Repetition in Words

by

Seyyed Hamoon Mousavi Haji

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2013

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 main topic of this thesis is combinatorics on words. The field of combinatorics on words dates back at least to the beginning of the 20th century when Axel Thue constructed an infinite squarefree sequence over a ternary alphabet. From this celebrated result also emerged the subfield of repetition in words which is the main focus of this thesis.

One basic tool in the study of repetition in words is the iteration of morphisms. In Chapter 1, we introduce this tool among other basic notions. In Chapter 2, we see applications of iterated morphisms in several examples. The second half of the chapter contains a survey of results concerning Dejean's conjecture. In Chapter 3, we generalize Dejean's conjecture to circular factors. We see several applications of iterated morphism in this chapter. We continue our study of repetition in words in Chapter 4, where we study the length of the shortest repetition-free word in regular languages. Finally, in Chapter 5, we conclude by presenting a number of open problems.

Acknowledgements

I would like to express my many thanks to Jeffrey Shallit, for his support, encouragement, and supervision of the research presented in this thesis. I also extend my appreciation to Jonathan Buss and Larry Cummings for their role as readers of the thesis. I would like to thank James Currie and Narad Rampersad for many fruitful discussions. The assistance from Susan Gow was particularly helpful.

Contents

List of Figures

Chapter 1

Combinatorics on Words

In this chapter, we give the basic definitions needed for this thesis. We define words: the subject being studied in this thesis. We then define the basic tools in the study of words such as morphisms. Repetition, a well-studied concept in combinatorics on words, is also introduced. The interested reader can find more in the papers [10, 17, 34, 30].

1.1 Words

A *word* is a finite or infinite sequence $(a_i)_{i \geq 0}$ where the symbols a_i (also called letters) are taken from a finite set called the *alphabet*. For example the alphabets for the finite word $acbab$ and the infinite word $0111 \cdots$ are $\{a, b, c\}$ and $\{0, 1\}$ respectively. Alphabets with two and three letters are called binary and ternary alphabets, respectively. The empty word ϵ is the empty sequence.

For an alphabet Σ, the notation Σ^* is used to denote the set of finite words over Σ. A language is any subset $L \subseteq \Sigma^*$. Let Σ^ω denote the set of infinite words over Σ, and let $\Sigma^\infty = \Sigma^\omega \cup \Sigma^*$. Let $w = a_0 a_1 \cdots \in \Sigma^\infty$ be a word. Let $w[i] = a_i$, and let $w[i..j] = a_i \cdots a_j$. By convention $w[i..j] = \epsilon$ for $i > j$.

A *prefix* (*suffix*) of the word w is a word x such that $w = xy$ ($w = yx$) for some word y. The word z is a factor of w if $w = xzy$, for some words x and y. For a word x, let $\mathrm{pref}(x)$ and $\mathrm{suff}(x)$, respectively, denote the set of prefixes and *suffixes* of x. For example $\mathrm{pref}(abc) = \{\epsilon, a, ab, abc\}$ and $\mathrm{suff}(abc) = \{\epsilon, c, bc, abc\}$. For words x, y, let $x \preceq y$ denote that x is a *factor* of y. A factor x of y is *proper* if $x \neq y$ and is denoted by $x \prec y$. For example $b \prec abc$ but $ac \not\prec abc$. Let $x \preceq_p y$ (resp., $x \preceq_s y$) denote that x is a prefix (resp., suffix) of y. Let $x \prec_p y$ (resp., $x \prec_s y$) denote that x is a *proper* prefix (resp., proper suffix) of y; that is, a prefix (resp., suffix) such that $x \neq y$. A prefix p of w is a *period* of w if $w[i+r] = w[i]$ for $0 \leq i < |w| - r$, where $r = |p|$.

The *concatenation* or *product* of two words x and y, denoted by xy, is the juxtaposition of the symbols of x followed by y. For example $(ab)(cab) = abcab$. The empty word is the identity element for concatenation. Concatenation is an associative operator, and thus, we can omit the brackets in products such as $(xy)z$. We use exponentiation to represent the concatenation of a word with itself for a certain number of times, that is $x^k = \overbrace{xx \cdots x}^{k}$.

For an integer $k \geq 2$, a *k-power* is a nonempty word of the form $w = x^k$. For the special cases of $k = 2, 3$, such a word is called *square* and *cube*, respectively. An example of a square is `blahblah`. A word is *k-power-free* if it has no k-powers as factors. For example, the word `square` is squarefree and the word `squarefree` is not since it contains the square `ee`. A word of the form $axaxa$, where a is a single letter, and x is a (possibly empty) word, is called an *overlap*. For example, $abbabba$ is an overlap. A word is *overlap-free* if it has no factor that is an overlap.

A word is *primitive* if it is not a k-power for any $k \geq 2$. Two words x, y are *conjugate* if one is a cyclic shift of the other; that is, if there exist words u, v such that $x = uv$ and $y = vu$. The two words `bookcase` and `casebook` are conjugates. One simple observation is that all conjugates of a k-power are k-powers.

1.2 Morphisms

It is easy to see that the set Σ^* together with concatenation forms a free monoid. The map $h : \Sigma^* \to \Gamma^*$ between two monoids is said to be a *monoid homomorphism* (or just *morphism*) if it respects concatenation $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$.

The fact that Σ^* and Γ^* are free monoids implies that for any mapping from $\Sigma \to \Gamma^*$, there exists a unique extension to a morphism between Σ^* and Γ^*. In other words, to specify a morphism, we just need to define its image for all the single letters. For example, $h : \{a, b, c\}^* \to \{a, b, c\}^*$ where

$$h(a) = bac$$
$$h(b) = aac$$
$$h(c) = ab$$

is a morphism, and we have $h(abc) = bacaacab$.

A morphism $h : \Sigma^* \to \Gamma^*$ is said to be *q-uniform* if $|h(a)| = q$ for all $a \in \Sigma$. A morphism is uniform if it is q-uniform for some q. Let $h : \Sigma^* \to \Sigma^*$ be a morphism, and suppose $h(a) = ax$ for some letter a. The *fixed point* of h, starting with $a \in \Sigma$, is denoted by $h^\omega(a) = a\, x\, h(x)\, h^2(x) \cdots$. A word w is *pure morphic* if a nontrivial morphism h exists such that $w = h(w)$.

Let $\Sigma_m = \{0, 1, \ldots, m-1\}$. Define the morphism $\mu : \Sigma_2^* \to \Sigma_2^*$ as follows:

$$\mu(0) = 01$$
$$\mu(1) = 10.$$

We call $\mathbf{t} = \mu^\omega(0) = 01101001 \cdots$ the *Thue-Morse word* [2]. It is easy to see that

$$\mu(\mathbf{t}[0..n-1]) = \mathbf{t}[0..2n-1] \text{ for } n \geq 0.$$

A morphism h is k-power-free (resp., overlap-free) if $h(w)$ is k-power-free (resp., overlap-free) if w is. From classical results of Thue [32, 33], we know that the morphism μ is overlap-free. From [6], we know that that $\mu(x)$ is k-power free for each $k > 2$.

An infinite word w is said to be *recurrent* if every factor of w occurs infinitely often. A finite or infinite word w is *uniformly recurrent*, if for every factor x of w, an integer l exists such that every factor of w of length l contains x. A uniformly recurrent word is *linearly recurrent* if a constant C exists such that for every factor x of w, every factor of w of length $C|x|$ contains x. The Thue-Morse word is linearly recurrent.

Chapter 2

Repetition Avoidance

In this chapter, we briefly survey some of the main avoidability results in the literature. In Section 2.1, we summarize the basic techniques for proving repetition-freeness of morphisms. In Section 2.2, we define repetition threshold, a variation of which is studied in detail in Chapter 3. At the end, we highlight some of the proof techniques developed in [14, 27, 24, 9, 23, 13, 28] over several decades that eventually proved the famous Dejean conjecture.

2.1 Constructing Repetition-Free Words

The study of repetitions and, in general, patterns in words is the heart of combinatorics on words. The basic idea is that a long word that is picked at random tends to contain repetitions. A simple example is words of length greater than 3 over a binary alphabet. It is an easy exercise to observe that all such words contain squares as factors.

One goal in the study of repetition avoidance is to construct an infinite word that avoids certain repetitions. The principal tool for constructing infinite repetition-free words so far is the morphism. In this section, we illustrate the applications of this tool by means of examples.

Perhaps the most convenient way of constructing an infinite k-power-free word is to give a k-power-free morphism. As defined in Chapter 1, a morphism h is k-power-free if it preserves k-power-freeness. Suppose we have a k-power-free morphism $h : \Sigma^* \to \Sigma^*$. Clearly for $a \in \Sigma$, the words $h^i(a)$ for all i are k-power-free. Therefore, if $h^\omega(a)$ exists, it is k-power-free. The fixed point starting from a of h exists if $h(a) = ax$ for some x. Finally, in order for $h^\omega(a)$ to be infinite, we need to guarantee $|h^{i+1}(a)| > |h^i(a)|$. Next, we see an example of this method applied to construct a squarefree word.

Example 1. *Thue [33] gave the morphism*

$$h(a) = abcab$$
$$h(b) = acabcb$$
$$h(c) = acbcacb.$$

It is easy to check that $h^\omega(a)$ exists and is infinite. Therefore, all we need is to prove that h is a squarefree morphism. We observe that images of all the letters are squarefree. It is effortless to check that the same quality holds for images of short squarefree words, but we need a systematic way of deciding whether a given morphism is squarefree. Crochemore [11], as stated more precisely in the following proposition, has shown that a morphism is squarefree if it's images of all squarefree words less than a certain constant in length are squarefree.

Proposition 1. *Let $\mu : \Sigma^* \to \Gamma^*$ be a morphism. Then μ is squarefree if $\mu(x)$ is squarefree for all squarefree words x of length*

$$k = \max\{3, \left\lceil \frac{M(\mu) - 3}{m(\mu)} \right\rceil + 1\}$$

where

$$M(f) = \max\{|\mu(a)| : a \in \Sigma\}$$
$$m(f) = \min\{|\mu(a)| : a \in \Sigma\}.$$

For the morphism h we have

$$M(h) = 7,$$
$$m(h) = 5,$$
$$k = 3,$$

and one can easily check that $h(x)$ is squarefree for all $|x| \leq 3$. Therefore h is squarefree, and hence the word

$$h^\omega(a) = abcabacabcbacbcacbabcabacabcb\cdots$$

is squarefree.

As the next example indicates, it is possible to construct squarefree words using using non-squarefree morphisms.

Example 2. *The morphism*

$$h(0) = 01$$
$$h(1) = 23$$
$$h(2) = 03$$
$$h(3) = 21$$

is not squarefree since, for example, $h(031) = 012123$ *is not squarefree. Still, we prove* $w = h^{\omega}(0)$ *is squarefree.*

We note three simple properties of w *and* h*:*

1. $w[i]$ *is even if and only if* i *is even.*
2. *For* $a, b \in \{0, 1, 2, 3\}$*, if* $h(a)$ *and* $h(b)$ *start with the same letter, then the parities of* a *and* b *are the same.*
3. *For* $a, b \in \{0, 1, 2, 3\}$*, if* $h(a)$ *and* $h(b)$ *are distinct but end with the same letter, then the parities of* a *and* b *are different.*

By way of contradiction, suppose that $w[i..j] = uu$ *for some* $i < j$ *that minimizes* $|u|$*. Using (1), the length* $|u|$ *is even since* $w[i] = w[i + |u|]$*, and hence* i *and* $i + |u|$ *have the same parity. There are two cases to consider:*

1. $uu = h(vv) = w[i..j]$ *which implies* vv *is a smaller square in* w*, a contradiction.*
2. $uu = ah(v)bah(v)b = w[i..j]$ *for some* $a, b \in \{0, 1, 2, 3\}$*. If we let* $c = w[i-1]$ *and* $d = w[j + 1]$ *we can write*

$$cuud = cah(v)bah(v)bd = w[i - 1..j + 1].$$

There exist $e, f, g \in \{0, 1, 2, 3\}$ *such that*

$$h(e) = ca,$$
$$h(f) = ba,$$
$$h(g) = bd.$$

Hence $cah(v)bah(v)bd = h(evfvg)$*, and therefore*

$$evfvg \prec w.$$

Note that $e \neq f$ *because otherwise* $h(ev)h(ev) \in w$*, which is a contradiction using case* (1)*. Now using* (3)*, the parities of* e *and* f *are different, and hence* $|v|$ *is even. On the other hand, using* (2)*, we get that the parities of* f *and* g *are the same, so* $|v|$ *is odd, a contradiction.*

This is a typical argument for proving the repetition-freeness of a fixed point of a morphism. We employ similar ideas in Chapter 3.

