Math. (Basel)* **29** (1977), 267–270.

[219] W. Schmidt. *Diophantine Approximation*, Vol. 785 of *Lecture Notes in Mathematics.* Springer-Verlag, 1980.

[220] P. Séébold. Morphismes itérés, mot de Morse et mot de Fibonacci. *C. R. Acad. Sci. Paris* **295** (1982), 439–441.

[221] P. Séébold. Overlap-free sequences. In M. Nivat and D. Perrin, editors, *Automata on Infinite Words*, Vol. 192 of *Lecture Notes in Computer Science*, pp. 207–215. Springer-Verlag, 1984.

[222] P. Séébold. Sequences generated by infinitely iterated morphisms. *Disc. Appl. Math.* **11** (1985), 255–264.

[223] P. Séébold. About some overlap-free morphisms on a n-letter alphabet. *J. Automata, Languages, and Combinatorics* **7** (2002), 579–597.

[224] P. Séébold. On some generalizations of the Thue–Morse morphism. *Theoret. Comput. Sci.* **292** (2003), 283–298.

[225] J. Shallit. Simultaneous avoidance of large squares and fractional powers in infinite binary words. *Internat. J. Found. Comp. Sci.* **15** (2004), 317–327.

[226] R. Shelton and R. Soni. Chains and fixing blocks in irreducible binary sequences. *Discrete Math.* **54** (1985), 93–99.

[227] A. M. Shur. The structure of the set of cube-free $\mathbb{Z}$-words in a two-letter alphabet (Russian). *Izv. Ross. Akad. Nauk Ser. Mat.* **64** (2000), 201–224. English translation in *Izv. Math.* **64** (2000), 847–871.

[228] J. Spencer. Asymptotic lower bounds for Ramsey functions. *Discrete Math.* **20** (1977), 69–76.

[229] X. Sun. New lower bound on the number of ternary square-free words. *J. Integer Sequences* **6** (2003), Article 03.3.2. Available electronically at `http://www.math.uwaterloo.ca/JIS/VOL6/Sun/sun.html`.

[230] T. Tapsoba. Automates calculant la complexité de suites automatiques. *J. Théorie Nombres Bordeaux* **6** (1994), 127–134.

[231] Y. Tarannikov. The minimal density of a letter in an infinite ternary square-free word is 0.2746... *J. Integer Sequences* **5** (2002), Article 02.2.2.

[232] K. Thomsen. Languages of finite words occurring infinitely many times in an infinite word. *RAIRO Inform. Théor. App.* **39** (2005), 641–650.

[233] A. Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell *et al.*, editors, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[234] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell *et al.*, editors, Universitetsforlaget, Oslo, 1977, pp. 413–477.

[235] O. Toeplitz. Beispiele zur Theorie der fastperiodischen Funktionen. *Math. Annalen* **98** (1928), 281–295.

[236] C. R. Tompkins. Latin square Thue–Morse sequences are overlap-free. Manuscript available at `http://arxiv.org/abs/0706.0907`, 2007.

[237] V. I. Trofimov. Growth functions of some classes of languages. *Kybernetika* (1981), no. 6, I, 9–12, 149.

[238] J. Tromp and J. Shallit. Subword complexity of a generalized Thue–Morse word. *Inform. Process. Lett.* **54** (1995), 313–316.

[239] B. L. van der Waerden. Beweis einer Baudetschen Vermutung. *Nieuw Arch. Wisk.* **15** (1927), 212–216.

[240] M. Waldschmidt. Words and transcendence. Manuscript available at `http://www.institut.math.jussieu.fr/~miw/articles/pdf/WordsTranscendence.pdf`, 2007.

[241] E. M. Wright. Prouhet's 1851 solution of the Tarry–Escott problem of 1910. *Amer. Math. Monthly* **66** (1959), 199–201.

[242] A. I. Zimin. Blocking sets of terms. *Mat. Sbornik* **119** (1982), 363–375.