2.2 Dejean's Conjecture

Repetition in words is an active research topic and has been studied for over a hundred years. For example, Axel Thue [32, 33] constructed an infinite word over a three-letter alphabet that contains no squares (i.e., no nonempty word of the form xx), and another infinite word over a two-letter alphabet that contains no cubes (i.e., no nonempty word of the form xxx).

In 1972, Dejean refined these results by considering fractional powers. An *α-power* for a rational number $\alpha \geq 1$ is a word of the form $w = x^{\lfloor \alpha \rfloor} x'$, where x' is a (possibly empty) prefix of x and $|w| = \alpha |x|$. The word w is a *repetition*, with a *period* x and an *exponent* α. Among all possible exponents, we let $\exp(w)$ denote the largest one, corresponding to the shortest period. For example, the word `alfalfa` has shortest period `alf` and exponent $\frac{7}{3}$. The *critical exponent* of a word w is the supremum, over all factors f of w, of $\exp(f)$. We write it as $\exp(w)$.

For a real number α, an *α^+-power* is a β-power where $\beta > \alpha$. For example $ababa = (ab)^{\frac{5}{2}}$ is a 2^+-power. A word w is

- *α^+-power-free*, if none of the factors of w is an α^+-power;
- *α-power-free* if, in addition to being α^+-power-free, the word w has no factor that is an α-power.

We also say that w *avoids* α^+-powers (resp., avoids α-powers). Dejean asked, what is the smallest real number r for which there exist infinite r^+-power-free words over an alphabet of size k? This quantity is known as the *repetition threshold* [6], and is denoted by $\mathrm{RT}(k)$. From results of Thue we know that $\mathrm{RT}(2) = 2$. Dejean [14] in 1972 proved $\mathrm{RT}(3) = \frac{7}{4}$, and conjectured that

$$\mathrm{RT}(k) = \begin{cases} \frac{7}{5}, & \text{if } k = 4; \\ \frac{k}{k-1}, & \text{if } k > 4. \end{cases}$$

This conjecture received much attention in the last forty years, and its proof was recently completed by Currie and Rampersad [13] and Rao [28], independently, based on work of Moulin-Ollagnier [24] and Carpi [9].

Thue [32] in 1906 proved $\mathrm{RT}(2) = 2$, and Dejean [14] in 1972 proved $\mathrm{RT}(3) = \frac{7}{4}$. They showed that there are only a finite number of $\mathrm{RT}(k)$-power-free words and gave $\mathrm{RT}(k)^+$-power-free morphisms for $k = 2, 3$. Thue's $\mathrm{RT}(2)^+$-power-free morphism is

$$\mu(a) = ab, \tag{2.1}$$
$$\mu(b) = ba, \tag{2.2}$$

as we introduced in Section 1.2, and Dejean's $\mathrm{RT}(3)^+$-power-free morphism is

$$\nu(a) = abcacbcabcbacbcacba, \tag{2.3}$$
$$\nu(b) = bcabacabcacbacabacb, \tag{2.4}$$
$$\nu(c) = cabcbabcabacbabcbac. \tag{2.5}$$

Brandenburg [6] realized that this approach, of using a repetition-free morphisms, cannot be applied to cases where $\mathrm{RT}(k) < \frac{3}{2}$.

Theorem 1 (Brandenburg)**.** *If* $\mathrm{RT}(k) < \frac{3}{2}$*, then there exists no growing* $\mathrm{RT}(k)^+$*-power-free morphism* $h : \Sigma_k^* \to \Sigma_k^*$*.*

Here, "growing" means that $|h(a)| > 1$ for all $a \in \Sigma_k$. Based on this theorem of Brandenburg and the conjecture of Dejean that predicts $\mathrm{RT}(k) < \frac{3}{2}$ for $k \geq 4$, researchers knew as early as 1981 that they needed a new method for $k \geq 4$.

Lemma 1 (Dejean)**.** *The repetition threshold,* $\mathrm{RT}(k)$*, is bounded below by* $\frac{k}{k-1}$*, that is,*

$$\mathrm{RT}(k) \geq \frac{k}{k-1}.$$

Proof. Let $w \in \Sigma_k^\omega$ be an arbitrary infinite word. We just need to prove that $\exp(w) \geq \frac{k}{k-1}$. If a subword of length k in w contains a repeated letter, then $\exp(w) \geq \frac{k}{k-1}$. Then we can assume that all subwords of length k have k distinct letters. Thus every subword of length $k+2$ in w begins and ends with the same word of length 2. See Figure 2.1.

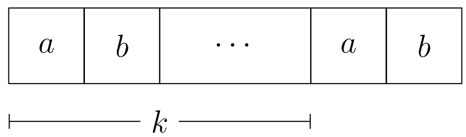


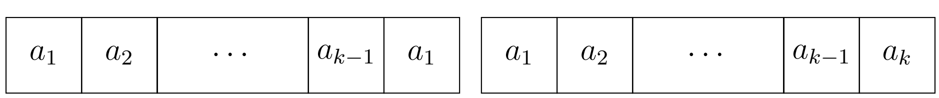
Figure 2.1: Subwords of length $k+2$ in w

So we have $\exp(w) \geq \frac{k+2}{k} \geq \frac{k}{k-1}$, provided $k \geq 2$. □

Based on this result, all that is needed to prove Dejean's conjecture is to prove that $\mathrm{RT}(k) \leq \frac{k}{k-1}$ for $k > 4$. In other words, all that is needed is to find a $(\frac{k}{k-1})^+$-power-free word over an alphabet of size k. The same is true for RT(4), where the conjectured value is $\frac{7}{5}$, since a computer search indicates that $\mathrm{RT}(4) \geq \frac{7}{5}$.

The first breakthrough in proving the Dejean's conjecture emerged in the work of Pansiot [27] in 1984. Pansiot [27] introduced a compact binary encoding of $\frac{k-1}{k-2}$-power-free words over Σ_k known as the *Pansiot encoding*. Since $(\frac{k}{k-1})^+$-power-free words are also $\frac{k-1}{k-2}$-power-free, the Pansiot encoding also exists for $(\frac{k}{k-1})^+$-power-free words. The Pansiot encoding is defined as follows.

Let $w \in \Sigma_k^*$ be a $\frac{k-1}{k-2}$-power-free word. Being $\frac{k-1}{k-2}$-power-free implies that every subword of w of length $k-1$ has $k-1$ distinct letters. This, in turn, implies that every subword of w of length k either contains $k-1$ distinct letters and the last letter is the same as the first letter or contains k distinct letters. The former is called type 0 factor, and the latter is called type 1. See Figure 2.2.



(a) Type 0 (b) Type 1

Figure 2.2: Subwords of w of length k are either of type 0 or 1, where $\{a_1, \ldots, a_k\} = \Sigma_k$

In other words, for a $\frac{k-1}{k-2}$-power-free word w and for every $i < |w| - k$, we have either

$$w[i + k - 1] = w[i], \text{ or} \tag{2.6}$$
$$w[i + k - 1] \in \Sigma_k - \{w[i], w[i + 1], \ldots, w[i + k - 2]\}\,. \tag{2.7}$$

Note that since $w[i..i+k-2]$ has $k-1$ distinct letters, the set $\Sigma_k - \{w[i], w[i + 1], \ldots, w[i + k - 2]\}$ is a singleton, so (2.7) determines $w[i + k - 1]$ uniquely. Let us record for every i whether $w[i] = w[i + k - 1]$. For this purpose define a new word b as follows:

$$b[i] = \begin{cases} 0, & \text{if } w[i] = w[i + k - 1]; \\ 1, & \text{if } w[i] \neq w[i + k - 1]; \end{cases}$$

for $0 \leq i \leq |w| - k$. We call this new word b the Pansiot encoding of w. Here is an example in Σ_5: Suppose that $w = 012304132$, then $b = 01101$.

A nice property of the Pansiot encoding is that using the first $k - 1$ letters of w and b, we can reconstruct w. For example if $b = 101011$ and $w \in \Sigma_4^*$ starts with 130, we can uniquely determine $w = 130231203$. Pansiot [27] gives the morphism

$$h(0) = 101101$$
$$h(1) = 10.$$

Suppose $w = 012 \cdots$ is the unique word over Σ_4 with Pansiot encoding $h^\omega(0)$. Pansiot then proves that w is $\frac{5}{4}^+$-power-free, and completes the proof of Dejean's conjecture for $k = 4$.

The next major step in proving Dejean's conjecture was taken by Moulin Ollagnier [24], by observing a connection between Pansiot encoding and the symmetric group. This connection relates repetitions in words to the identity element in the symmetric group. Let σ_0 and σ_1 be two permutations in the symmetric group on Σ_k defined by

$$\sigma_0 = \begin{pmatrix} 0 & 1 & 2 & \cdots & k-3 & k-2 & k-1 \\ 1 & 2 & 3 & \cdots & k-2 & 0 & k-1 \end{pmatrix}$$
$$\sigma_1 = \begin{pmatrix} 0 & 1 & 2 & \cdots & k-3 & k-2 & k-1 \\ 1 & 2 & 3 & \cdots & k-2 & k-1 & 0 \end{pmatrix}.$$

The permutation σ_0 is the cycle on the first $k-1$ elements of Σ_k, and σ_1 is the cycle on all the k elements. Define the monoid morphism $\eta : \Sigma_2^* \to S_k$ where $\eta(0) = \sigma_0$ and $\eta(1) = \sigma_1$.

Let us illustrate these definitions in an example in Σ_5. Suppose that $w = 01230423140$. The Pansiot encoding is then $b = 0100101$. We apply η on all prefixes of b and we obtain

$$\eta(\epsilon) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix}$$
$$\eta(0) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 0 & 4 \end{pmatrix}$$
$$\eta(01) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 2 & 3 & 0 & 4 & 1 \end{pmatrix}$$
$$\eta(010) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 3 & 0 & 4 & 2 & 1 \end{pmatrix}$$
$$\eta(0100) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 4 & 2 & 3 & 1 \end{pmatrix}$$
$$\eta(01001) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 2 & 3 & 1 & 0 \end{pmatrix}$$
$$\eta(010010) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 2 & 3 & 1 & 4 & 0 \end{pmatrix}$$
$$\eta(0100101) = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 0 & 2 \end{pmatrix}.$$

The second row and first four columns of each of the above permutations is a factor of w of length 4. This is no coincidence. In fact, a simple induction proves that for a $\frac{k-1}{k-2}$-power-free word $w = 0123 \cdots k-2 \cdots$, we can write

$$\eta(b[0..i]) = \begin{pmatrix} 0 & 1 & 2 & \cdots & k-3 & k-2 & k-1 \\ w[i] & w[i+1] & w[i+2] & \cdots & w[i+k-3] & w[i+k-2] & a \end{pmatrix}$$

where a is the unique letter in $\Sigma_k - \{w[i], w[i+1], \ldots, w[i+k-2]\}$.

Now if $w[i..i+k-1] = w[j..j+k-1]$, then $\eta(b[0..i]) = \eta(b[0..j])$. It follows immediately that $\eta(b[i+1..j]) = id_k$ where id_k is the identity element of S_k. In other words, the Pansiot encoding of repetitions in words (or at least those that are long) are kernels of the morphism η. To put it simply, in order to avoid repetitions, we need to control kernels in Pansiot encodings. This view enabled Moulin Ollagnier to prove Dejean's conjecture for $5 \leq k \leq 11$ in 1989.

The last major step was taken by Carpi [9] in 2007. Carpi proved that Dejean's conjecture holds for $k \geq 33$ by extending the work of Moulin Ollagnier. The remaining cases, i.e., the cases $11 < k < 33$ were proved independently by Currie, Rampersad and Mohammad-Noori [13, 23] and Rao [28] in 2009, along the lines of the proof by Carpi.

Chapter 3

Repetition Avoidance in Circular Factors

In this chapter[1], we consider the following novel variation on a classical avoidance problem from combinatorics on words: instead of avoiding repetitions in all factors of a word, we avoid repetitions in all factors where each individual factor is considered as a "circular word", i.e., the end of the word wraps around to the beginning. We determine the best possible avoidance exponent for alphabet size 2 and 3, and provide a lower bound for larger alphabets. The main result of this chapter is Theorem 4.

3.1 Introduction

We consider the following novel variation on Dejean, which we call "circular α-power avoidance". We consider each finite factor x of a word w, but interpret such a factor as a "circular" word, where the end of the word wraps around to the beginning. Then we consider each factor f of this interpretation of x; for w to be circularly α-power-free, each such f must be α-power-free. For example, consider the English word $w = \texttt{dividing}$ with factor $x = \texttt{dividi}$. The circular shifts of x are

$$\texttt{dividi}, \texttt{ividid}, \texttt{vididi}, \texttt{ididiv}, \texttt{didivi}, \texttt{idivid},$$

and (for example) the word `ididiv` contains a factor `ididi` that is a $\frac{5}{2}$-power. In fact, w is circularly cubefree and circularly $(\frac{5}{2})^+$-power-free.

[1]The contents of this chapter are taken largely verbatim from Mousavi and Shallit [25].

To make this more precise, we recall the notion of conjugacy. Two words x, y are *conjugate* if one is a cyclic shift of the other; that is, if there exist words u, v such that $x = uv$ and $y = vu$.

Definition 1. *Let w be a finite or infinite word. The largest circular α-power in a word w is defined to be the supremum of* $\exp(f)$ *over all factors f of conjugates of factors of w. We write it as* $\text{cexp}(w)$.

Although Definition 1 characterizes the subject of this chapter, we could have used a different definition, based on the following.

Proposition 2. *Let w be a finite word or infinite word. The following are equivalent:*

(a) s is a factor of a conjugate of a factor of w;

(b) s is a prefix of a conjugate of a factor of w;

(c) s is a suffix of a conjugate of a factor of w;

(d) $s = vt$ for some factor tuv of w.

Proof. (a) $\implies$ (b): Suppose $s = y''x'$, where xy is a factor of w and $x = x'x''$ and $y = y'y''$. Another conjugate of xy is then $y''x'x''y'$ with prefix $y''x'$.

(b) $\implies$ (c): Such a prefix s is either of the form y' or yx', where xy be a factor of w and $x = x'x''$ and $y = y'y''$. Considering the conjugate $y''xy'$ of yx, we get a suffix y', and consider the conjugate $x''yx'$ we get a suffix yx'.

(c) $\implies$ (d): Such a suffix s is either of the form $s = x''$ or $s = y''x$, where xy be a factor of w and $x = x'x''$ and $y = y'y''$. In the former case, let $t = x''$, $u = v = \epsilon$. In the latter case, let $t = x$, $u = y'$, and $v = y''$.

(d) $\implies$ (a): Let tuv be a factor of w. Then vtu is a conjugate of tuv, and vt is a factor of it.

$\square$

Let $\Sigma_k = \{0, 1, \ldots, k-1\}$. Define $\text{RTC}(k)$, the *repetition threshold for circular factors*, to be the smallest real number r for which there exist infinite circularly r^+-power-free words in Σ_k. Clearly we have

$$\text{RTC}(k) \geq \text{RT}(k).$$

In this paper we prove that RTC(2) = 4 and RTC(3) = $\frac{13}{4}$. For larger alphabets, we conjecture that

$$\mathrm{RTC}(k) = \begin{cases} \frac{5}{2}, & \text{if } k = 4; \\ \frac{105}{46}, & \text{if } k = 5; \\ \frac{2k-1}{k-1}, & \text{if } k \geq 6. \end{cases}$$

In the next section, we prove some preliminary results. We get some bounds for RTC(k), and in particular, we prove that RTC(2) = 2 RT(2) = 4. In Section 3.3, we study the three-letter alphabet, and we prove that RTC(3) = $\frac{13}{4}$. Finally, in Section 3.4, we give another interpretation for repetition threshold for circular factors.

Finally, we point out that the quantities we study here are *not* closely related to the notion of *avoidance in circular words*, studied previously in [1, 15, 18]. Aberkane and Currie [1] proved a conjecture in Alon et al. [3]. Alon et al. introduced the concept of nonrepetitive coloring of graphs. A nonrepetitive coloring of a graph is a coloring for which the sequence of colors in every cycle-free path contains no square. They conjectured there exist nonrepetitive coloring of C_n, cycle on n vertices, for every $n \geq 18$.

Related to C_n is the notion of circular words. A circular word is a word that the end is linked to the beginning, forming a cycle. Gorbunova [15] studied repetition threshold on circular words, and proved for every $k \geq 6$, there exist $\left(\frac{\lceil \frac{k}{2} \rceil + 1}{\lceil \frac{k}{2} \rceil}\right)^+$-power-free circular words of every length.

3.2 Binary Alphabet

First of all, we prove a bound on RTC(k).

Theorem 2. $1 + \mathrm{RT}(k) \leq \mathrm{RTC}(k) \leq 2\,\mathrm{RT}(k)$.

Proof. Let $r = \mathrm{RT}(k)$. We first prove that $\mathrm{RTC}(k) \leq 2r$. Let $w \in \Sigma_k^\omega$ be an r^+-power-free word. We prove that w is circularly $(2r)^+$-power-free. Suppose that $xty \preceq w$, such that yx is $(2r)^+$-power. Now either y or x is an r^+-power. This implies that w contains an r^+-power, a contradiction.

Now we prove that $1 + r \leq \mathrm{RTC}(k)$. Let l be the length of the longest r-power-free word over Σ_k, and let $w \in \Sigma_k^\omega$. Considering the factors of length $n = l + 1$ of w, we know some factor f must occur infinitely often. This f contains an r-power: z^r. Therefore

$z^r tz$ is a factor of w. Therefore w contains a circular $(1+r)$-power. This proves that $1 + r \leq \mathrm{RTC}(k)$. □

Note that since $\mathrm{RT}(k) > 1$, we have $\mathrm{RTC}(k) > 2$.

Lemma 2. $\mathrm{RTC}(2) \geq 4$.

Proof. Let $w \in \Sigma_2^\omega$ be an arbitrary word. It suffices to prove that w contains circular 4-powers. There are two cases: either 00 or 11 appears infinitely often, or w ends with a suffix of the form $(01)^\omega$. In the latter case, obviously there are circular 4-powers; in the former there are words of the form $aayaa$ for $a \in \Sigma_2$ and $y \in \Sigma_2^*$ and hence circular 4-powers. □

Theorem 3. $\mathrm{RTC}(2) = 4$.

Proof. A direct consequence of Theorem 2 and Lemma 2. □

The Thue-Morse word is an example of a binary word that avoids circular 4^+-powers.

3.3 Ternary Alphabet

Our goal in this section is to show that $\mathrm{RTC}(3) = \frac{13}{4}$. For this purpose, we frequently use the notion of synchronizing morphism, which was introduced in Ilie et al. [20].

Definition 2. *A morphism $h : \Sigma^* \to \Gamma^*$ is said to be* synchronizing *if for all $a, b, c \in \Sigma$ and $s, r \in \Gamma^*$, if $h(ab) = rh(c)s$, then either $r = \epsilon$ and $a = c$ or $s = \epsilon$ and $b = c$.*

Definition 3. *A synchronizing morphism $h : \Sigma^* \to \Gamma^*$ is said to be* strongly synchronizing *if for all $a, b, c \in \Sigma$, if $h(c) \in \mathrm{pref}(h(a))\,\mathrm{suff}(h(b))$, then either $c = a$ or $c = b$.*

The following technical lemma is applied several times throughout the paper.

Lemma 3. *Let $h : \Sigma^* \to \Gamma^*$ be a synchronizing q-uniform morphism. Let $n > 1$ be an integer, and let $w \in \Sigma^*$. If $z^n \preceq_p h(w)$ and $|z| \geq q$, then $u^n \preceq_p w$ for some u. Furthermore $|z| \equiv 0 \pmod{q}$.*

Proof. Let $z = h(u)z'$, where $|z'| < q$ and $u \in \Sigma^*$. Note that $u \neq \epsilon$, since $|z| \geq q$. Clearly, we have $z'h(u[0]) \preceq_p h(w[|u|..|u|+1])$. Since h is synchronizing, the only possibility is that $z' = \epsilon$, so $|z| \equiv 0 \pmod{q}$. Now we can write $z^n = h(u^n) \preceq_p h(w)$. Therefore $u^n \preceq_p w$. □

The next lemma states that if the fixed point of a strongly synchronizing morphism (SSM) avoids small n'th powers, where n is an integer, it avoids n'th powers of all lengths.

Lemma 4. *Let $h : \Sigma^* \to \Sigma^*$ be a strongly synchronizing q-uniform morphism. Let $n > 1$ be an integer. If $h^\omega(0)$ avoids factors of the form z^n, where $|z^n| < 2nq$, then $h^\omega(0)$ avoids n'th powers.*

Proof. Let $w = a_0 a_1 a_2 \cdots = h^\omega(0)$. Suppose w has n'th powers of length greater than or equal to $2nq$. Let z be the shortest such word, i.e., $|z^n| \geq 2nq$ and $z^n \preceq w$. We can write

$$\begin{aligned} &z^n = xh(w[i..j])y, \\ &x \preceq_s h(a_{i-1}), \\ &y \preceq_p h(a_{j+1}), \\ &|x|, |y| < q, \end{aligned}$$

for some integers $i, j \geq 0$. If $x = y = \epsilon$, then using Lemma 3, since $|z| \geq q$, the word $w[i..j]$ contains an n'th power. Therefore w contains an n'th power of length smaller than $|z^n|$, a contradiction. Now suppose that $xy \neq \epsilon$. Since $|z| \geq \frac{2nq}{n} = 2q$, and $|xh(w[i])|, |h(w[j])y| < 2q$, we can write

$$\begin{aligned} &xh(w[i]) \preceq_p z, \\ &h(w[j])y \preceq_s z. \end{aligned}$$

Therefore $h(w[j])yxh(w[i]) \preceq z^2 \preceq z^n$. Since h is synchronizing

$$h(w[j])yxh(w[i]) \preceq h(w[i..j]).$$

Hence $yx = h(a)$ for some $a \in \Sigma$. Since h is an SSM, we have either $a = a_{i-1}$ or $a = a_{j+1}$. Without loss of generality, suppose that $a = a_{i-1}$. Then we can write $h(w[i-1..j]) = yxh(w[i..j])$. The word $yxh(w[i..j])$ is an n'th power, since it is a conjugate of $xh(w[i..j])y$. So we can write

$$h(w[i-1..j]) = z_1^n$$

where z_1 is a conjugate of z. Note that $|z_1| = |z| \geq 2q$. Now using Lemma 3, the word $w[i-1..j]$ contains an n'th power, and hence w contains an n'th power of length smaller than $|z^n|$, a contradiction. ☐

The following lemma states that, for an SSM h and a well-chosen word w, all circular $(\frac{13}{4})^+$-powers in $h(w)$ are small.

Lemma 5. *Let $h : \Sigma^* \to \Gamma^*$ be a strongly synchronizing q-uniform morphism. Let $w = a_0a_1a_2 \cdots \in \Sigma^\omega$ be a circularly cubefree word. In addition, suppose that w is squarefree. If $x_1tx_2 \preceq h(w)$ for some words t, x_1, x_2, and x_2x_1 is a $(13/4)^+$-power, then $|x_2x_1| < 22q$.*

Proof. The proof is by contradiction. Suppose there are words t, x_1, x_2, and z in Γ^* and a rational number $\alpha > \frac{13}{4}$ such that

$$x_1tx_2 \preceq h(w)$$

$$|x_2x_1| \geq 22q$$

$$x_2x_1 = z^\alpha.$$

Suppose $|z| < q$. Let k be the smallest integer for which $|z^k| \geq q$. Then $|z^k| < 2q$, because otherwise $|z^{k-1}| \geq q$, a contradiction. We can write $x_2x_1 = (z^k)^\beta$, where $\beta = \frac{|x_2x_1|}{|z^k|} > \frac{22q}{2q} > \frac{13}{4}$. Therefore we can assume that $|z| \geq q$, since otherwise we can always replace z with z^k, and α with β.

There are three cases to consider.

(a) Suppose that x_1 and x_2 are both long enough, so that each contains an image of a word under h. More formally, suppose that

$$
\begin{aligned}
& x_1 = y_1h(w[i_1..j_1])y_2, && (3.1)\\
& x_2 = y_3h(w[i_2..j_2])y_4, && (3.2)\\
& i_1 \leq j_1, i_2 \leq j_2,\\
& y_1 \preceq_s h(a_{i_1-1}),\\
& y_2 \preceq_p h(a_{j_1+1}),\\
& y_3 \preceq_s h(a_{i_2-1}),\\
& y_4 \preceq_p h(a_{j_2+1}),\\
& |y_1|, |y_2|, |y_3|, \text{ and } |y_4| < q, \text{ and}\\
& y_2ty_3 = h(w[j_1+1..i_2-1]).
\end{aligned}
$$

Let $v_1 = w[i_1..j_1]$ and $v_2 = w[i_2..j_2]$. See Figure 3.1.

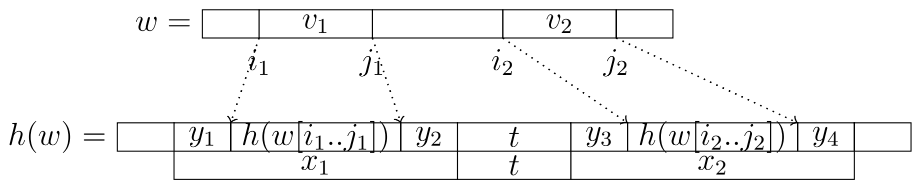


Figure 3.1: x_1tx_2 is a factor of $h(w)$

There are two cases to consider.

(1) Suppose that $y_4y_1 = \epsilon$. Let $v = w[i_2..j_2]w[i_1..j_1]$.
The word $h(v)y_2$ is a factor of $y_3h(v)y_2 = z^\alpha$ of length $\geq 22q - q = 21q$, and so

$$h(v)y_2 = z_1^\beta,$$

where z_1 is a conjugate of z, and $\beta \geq \frac{21}{22}\alpha > 3$. Therefore we can write

$$z_1^3 \preceq_p h(v)y_2 \preceq_p h(vw[j_1 + 1]).$$

Note that $|z_1| = |z| \geq q$, so using Lemma 3, we can write $|z_1| \equiv 0 \pmod{q}$. Therefore

$$z_1^3 \preceq_p h(v).$$

Using Lemma 3 again, the word v contains a cube, which means that the word w contains a circular cube, a contradiction.

(2) Suppose that $y_4y_1 \neq \epsilon$. We show how to get two new factors $x'_1 = h(v'_1)y'_2$ and $x'_2 = y'_3h(v'_2)$, with v'_1, v'_2 nonempty, such that $x'_2x'_1 = x_2x_1$. Then we use case (1) above to get a contradiction.
Let $s = h(w[j_2])y_4y_1h(w[i_1])$, and let m be the smallest integer for which $|z^m| \geq |s|$. Note that if $|z| < |s|$, then

$$|z^m| < 2|s| < 8q. \tag{3.3}$$

We show that at least one of the following inequalities holds:

$$|h(v_1)| \geq q + |z^m|,$$
$$|h(v_2)| \geq q + |z^m|.$$

Suppose that both inequalities fail. Then using (3.1) and (3.2) we can write

$$|x_2x_1| < 2q + 2|z^m| + |y_1y_2y_3y_4| < 6q + 2|z^m|. \tag{3.4}$$

If $|z| < |s|$, then by (3.3) and (3.4) one gets $|x_2x_1| < 22q$, contradicting our assumption. Otherwise $|z| \geq |s|$, and hence $m = 1$. Then

$$|x_2x_1| = \alpha|z| < 2q + 2|z| + |y_1y_2y_3y_4| < 6q + 2|z|,$$

and hence $|z| < 6q$. So $|x_2x_1| < 6q + 2|z| < 18q$, contradicting our assumption. Without loss of generality, suppose that $|h(v_1)| \geq q + |z^m|$.
Using the fact that z is a period of x_2x_1, we can write

$$h(v_1)[q + |z^m| - |s|..q + |z^m| - 1] = s,$$

or, in other words,

$$s \preceq h(v_1).$$

See Figure 3.2.

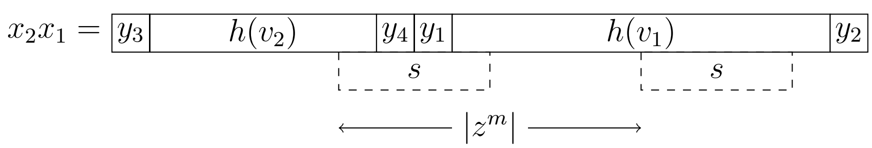


Figure 3.2: $h(v_1)$ contains a copy of s

Using the fact that h is synchronizing, we get that $y_4y_1 = h(a)$ for some $a \in \Sigma$. Since h is an SSM, we have either $a = a_{i_1-1}$ or $a = a_{j_2+1}$. Without loss of generality, suppose that $a = a_{j_2+1}$. Now look at the following factors of $h(w)$, which can be obtained from x_1 and x_2 by extending x_2 to the right and shrinking x_1 from the left:

$$x'_1 = h(w[i_1..j_1])y_2$$
$$x'_2 = y_3h(w[i_2..j_2 + 1]).$$

See Figure 3.3.

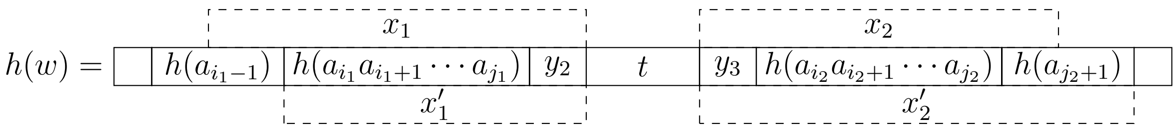


Figure 3.3: x'_1 and x'_2 are obtained from x_1 and x_2

We can see that

$$x'_2x'_1 = y_3h(w[i_2..j_2+1])h(w[i_1..j_1])y_2 = y_3h(w[i_2..j_2])y_4y_1h(w[i_1..j_1])y_2 = x_2x_1.$$

Now using case (1) we get a contradiction.

(b) Suppose that x_2 is too short to contain an image of a word under h. More formally, we can write

$$x_1 = y_1h(v)y_2 \text{ where } |x_2| < 2q \text{ and } |y_1|, |y_2| < q$$

for some words $y_1, y_2 \in \Gamma^*$ and a word $v \preceq w$. Then $h(v)$ is a factor of $x_2x_1 = z^\alpha$ of length $\geq 22q - 4q = 18q$, and so

$$h(v) = z_1^\beta,$$

where z_1 is a conjugate of z, and $\beta \geq \frac{18}{22}\alpha > 2$. Note that $|z_1| = |z| \geq q$, so using Lemma 3, the word v contains a square, a contradiction.

(c) Suppose that x_1 is not long enough to contain an image of a word under h. An argument similar to (b) applies here to get a contradiction.

□

The following 15-uniform morphism is an example of an SSM:

$$\mu(0) = 012102120102012$$
$$\mu(1) = 201020121012021$$
$$\mu(2) = 012102010212010$$
$$\mu(3) = 201210212021012$$
$$\mu(4) = 102120121012021$$
$$\mu(5) = 102010212021012.$$

Another example of an SSM is the 4-uniform morphism $\psi : \Sigma_6^* \to \Sigma_6^*$ as follows:

$$\begin{aligned}\psi(0) &= 0435\\ \psi(1) &= 2341\\ \psi(2) &= 3542\\ \psi(3) &= 3540\\ \psi(4) &= 4134\\ \psi(5) &= 4105.\end{aligned}$$

Our goal is to show that $\mu(\psi^\omega(0))$ is circularly $(\frac{13}{4})^+$-power-free. For this purpose, we first prove that $\psi^\omega(0)$ is circularly cubefree. Then we apply Lemma 5, for $h = \mu$ and $w = \psi^\omega(0)$.

Lemma 6. *The fixed point $\psi^\omega(0)$ is squarefree.*

Proof. Suppose that $\psi^\omega(0)$ contains a square. Using Lemma 4, there is a square $zz \preceq \psi^\omega(0)$ such that $|zz| < 16$. Using a computer program, we checked all factors of length smaller than 16 in $\psi^\omega(0)$, and none of them is a square. This is a contradiction. □

Lemma 7. *The fixed point $\psi^\omega(0)$ is circularly cubefree.*

Proof. By contradiction. Let $w = a_0a_1a_2\cdots = \psi^\omega(0)$. Suppose $x_1tx_2 \preceq w$, and $x_2x_1 = z^3$ for some words t, x_1, x_2, z, where

$$\begin{aligned}&x_1 = y_1\psi(w[i_1..j_1])y_2,\\ &x_2 = y_3\psi(w[i_2..j_2])y_4,\\ &y_1 \preceq_s \psi(a_{i_1-1}),\\ &y_2 \preceq_p \psi(a_{j_1+1}),\\ &y_3 \preceq_s \psi(a_{i_2-1}),\\ &y_4 \preceq_p \psi(a_{j_2+1}),\\ &|y_1|, |y_2|, |y_3|, \text{ and } |y_4| < 4,\\ &y_2ty_3 = \psi(w[j_1+1..i_2-1]),\end{aligned}$$

for proper choices of the integers i_1, i_2, j_1, j_2. Let $v_1 = w[i_1..j_1]$ and $v_2 = w[i_2..j_2]$.

Using a computer program, we searched for circular cubes of length not greater than 66, and it turns out that there is no such circular cube in w. Thus we can assume that $|x_2x_1| > 66$ so $|z| > 22$. Moreover suppose that x_2x_1 has the smallest possible length.

There are two cases to consider.

(a) Suppose that $y_4y_1 = \epsilon$. If $y_2y_3 = \epsilon$, then $\psi(v_2v_1) = z^3$. Using Lemma 3, we get that v_2v_1 contains a cube. Hence w contains a smaller circular cube, a contradiction.

Suppose that $y_2y_3 \neq \epsilon$. Since $|y_3\psi(w[i_2])|, |\psi(w[j_1])y_2| < 8$ and $|z| > 22$, we can write

$$y_3\psi(w[i_2]) \preceq_p z,$$
$$\psi(w[j_1])y_2 \preceq_s z.$$

Therefore $\psi(w[j_1])y_2y_3\psi(w[i_2]) \preceq z^3$, and since ψ is synchronizing

$$\psi(w[j_1])y_2y_3\psi(w[i_2]) \preceq \psi(v_2v_1).$$

Hence $y_2y_3 = \psi(b)$ for some $b \in \Sigma_6$. Since ψ is an SSM, we have either $b = a_{i_2-1}$, or $b = a_{j_1+1}$. Without loss of generality, suppose that $b = a_{i_2-1}$. So we can write

$$\psi(w[i_2 - 1..j_2]w[i_1..j_1]) = y_2y_3\psi(w[i_2..j_2]w[i_1..j_1]).$$

The word $y_2y_3\psi(v_2v_1)$ is a cube, since it is a conjugate of $y_3\psi(v_2v_1)y_2$. So we can write

$$\psi(w[i_2 - 1..j_2]w[i_1..j_1]) = z_1^3$$

where z_1 is a conjugate of z. Then using Lemma 3, the word $w[i_2 - 1..j_2]w[i_1..j_1]$ contains a cube. Note that since $y_2y_3 \neq \epsilon$ we have $j_1 < i_2 - 1$. Hence $w[i_2 - 1..j_2]w[i_1..j_1]$ is a circular cube of w, a contradiction.

(b) Suppose that $y_4y_1 \neq \epsilon$. We show how to get two new factors $x'_1 = h(v'_1)y'_2$ and $x'_2 = y'_3h(v'_2)$ of w, for nonempty words v'_1, v'_2, such that $x'_2x'_1 = x_2x_1$. Then we use case (a) above to get a contradiction.

The word w is squarefree due to Lemma 6. Therefore $|x_1|, |x_2| > |z| > \frac{66}{3}$ and hence $|v_1|, |v_2| > 0$. One can observe that either $|\psi(v_1)| \geq 4 + |z|$ or $|\psi(v_2)| \geq 4 + |z|$. Without loss of generality, suppose that $|\psi(v_1)| \geq 4 + |z|$. Let $s = w[j_2]y_4y_1w[i_1]$. Now, using the fact that z is a period of x_2x_1, we can write

$$\psi(v_1)[4 + |z| - |s|..4 + |z| - 1] = s,$$

or, in other words,

$$s \preceq \psi(v_1).$$

Using the fact that ψ is synchronizing, we get that $y_4 y_1 = \psi(a)$ for some $a \in \Sigma_6$. Since ψ is an SSM, we have either $a = a_{i_1-1}$, or $a = a_{j_2+1}$. Without loss of generality, suppose that $a = a_{j_2+1}$. Now look at the following factors of w, which can be obtained from x_1 and x_2 by extending x_2 to the right and shrinking x_1 from the left

$$x_1' = \psi(w[i_1..j_1])y_2$$
$$x_2' = y_3\psi(w[i_2..j_2+1]).$$

We can write

$$x_2'x_1' = y_3\psi(w[i_2..j_2+1])\psi(w[i_1..j_1])y_2 = y_3\psi(v_2)y_4y_1\psi(v_1)y_2 = x_2x_1 = z^3.$$

So using case (a) we get a contradiction.

□

Theorem 4. $\mathrm{RTC}(3) = \frac{13}{4}$.

Proof. First let us show that $\mathrm{RTC}(3) \geq \frac{13}{4}$.

Suppose there exists an infinite word w that avoids circular α-powers, for $\alpha < 4$. We now argue that for every integer C, there exists an infinite word w' that avoids both squares of length $\leq C$ and circular α-powers. Note that none of the factors of w looks like $xxyxx$, since w avoids circular 4-powers. Therefore, every square in w occurs only finitely many times. Therefore w' can be obtained by removing a sufficiently long prefix of w.

Computer search verifies that the longest circularly $\frac{13}{4}$-power-free word over a 3-letter alphabet that avoids squares xx where $|xx| < 147$ has length 147. Therefore the above argument for $C = 147$ shows that circular $\frac{13}{4}$-powers are unavoidable over a 3-letter alphabet.

Now to prove $\mathrm{RTC}(3) = \frac{13}{4}$, it is sufficient to give an example of an infinite word that avoids circular $(\frac{13}{4})^+$-powers. We claim that $\mu(\psi^\omega(0))$ is such an example. We know that $\psi^\omega(0)$ is circularly cubefree. Therefore we can use Lemma 5 for $w = \psi^\omega(0)$ and $h = \mu$. So if $xty \preceq \mu(\psi^\omega(0))$, and yx is a $(\frac{13}{4})^+$-power, then $|yx| < 22 \times 15$. Now there are finitely many possibilities for x and y. Using a computer program, we checked that none of them leads to a $(\frac{13}{4})^+$-power. This completes the proof. □

3.4 Another Interpretation

We could, instead, consider the supremum of $\exp(p)$ over all products of i factors of w. Call this quantity $\text{pexp}_i(w)$.

Proposition 3. *If w is a recurrent infinite word, then* $\text{pexp}_2(w) = \text{cexp}(w)$.

Proof. Let s be a product of two factors of w, say $s = xy$. Let y occur for the first time at position i of w. Since w is recurrent, x occurs somewhere after position $i + |y|$ in w. So there exists z such that yzx is a factor of w. Then xy is a factor of a conjugate of a factor of w.

On the other hand, from Proposition 2, we know that if s is a conjugate of a factor of w, then $s = vt$ where tuv is a factor of w. Then s is the product of two factors of w. $\square$

We can now study the repetition threshold for i-term products, $\text{RT}_i(k)$, which is the infimum of $\text{pexp}_i(w)$ over all words $w \in \Sigma_k^\omega$. Note that

$$\text{RT}_2(k) \geq \text{RTC}(k).$$

The two recurrent words, the Thue-Morse word and $\mu(\psi^\omega(0))$, introduced in Section 3.3, are circularly $\text{RTC}(2)^+$-power-free and circularly $\text{RTC}(3)^+$-power-free, respectively. Using Proposition 3, we get that $\text{RT}_2(k) = \text{RTC}(k)$ for $k = 2, 3$.

Theorem 5. *For $i \geq 1$ we have* $\text{RT}_i(2) = 2i$.

Proof. From Thue we know there exists an infinite 2^+-power-free word. If some product of factors $x_1 x_2 \cdots x_i$ contains a $(2i)^+$-power, then some factor contains a 2^+-power, a contradiction. So $\text{RT}_i(2) \leq 2i$.

For the lower bound, fix $i \geq 2$, and let $w \in \Sigma_2^\omega$ be an arbitrary word. Either 00 or 11 appears infinitely often, or w ends in a suffix of the form $(01)^\omega$. In the latter case we get arbitrarily high powers, and the former case there is a product of i factors with exponent $2i$. $\square$

It would be interesting to obtain more values of $\text{RT}_i(k)$. We propose the following conjectures which are supported by numerical evidence:

$$\begin{aligned}
\mathrm{RT}_2(4) &= \mathrm{RTC}(4) = \frac{5}{2}\ , \\
\mathrm{RT}_2(5) &= \mathrm{RTC}(5) = \frac{105}{46}\ , \text{ and} \\
\mathrm{RT}_2(k) &= \mathrm{RTC}(k) = 1 + \mathrm{RT}(k) = \frac{2k-1}{k-1} \text{ for } k \geq 6.
\end{aligned}$$

We know that the values given above are lower bounds for $\mathrm{RTC}(k)$.

Chapter 4

Automata Accepting Repetition-Free Words

In this chapter[1] we consider the following problem: given that a finite automaton M of N states accepts at least one k-power-free (resp., overlap-free) word, what is the length of the shortest such word accepted? We give upper and lower bounds which, unfortunately, are widely separated. The main results of this chapter are Theorem 10 and 13.

4.1 Introduction

For a DFA $D = (Q, \Sigma, \delta, q_0, F)$, the set of states, input alphabet, transition function, set of final states, and initial state are denoted by Q, Σ, δ, F, and q_0, respectively. Let $L(D)$ denote the language accepted by D. As usual, we write $\delta(q, wa) = \delta(\delta(q, w), a)$ for a word w.

Let L be an interesting language, such as the language of primitive words, or the language of non-palindromes. We are interested in the following kind of question: *given that an automaton M of N states accepts a member of L, what is a good bound on the length $\ell(N)$ of the shortest word accepted?*

For example, Ito et al. [21] proved that if L is the language of primitive words, then $\ell(N) \leq 3N - 3$. Horváth et al. [19] proved that if L is the language of non-palindromes, then $\ell(N) \leq 3N$. For additional results along these lines, see [4].

[1] The contents of this chapter are taken largely verbatim from Mousavi and Shallit [26].

In this paper we address two open questions left unanswered in [4], corresponding to the case where L is the language of k-power-free (resp., overlap-free) words. For these words we give a class of DFAs of N states for which the shortest k-power (resp., overlap) is of length $N^{\frac{1}{4}(\log N)+O(1)}$. For overlaps over a binary alphabet we give an upper bound of $2^{O(N^{4N})}$.

We state the following basic result without proof.

Proposition 4. *Let $D = (Q, \Sigma, \delta, q_0, F)$ be a (deterministic or nondeterministic) finite automaton. If $L(D) \neq \emptyset$, then D accepts at least one word of length smaller than $|Q|$.*

4.2 Special cases

In this section, we study the cases of the original problem where the shortest word in L that is accepted by M has an additional property. Proving upper bounds for these special cases are easier. In fact, we are not aware of any good upper bound for the original problem. The first special case we study is when the shortest repetition-free word accepted is also circularly repetition-free as defined in Chapter 3. The second case we study is when the shortest repetition-free word is a linearly recurrent. In the former case we prove a linear upper bound, and in the latter case we prove an exponential upper bound. Before we start proving the main theorem of this section, we recall an important theorem of automata theory.

We recall that a relation R on a set S is a subset of $S \times S$. We denote by xRy the fact that $(x, y) \in R$. A relation R is an equivalence relation if R is reflexive, symmetric, and transitive. Index of an equivalence relation R is the number of equivalence classes of R.

One important equivalence relations in formal languages is the Myhill-Nerode relation. For a language $L \subseteq \Sigma^*$ the Myhill-Nerode relation R_L on Σ^* is defined as follows

$$R_L = \{(x, y) | xz \in L \iff yz \in L \text{ for all } z \in \Sigma^*\}.$$

For example for $L = a(a + b)^*$, we have $(a, b) \notin R_L$ since $a \in L$ but $b \notin L$, whereas clearly $(a, aa) \in R_L$. It is easy to see that R_L is an equivalence relation. The famous Myhill-Nerode Theorem states a necessary and sufficient condition of when L is a regular language.

Theorem 6 (Myhill-Nerode)**.** *The language L is regular if and only if R_L is of finite index. Furthermore, if $L = L(M)$ where M is a DFA with N states and R_L has index n, then $n \leq N$.*

Now we can state the main theorem of this section which enables us to prove upper bound on the length of the shortest k-power-free word accepted by a DFA in special cases.

Theorem 7. *Suppose w is the shortest k-power-free word for some $k \geq 2$, accepted by DFA M with N states. Suppose also that there exist an integer l and words $p_1, q_1, p_2, q_2, \ldots, p_l, q_l$ such that*

$$p_1 \prec_p p_2 \prec_p \cdots \prec_p p_l,$$
$$w = p_1 q_1 = p_2 q_2 = \cdots = p_l q_l,$$

and $p_i q_j$ for all $i < j$ are k-power-free, then $l \leq N$.

Proof. Let $L = L(M)$. We show that the index of R_L is $\geq l$. The theorem then follows immediately using the Myhill-Nerode Theorem. To prove that R_L has at least l equivalence classes, we show that $(p_i, p_j) \notin R_L$ for all $i \neq j$. The relation R_L is symmetric, so without loss of generality suppose that $i < j$. We know that $p_i q_j$ is k-power-free and that $|p_i q_j| < |w|$. Based on the assumption that w is the shortest k-power-free word in L, we get that $p_i q_j \notin L$. Now since $p_j q_j = w \in L$, we get that $(p_i, p_j) \notin R_L$ by definition of Myhill-Nerode relation. □

The next theorem states a linear upper bound on the length of the shortest repetition-free word accepted by a DFA when the word is also circularly repetition-free.

Theorem 8. *Let M be a DFA with N states, and let w be the shortest k-power-free word in $L(M)$. If w is circularly k-power-free, then $|w| \leq N$.*

Proof. Let $l = |w|$ and p_i be the prefix of length i of w. Since w is circularly k-power-free, the words $p_i q_j$, for $i < j$, are all k-power-free. Therefore $(p_i, p_j) \notin R_{L(M)}$ and the conditions of Theorem 7 hold. Thus the theorem follows immediately. □

Next we consider the case where the shortest repetition-free word is linearly recurrent. We say w is c-linearly recurrent if for every factor x of w the distance between two consecutive occurrences of x in w is $\leq c|x|$.

Theorem 9. *Let M be a DFA with N states, and let w be the shortest k-power-free word in $L(M)$. If w is c-linearly recurrent, then $|w| < (1 + c)^N$.*

Proof. First note that if we take $w = p_1q_1 = p_2q_2 = \cdots = p_lq_l$ such that p_i is a proper suffix of p_{i+1}, for all i, then the conditions in Theorem 7 are all satisfied. The reason is that $w = p_jq_j$ is k-power-free and therefore all factors of p_jq_j, including p_iq_j, are k-power-free. Thus we have $(p_i, p_j) \notin R_{L(M)}$.

Let l be the integer for which

$$(1+c)^{l-1} \leq |w| \tag{4.1}$$

$$(1+c)^{l} > |w|. \tag{4.2}$$

Let $p_1 = w[0]$. Since w is c-linearly recurrent, there exists p_2 of length $\leq 1 + c$ such that $p_1 \prec_s p_2$. Likewise, there exist $p_3, \ldots, p_l$ such that $p_1 \prec_p p_2 \prec_p p_3 \cdots \prec_p p_l \prec_p w$ and $p_1 \prec_s p_2 \prec_s p_3 \cdots \prec_s p_l$. Note that $|p_l| \leq (1+c)^{l-1}$.

Now using Theorem 7, we get that $l \leq N$. On the other hand, using (4.2), we get that $\log_{1+c} |w| < l$. Thus we can write $|w| < (1+c)^N$. □

4.3 Lower bound

In this section, we construct an infinite family of DFAs such that the shortest k-power-free word accepted is rather long, as a function of the number of states. Up to now only a linear bound was known.

For a word w of length n and $i \geq 1$, let

$$\mathrm{cyc}_i(w) = w[i..n-1]\, w[0..i-2]$$

denote w's ith cyclic shift to the left, followed by removing the last symbol. Also define

$$\mathrm{cyc}_0(w) = w[0..n-2].$$

For example, we have

$$\mathrm{cyc}_2(\texttt{recompute}) = \texttt{computer},$$
$$\mathrm{cyc}_4(\texttt{richly}) = \texttt{lyric}.$$

We call each $\mathrm{cyc}_i(w)$ a *partial conjugate* of w, which is not a reflexive, symmetric, or transitive relation.

A word w is a *simple k-power* if it is a k-power and it contains no k-power as a proper factor.

We start with a few lemmas.

Lemma 8. *Let $w = p^k$ be a simple k-power. Then the word p has $|p|$ distinct conjugates.*

Proof. By contradiction. If p^k is a simple k-power, then p is a primitive word. Suppose that $p = uv = xy$ such that $x \prec_p u$ and $vu = yx$. Without loss of generality, we can assume that $xv \neq \epsilon$. Then there exists a word $t \neq \epsilon$ such that $u = xt$ and $y = tv$. From $vu = yx$ we get

$$vxt = tvx.$$

Using a theorem of Lyndon and Schützenberger [22], we get that there exists $z \neq \epsilon$ such that

$$vx = z^i$$

$$t = z^j$$

for some positive integers i, j. So $yx = z^{i+j}$. Hence $p = xy$ is not primitive, a contradiction. □

Lemma 9. *Let w be a simple k-power of length n. Then we have*

$$\text{cyc}_i(w) = \text{cyc}_j(w) \text{ iff } i \equiv j \pmod{\frac{n}{k}}. \tag{4.3}$$

Proof. Let $w = p^k$. If $i \equiv i' \pmod{\frac{n}{k}}$ and $i' < \frac{n}{k}$, then

$$\text{cyc}_i(w) = (p[i'..\frac{n}{k} - 1]\, p[0..i' - 1])^{k-1}\, \text{cyc}_{i'}(p).$$

Similarly, if $j \equiv j' \pmod{\frac{n}{k}}$ and $j' < \frac{n}{k}$, then

$$\text{cyc}_j(w) = (p[j'..\frac{n}{k} - 1]\, p[0..j' - 1])^{k-1}\, \text{cyc}_{j'}(p).$$

So if $i' = j'$, we get $\text{cyc}_i(w) = \text{cyc}_j(w)$. On the other hand, if $i' \neq j'$, we get

$$p[i'..\frac{n}{k} - 1]\, p[0..i' - 1] \neq p[j'..\frac{n}{k} - 1]\, p[0..j' - 1]$$

using Lemma 8, and hence $\text{cyc}_i(w) \neq \text{cyc}_j(w)$.

□

Lemma 10. *All conjugates of a simple k-power are simple k-powers.*

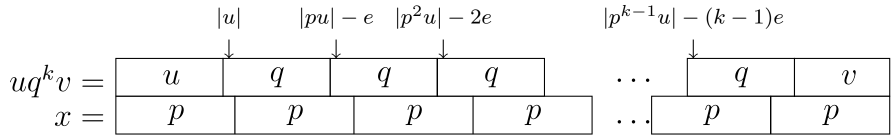


Figure 4.1: Starting positions of the occurrences of q inside x

Proof. By contradiction. Let $w = p^k$ be a simple k-power, and let $z \neq w$ be a conjugate of w. Clearly z is a k-power. Suppose z contains q^k and $z \neq q^k$. Thus $|q| < |p|$. Since w is simple $q^k \not\preceq w = p^k$. The word $x = p^{k+1}$ contains z as a factor. So $x = uq^k v$, for some words $u, v \preceq p$.
Note that u and v are nonempty and not equal to p since $q^k \not\preceq p^k$. Letting $e := |p| - |q|$, and considering the starting positions of the occurrences of q in x (see Figure 4.1), we can write

$$x\left[|p^i u| - ie..|p^i u| - (i-1)e - 1\right] = x\left[|p^j u| - je..|p^j u| - (j-1)e - 1\right]$$

for every $0 \leq i, j < k$. Since p is a period of x, we can write

$$x\left[|u| - ie..|u| - (i-1)e - 1\right] = x\left[|u| - je..|u| - (j-1)e - 1\right]$$

which means $x[u - (k-1)e..u + e - 1] \preceq w$ is a k-power. Therefore w contains a k-power other than itself, a contradiction. □

Corollary 1. *Partial conjugates of simple k-powers are k-power-free.*

The next lemma shows that there are infinitely many simple k-powers over a binary alphabet for $k > 2$. We also show that there are infinitely many simple squares over a ternary alphabet, using a result of Currie [8].

Lemma 11.

(i) *Let $p = \mathbf{t}[0..2^n - 1]$ where $n \geq 0$. For every $k > 2$, the word p^k is a simple k-power.*

(ii) *There are infinitely many simple squares over a ternary alphabet.*

Proof.

(i) By induction on n. For $n = 0$ we have $p^k = 0^k$ which is a simple k-power. Suppose $n > 0$. To get a contradiction, suppose that there exist words u, v, x with $uv \neq \epsilon$

and $x \neq \epsilon$ such that $p^k = ux^kv$. Note that $|x| < |p|$, so $|uv| \geq k$. Without loss of generality, we can assume that $|v| \geq \lceil \frac{k}{2} \rceil \geq 2$. Let $q = \mathbf{t}[0..2^{n-1} - 1]$. We know that

$$p^k = \mu(q^k).$$

We can write

$$w = ux^k \preceq_p \mu(q^{k-1}q[0..|q| - 2]).$$

Since μ is k-power-free, the word $q^{k-1}q[0..|q| - 2]$ contains a k-power. Hence q^k contains at least two k-powers, a contradiction.

(ii) Currie [8] proved that over a ternary alphabet, for every $n \geq 18$, there is a word p of length n such that all its conjugates are squarefree. Such squarefree words are called *circularly squarefree words*.

We claim that for every circularly squarefree word p, the word p^2 is a simple square. To get a contradiction, let q^2 be the smallest square in p^2. So there exist words u, y with $uy \neq \epsilon$ such that $p^2 = uq^2y$. We have $|q^2| > |p|$ since p is circularly squarefree. Therefore, if we let $p = uv = xy$, then $|x| > |u|$ and $|v| > |y|$. So there exists t such that $x = ut$ and $v = ty$. We can assume $|t| < |q|$, since otherwise $|t| = |q|$ and $|uy| = 0$, a contradiction. Now since $q^2 = vx = tyut$, we get that q begins and ends with t, which means $t^2 \prec q^2$. Therefore p^2 has a smaller square than q^2, a contradiction.

□

Next we show how to construct arbitrarily long simple k-powers from smaller ones. Fix $k = 2$ (resp., $k \geq 3$) and $m = 3$ (resp., $m = 2$). Let $w_1 \in \Sigma_m^*$ be a simple k-power. Using the previous lemma, there are infinitely many choices for w_1. Let w_1 be of length n. Define $w_{i+1} \in \Sigma_{m+i}^*$ for $i \geq 1$ recursively by

$$w_{i+1} = \mathrm{cyc}_0(w_i)a_i\, \mathrm{cyc}_{n^{i-1}}(w_i)a_i\, \mathrm{cyc}_{2n^{i-1}}(w_i)a_i \cdots \mathrm{cyc}_{(n-1)n^{i-1}}(w_i)a_i \qquad (4.4)$$

where $a_i = m + i - 1$. The next lemma states that w_i, for $i \geq 1$, is a simple k-power. Therefore, using Corollary 1, each word $\mathrm{cyc}_0(w_i)$ is k-power-free. For $i \geq 1$, it is easy to see that

$$|w_i| = n|w_{i-1}| = n^i. \qquad (4.5)$$

Lemma 12. *For every $i \geq 1$, the word w_i is a simple k-power.*

Proof. By induction on i. The word w_1 is a simple k-power. Now suppose that w_i is a simple k-power for some $i \geq 1$. Using Lemma 9, we have $\mathrm{cyc}_{jn^{i-1}}(w_i) = \mathrm{cyc}_{(j+\frac{n}{k})n^{i-1}}(w_i)$, since $\frac{|w_i|}{k} = \frac{n^i}{k}$.

We get that w_{i+1} is a k-power since

$$w_{i+1} = (\mathrm{cyc}_0(w_i)a_i \,\mathrm{cyc}_{n^{i-1}}(w_i)a_i \,\mathrm{cyc}_{2n^{i-1}}(w_i)a_i \cdots \mathrm{cyc}_{(\frac{n}{k}-1)n^{i-1}}(w_i)a_i)^k.$$

We now claim that w_{i+1} is a simple k-power. To see this, suppose that w_{i+1} contains a k-power y^k such that $w_{i+1} \neq y^k$.

If y contains more than one occurrence of a_i, then $y = ua_i \,\mathrm{cyc}_j(w_i)a_i v$ for some words u, v and an integer j. Since $y^2 = ua_i \,\mathrm{cyc}_j(w_i)a_i vua_i \,\mathrm{cyc}_j(w_i)a_i v \preceq w_{i+1}$, using (4.4) and Lemma 9, we get that

$$|y| = \left|\mathrm{cyc}_j(w_i)a_i vua_i\right| \geq \frac{n}{k}n^i = \frac{|w_{i+1}|}{k},$$

and hence $y^k = w_{i+1}$, a contradiction.

If y contains just one a_i, then $y = ua_i v$ for some words u, v which contain no a_i. So $y^k = u(avu)^{k-1}av$ for $a = a_i$. Therefore vu is a partial conjugate of w_i. However the distance between two equal partial conjugates of w_i in w_{i+1} is longer than just one letter, using (4.4) and Lemma 9.

Finally, if y contains no a_i, then a partial conjugate of w_i contains a k-power, which is impossible due to Corollary 1. □

To make our formulas easier to read, we define $a_0 = w_1[n-1]$.

Theorem 10. *For $i \geq 1$, there is a DFA D_i with $2^{i-1}(n-1)+2$ states such that $\mathrm{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$.*

Proof. Define $D_1 = (Q_1, \Sigma_{a_1}, \delta_1, q_{1,0}, F_1)$ where

$$\begin{aligned}
Q_1 &:= \{q_{1,0}, q_{1,1}, q_{1,2}, \ldots, q_{1,n-1}, q_d\},\\
F_1 &:= \{q_{1,n-1}\},\\
\delta_1(q_{1,j}, w_1[j]) &:= q_{1,j+1} \text{ for } 0 \leq j < n-1,
\end{aligned}$$

and the rest of the transitions go to the dead state q_d. Clearly we have $|Q_1| = n+1$ and $L(D_1) = \{\mathrm{cyc}_0(w_1)\}$.

We define $D_i = (Q_i, \Sigma_{a_i}, \delta_i, q_{1,0}, F_i)$ for $i \geq 2$ recursively. We recall that $a_i = m + i - 1$ for $i \geq 1$ and $a_0 = w_1[n-1]$. For the rest of the proof s and t denote (possibly empty) sequences of integers and j denotes a single integer (a sequence of length 1). We use integer sequences as subscripts of states in Q_i. For example, $q_{1,0}$, $q_{s,j}$, and $q_{s,1,t}$ might denote states of D_i. For $i \geq 1$, define

$$Q_{i+1} := Q_i \cup \{q_{i+1,t} : q_t \in (Q_i - F_i) - \{q_d\}\}, \tag{4.6}$$

$$F_{i+1} := \{q_{i+1,i,t} : \delta_i(q_{i,t}, c) = q_{1,n-1} \text{ for some } c \in \Sigma_{a_i}\}, \tag{4.7}$$

$$\text{if } q_t \in Q_i \text{ and } c \in \Sigma_{a_i}, \text{ then } \delta_{i+1}(q_t, c) := \delta_i(q_t, c) \tag{4.8}$$

$$\text{if } q_t, q_s \in (Q_i - F_i) - \{q_d\}, c \in \Sigma_{a_i}, \text{ and } \delta_i(q_t, c) = q_s, \quad \text{then } \delta_{i+1}(q_{i+1,t}, c) := q_{i+1,s} \tag{4.9}$$

$$\text{if } q_t \in F_i, \text{ then } \delta_{i+1}(q_t, a_i) := q_{1,1} \text{ and } \delta_{i+1}(q_t, a_{i-1}) := q_{i+1,1,0} \tag{4.10}$$

$$\text{if } i > 1, q_{i+1,t} \notin F_{i+1}, \text{ and } \delta_i(q_t, a_{i-1}) = q_{1,j}, \quad \text{then } \delta_{i+1}(q_{i+1,t}, a_i) := q_{1,j+1} \tag{4.11}$$

and finally for the special case of $i = 1$,

$$\delta_2(q_{2,1,j}, a_1) := q_{1,j+2} \text{ for } 0 \leq j < n-2. \tag{4.12}$$

The rest of the transitions, not indicated in (4.8)–(4.12), go to the dead state q_d. Figure 4.2b depicts D_2 and D_3. Using (4.6), we have $|Q_{i+1}| = 2|Q_i| - 2 = 2^i(n-1) + 2$ by a simple induction.

An easy induction on i proves that $|F_i| = 1$. So let f_i be the appropriate integer sequence for which $F_i = \{q_{f_i}\}$. Using (4.8)–(4.12), we get that for every $1 \leq j < n$, there exists exactly one state $q_t \in Q_i$ for which $\delta_i(q_t, a_{i-1}) = q_{1,j}$.

By induction on i, we prove that for $i \geq 2$ if $\delta_i(q_t, a_{i-1}) = q_{1,j}$, then

$$x_1 = \text{cyc}_{(j-1)n^{i-2}}(w_{i-1}), \tag{4.13}$$

$$x_2 = w_i[0..jn^{i-1} - 2], \tag{4.14}$$

$$x_3 = w_i[(j-1)n^{i-1}..n^i - 2]. \tag{4.15}$$

are the shortest k-power-free words for which

$$\delta_i(q_{1,j-1}, x_1) = q_t, \tag{4.16}$$

$$\delta_i(q_{1,0}, x_2) = q_t, \tag{4.17}$$

$$\delta_i(q_{1,j-1}, x_3) = q_{f_i}. \tag{4.18}$$

In particular, from (4.15) and (4.18), for $j = 1$, we get that $\text{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$.

The fact that our choices of x_1, x_2, and x_3 are k-power-free follows from the fact that proper factors of simple k-powers are k-power-free. For $i = 2$ the proofs of (4.16)–(4.18) are easy and left to the readers.

Suppose that (4.16)–(4.18) hold for some $i \geq 2$. Let us prove (4.16)–(4.18) for $i + 1$. Suppose that

$$\delta_{i+1}(q_t, a_i) = q_{1,j}. \tag{4.19}$$

First we prove that the shortest k-power-free word x for which

$$\delta_{i+1}(q_{1,j-1}, x) = q_t,$$

is $x = \text{cyc}_{(j-1)n^{i-1}}(w_i)$.

If $q_t \in Q_i$, from (4.10) and (4.19), we have

$$\begin{aligned} &q_t = q_{f_i}, \text{ and} \\ &\delta_{i+1}(q_t, a_i) = q_{1,1}. \end{aligned}$$

By induction hypothesis, the $\text{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$. In other words, we have $\delta_i(q_{1,0}, \text{cyc}_0(w_i)) = q_{f_i} = q_t$, which can be rewritten using (4.8) as $\delta_{i+1}(q_{1,0}, \text{cyc}_0(w_i)) = q_t$.

Now suppose $q_t \notin Q_i$. Then by (4.11) and (4.19), we get that there exists t' such that $q_{t'} \in Q_i$ and

$$\begin{aligned} &t = i + 1, t'; \\ &\delta_i(q_{t'}, a_{i-1}) = q_{1,j-1}. \end{aligned}$$

From the induction hypothesis, i.e., (4.17) and (4.18), we can write

$$\delta_i(q_{1,0}, w_i[0..(j-1)n^{i-1} - 2]) = q_{t'}, \tag{4.20}$$

$$\delta_i(q_{1,j-1}, w_i[(j-1)n^{i-1}..n^i - 2]) = q_{f_i}. \tag{4.21}$$

In addition $w_i[0..(j-1)n^{i-1} - 2]$ and $w_i[(j-1)n^{i-1}..n^i - 2]$ are the shortest k-power-free transitions from $q_{1,0}$ to $q_{t'}$ and from $q_{1,j-1}$ to q_{f_i} respectively. Using (4.8), we can rewrite (4.20) and (4.21) for δ_{i+1} as follows:

$$\delta_{i+1}(q_{1,0}, w_i[0..(j-1)n^{i-1} - 2]) = q_{t'}, \tag{4.22}$$

$$\delta_{i+1}(q_{1,j-1}, w_i[(j-1)n^{i-1}..n^i - 2]) = q_{f_i}. \tag{4.23}$$

Note that from (4.9) and (4.22), we get

$$\delta_{i+1}(q_{i+1,1,0}, w_i[0..(j-1)n^{i-1}-2]) = q_{i+1,t'} = q_t. \tag{4.24}$$

We also have $\delta_{i+1}(q_{f_i}, a_i) = q_{i+1,1,0}$, using (4.10). So together with (4.23) and (4.24), we get

$$\delta_{i+1}(q_{1,j-1}, \mathrm{cyc}_{(j-1)n^{i-1}}(w_i)) = q_t$$

and $\mathrm{cyc}_{(j-1)n^{i-1}}(w_i)$ is the shortest k-power-free transition from $q_{1,j-1}$ to q_t.

The proofs of (4.17) and (4.18) are similar. □

In what follows, all logarithms are to the base 2.

Corollary 2. *For infinitely many N, there exists a* DFA *with N states such that the shortest k-power-free word accepted is of length $N^{\frac{1}{4}\log N+O(1)}$.*

Proof. Let $i = \lfloor \log n \rfloor$ in Theorem 10. Then $D = D_i$ has

$$N = 2^{\lfloor \log n \rfloor -1}(n-1) + 2 = \Omega(n^2)$$

states. In addition, the shortest k-power-free word in $L(D)$ is $\mathrm{cyc}_0\left(w_{\lfloor \log n \rfloor}\right)$. Now, using (4.5) we can write

$$\left|\mathrm{cyc}_0(w_{\lfloor \log n \rfloor})\right| = n^{\lfloor \log n \rfloor} - 1.$$

Suppose $2^t \leq n < 2^{t+1}-1$, so that $t = \lfloor \log n \rfloor$ and Then $\log N = 2t + O(1)$, so $\frac{1}{4}(\log N)^2 = t^2 + O(t)$. On the other hand $\log |w| = \lfloor \log n \rfloor (\log n) = t(t+O(1)) = t^2 + O(t)$. Now $2^{O(t)} = n^{O(1)} = N^{O(1)}$, and the result follows. □

Remark 1. *The same bound holds for overlap-free words. To do so, we define a simple overlap as a word of the form $axaxa$ where $axax$ is a simple square. In our construction of the DFAs, we use complete conjugates of $(ax)^2$ instead of partial conjugates.*

Remark 2. *The D_i in Theorem 10 are defined over the growing alphabet Σ_{m+i-1}. However, we can fix the alphabet to be Σ_{m+1}. For this purpose, we introduce w'_i which is quite similar to w_i:*

$$w'_1 = w_1,$$
$$w'_{i+1} = \mathrm{cyc}_0(w'_i)b_i\, \mathrm{cyc}_{n^{i-1}}(w'_i)b_i\, \mathrm{cyc}_{2n^{i-1}}(w'_i)b_i \cdots \mathrm{cyc}_{(n-1)n^{i-1}}(w'_i)b_i,$$

where $b_i = mc_im$ such that c_i is (any of) the shortest nonempty k-power-free word over Σ_m not equal to $c_1, \ldots, c_{i-1}$. Clearly we have $|b_i| \leq |b_{i-1}| + 1 = O(i)$, and hence $w'_i = \Theta(n^i)$.

One can then prove Lemma 12 and Theorem 10 for w_i' with minor modifications of the argument above. In particular, we construct DFA D_i' *that accepts* $\text{cyc}_0(w_i')$ *as the shortest k-power-free word accepted, and a* D_i' *that is quite similar to* D_i. *In particular, they have asymptotically the same number of states.*

4.4 Upper bound for overlap-free words

In this section, we prove an upper bound on the length of the shortest overlap-free word accepted by a DFA D over a binary alphabet.

Let $L = L(D)$ and let R be the set of overlap-free words over Σ_2^*. Carpi [7] defined a certain operation Ψ on binary languages, and proved that $\Psi(R)$ is regular. We prove that $\Psi(L)$ is also regular, and hence $\Psi(L) \cap \Psi(R)$ is regular. Then we apply Proposition 4 to get an upper bound on the length of the shortest word in $\Psi(L) \cap \Psi(R)$. This bound then gives us an upper bound on the length of the shortest overlap-free word in L.

Let $H = \{\epsilon, 0, 1, 00, 11\}$. Carpi defines maps

$$\Phi_l, \Phi_r : \Sigma_{25} \to H$$

such that for every pair $h, h' \in H$, one has

$$h = \Phi_l(a), h' = \Phi_r(a)$$

for exactly one letter $a \in \Sigma_{25}$.

For every word $w \in \Sigma_{25}^*$, define $\Phi(w) \in \Sigma_2^*$ inductively by

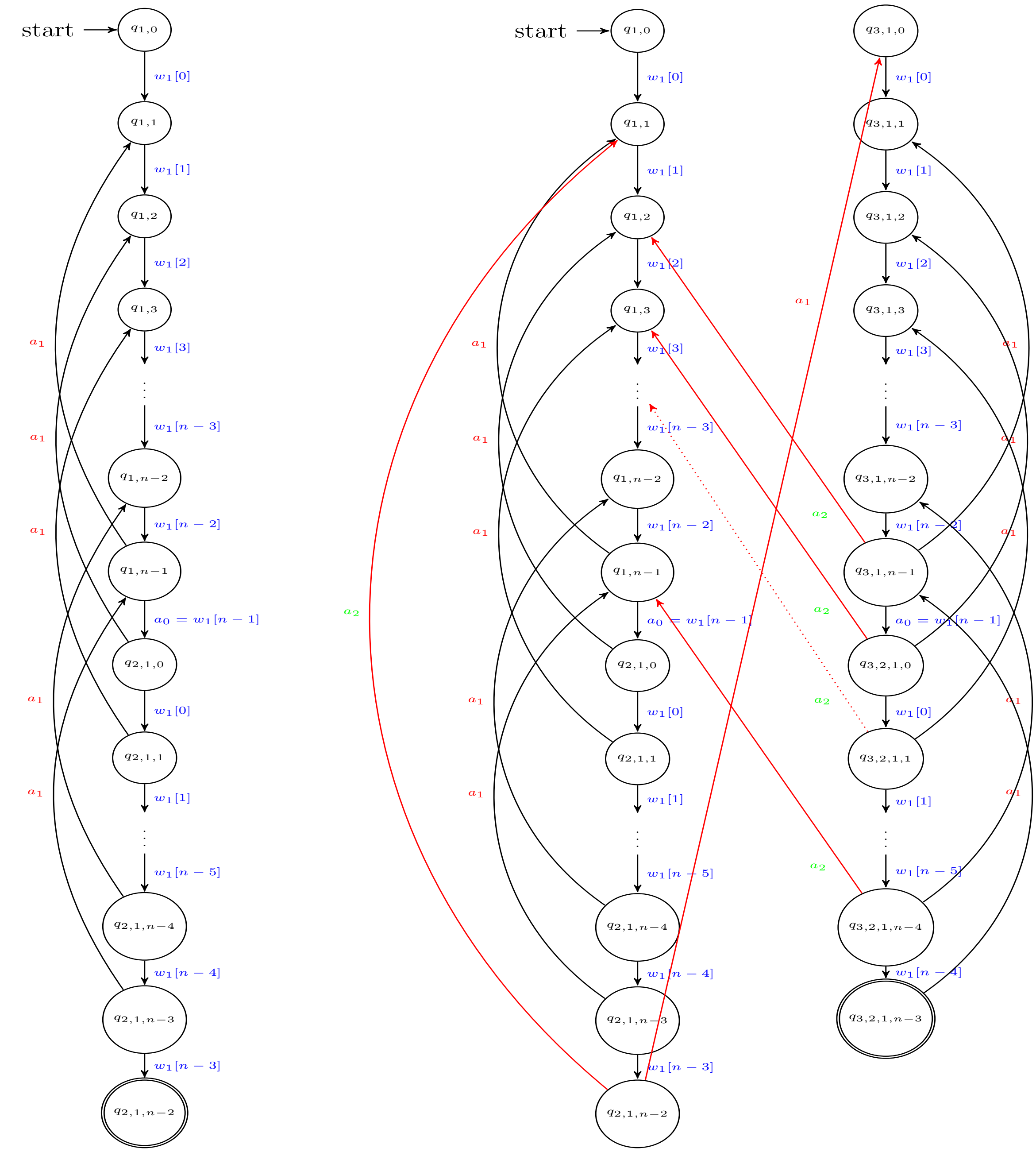
$$\Phi(\epsilon) = \epsilon, \Phi(aw) = \Phi_l(a)\mu(\Phi(w))\Phi_r(a) \qquad (w \in \Sigma_{25}^*, a \in \Sigma_{25}). \tag{4.25}$$

Expanding (4.25) for $w = a_0 a_1 \cdots a_{n-1}$, we get

$$\Phi_l(a_0)\mu(\Phi_l(a_1)) \cdots \mu^{n-1}(\Phi_l(a_{n-1}))\mu^{n-1}(\Phi_r(a_{n-1})) \cdots \mu(\Phi_r(a_1))\Phi_r(a_0). \tag{4.26}$$

For $L \subseteq \Sigma_2^*$ define $\Psi(L) = \bigcup_{x \in L} \Phi^{-1}(x)$. Based on the decomposition of Restivo and Salemi [29] for finite overlap-free words, the language $\Psi(\{x\})$ is always nonempty for an overlap-free word $x \in \Sigma_2^*$. The next theorem is due to Carpi [7].

Theorem 11. $\Psi(R)$ *is regular.*



(a) Transition Diagram of D_2

(b) Transition Diagram of D_3

Figure 4.2: Transition Diagrams

Carpi constructed a DFA A with less than 400 states that accepts $\Psi(R)$. We prove that Ψ preserves regular languages.

Theorem 12. *Let $D = (Q, \Sigma_2, \delta, q_0, F)$ be a* DFA *with N states, and let $L = L(D)$. Then $\Psi(L)$ is regular and is accepted by a* DFA *with at most N^{4N} states.*

Proof. Let $\iota : Q \to Q$ denote the identity function, and define $\eta_0, \eta_1 : Q \to Q$ as follows

$$\eta_i(q) = \delta(q, i) \text{ for } i = 0, 1. \tag{4.27}$$

For functions $\zeta_0, \zeta_1 : Q \to Q$, and a word $x = b_0 b_1 \cdots b_{n-1} \in \Sigma_2^*$, define $\zeta_x = \zeta_{b_{n-1}} \circ \cdots \circ \zeta_{b_1} \circ \zeta_{b_0}$. Therefore we have $\zeta_y \circ \zeta_x = \zeta_{xy}$. Also by convention $\zeta_\epsilon = \iota$. So for example $x \in L(D)$ if and only if $\eta_x(q_0) \in F$.

We create DFA $D' = (Q', \Sigma_{25}, \delta', q_0', F')$ where

$$Q' = \{[\kappa, \lambda, \zeta_0, \zeta_1] : \kappa, \lambda, \zeta_0, \zeta_1 : Q \to Q\},$$
$$\delta'([\kappa, \lambda, \zeta_0, \zeta_1], a) = \left[\zeta_{\Phi_l(a)} \circ \kappa, \lambda \circ \zeta_{\Phi_r(a)}, \zeta_1 \circ \zeta_0, \zeta_0 \circ \zeta_1\right].$$

Also let

$$q_0' = [\iota, \iota, \eta_0, \eta_1],$$
$$F' = \{[\kappa, \lambda, \zeta_0, \zeta_1] : \lambda \circ \kappa(q_0) \in F\}. \tag{4.28}$$

We can see that $|Q'| = N^{4N}$. We claim that D' accepts $\Psi(L)$. Indeed, on input w, the DFA D' simulates the behavior of D on $\Phi(w)$.

Let $w = a_0 a_1 \cdots a_{n-1} \in \Sigma_{25}^*$, and define

$$\Phi_1(w) = \Phi_l(a_{a_0}) \mu(\Phi_l(a_1)) \cdots \mu^{n-1}(\Phi_l(a_{n-1})),$$
$$\Phi_2(w) = \mu^{n-1}(\Phi_r(a_{n-1})) \cdots \mu(\Phi_r(a_1)) \Phi_r(a_0).$$

Using (4.26), we can write

$$\Phi(w) = \Phi_1(w)\Phi_2(w).$$

We prove by induction on n that

$$\delta'(q_0', w) = \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^n(0)}, \eta_{\mu^n(1)}\right]. \tag{4.29}$$

For $n = 0$, we have $\Phi(w) = \Phi_1(w) = \Phi_2(w) = \epsilon$. So

$$\delta'(q_0', \epsilon) = q_0' = [\iota, \iota, \eta_0, \eta_1] = [\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^0(0)}, \eta_{\mu^0(1)}].$$

So we can assume (4.29) holds for some $n \geq 0$. Now suppose $w = a_0 a_1 \cdots a_n$ and write

$$\begin{aligned}
&\delta'(q_0', a_0 a_1 \cdots a_n) \\
&= \delta'(\delta'(q_0', a_0 a_1 \cdots a_{n-1}), a_n) \\
&= \delta'\left(\left[\eta_{\Phi_1(w[0..n-1])}, \eta_{\Phi_2(w[0..n-1])}, \eta_{\mu^n(0)}, \eta_{\mu^n(1)}\right], a_n\right) \\
&= \Big[\eta_{\mu^n(\phi_l(a_n))} \circ \eta_{\Phi_1(w[0..n-1])}, \eta_{\Phi_2(w[0..n-1])} \circ \eta_{\mu^n(\phi_r(a_n))}, \\
&\qquad\qquad \eta_{\mu^n(1)} \circ \eta_{\mu^n(0)}, \eta_{\mu^n(0)} \circ \eta_{\mu^n(1)}\Big] \\
&= \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^{n+1}(0)}, \eta_{\mu^{n+1}(1)}\right], \qquad (4.30)
\end{aligned}$$

and equality (4.30) holds because

$$\begin{aligned}
&\Phi_1(w[0..n-1])\mu^n(\phi_l(a_n)) = \Phi_1(w), \\
&\mu^n(\phi_r(a_n))\Phi_2(w[0..n-1]) = \Phi_2(w), \\
&\mu^n(0)\mu^n(1) = \mu^n(01) = \mu^n(\mu(0)) = \mu^{n+1}(0), \text{ and similarly} \\
&\mu^n(1)\mu^n(0) = \mu^{n+1}(1).
\end{aligned}$$

Finally, using (4.28), we have

$$\begin{aligned}
w \in L(D') &\iff \delta'(q_0', w) = \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \zeta_0, \zeta_1\right] \in F' \\
&\iff \eta_{\Phi_2(w)} \circ \eta_{\Phi_1(w)}(q_0) \in F \\
&\iff \Phi(w) = \Phi_1(w)\Phi_2(w) \in L(D).
\end{aligned}$$

□

Theorem 13. *Let $D = (Q, \Sigma_2, \delta, q_0, F)$ be a* DFA *with N states. If D accepts at least one overlap-free word, then the length of the shortest overlap-free word accepted is $2^{O(N^{4N})}$.*

Proof. Let $L = L(D)$. Using Theorem 12, there exists a DFA D' with N^{4N} states that accepts the language $\Psi(L)$.

Since $\Psi(R)$ is regular and is accepted by a DFA with at most 400 states, we see that

$$K = \Psi(L) \cap \Psi(R)$$

is regular and is accepted by a DFA with $O\left(N^{4N}\right)$ states.

Since L accepts an overlap-free word, the language K is nonempty. Using Proposition 4, we see that K contains a word w of length $O\left(N^{4N}\right)$.

Therefore $\Phi(w)$ is an overlap-free word in L. By induction, one can easily prove that $|\Phi(w)| = O\left(2^{|w|}\right)$. Hence we have $|\Phi(w)| = 2^{O(N^{4N})}$. □

Chapter 5

Open Problems

We state a number of open problems in this chapter.

In Chapter 3, we introduced the quantity $\exp(w)$. The naive algorithm to compute $\exp(w)$ takes cubic time. Badkobeh et al. [5] give a linear algorithm that computes $\exp(w)$.

Problem 1. *How fast can we compute the circular exponent,* $\mathrm{cexp}(w)$*?*

We introduce the notion of RT_k in Chapter 3.

Problem 2. *Prove or disprove any of the following equalities*

$$\begin{aligned}
\mathrm{RT}_2(4) &= \mathrm{RTC}(4) = \frac{5}{2}\ , \\
\mathrm{RT}_2(5) &= \mathrm{RTC}(5) = \frac{105}{46}\ , \textit{ and} \\
\mathrm{RT}_2(n) &= \mathrm{RTC}(n) = 1 + \mathrm{RT}(n) = \frac{2n-1}{n-1} \textit{ for } n \geq 6.
\end{aligned}$$

Problem 3. *Compute* $\mathrm{RT}_k(n)$ *for all* k *and* n.

In Chapter 4, we gave a doubly exponential upper bound on the length of the shortest binary overlap-free word accepted by a DFA. We are not aware of any upper bound on the length of the shortest k-power-free word accepted by a DFA.

Problem 4. *Either prove a sharp lower bound on the length of the shortest binary overlap-free word accepted by a DFA, or improve the upper bound.*

Problem 5. *Prove any upper bound on the length of the shortest* k*-power-free word accepted by a DFA, for any* k.

References

[1] A. Aberkane and J. D. Currie. There exist binary circular $5/2^+$ power free words of every length. *Electronic J. Combinatorics*, 11(10), 2004.

[2] J.-P. Allouche and J. O. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and Their Applications, Proceedings of SETA '98*, pages 1–16. Springer-Verlag, 1999.

[3] N. Alon, J. Grytczuk, M. Hauszczak, and O. Riordan. Nonrepetitive colorings of graphs. *Random Structures & Algorithms*, 21(3-4):336–346, 2002.

[4] T. Anderson, J. Loftus, N. Rampersad, N. Santean, and J. Shallit. Detecting palindromes, patterns and borders in regular languages. *Info. Comput.*, 207:1096–1118, 2009.

[5] G. Badkobeh, M. Crochemore, and C. Toopsuwan. Computing the maximal-exponent repeats of an overlap-free string in linear time. In Liliana Caldern-Benavides, Cristina Gonzlez-Caro, Edgar Chvez, and Nivio Ziviani, editors, *String Processing and Information Retrieval*, volume 7608 of *Lecture Notes in Computer Science*, pages 61–72. Springer-Verlag, 2012.

[6] F.-J. Brandenburg. Uniformly growing k-th power-free homomorphisms. *Theoret. Comput. Sci.*, 23:69–82, 1983.

[7] A. Carpi. Overlap-free words and finite automata. *Theoret. Comput. Sci.*, 115:243–260, 1993.

[8] A. Carpi. There are ternary circular square-free words of length n for $n \geq 18$. *Electronic J. Combinatorics*, 9(10), 2002.

[9] A. Carpi. On Dejean's conjecture over large alphabets. *Theoret. Comput. Sci.*, 385:137–151, 2007.

[10] C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer-Verlag, 1997.

[11] M. Crochemore. Sharp characterizations of squarefree morphisms. *Theoret. Comput. Sci.*, 18:221–226, 1982.

[12] M. Crochemore. Linear searching for a square in a word. In Jan Paredaens, editor, *Automata, Languages and Programming*, volume 172 of *Lecture Notes in Computer Science*, pages 137–137. Springer-Verlag, 1984.

[13] J. Currie and N. Rampersad. A proof of Dejean's conjecture. *Math. Comp.*, 80:1063–1070, 2011.

[14] F. Dejean. Sur un théorème de Thue. *J. Combin. Theory. Ser. A*, 13:90–99, 1972.

[15] I. A. Gorbunova. Repetition threshold for circular words. *Electronic J. Combinatorics*, 19(11), 2012.

[16] T. Harju. *On cyclically overlap-free words in binary alphabets*, pages 125–130. Springer-Verlag, 1986.

[17] T. Harju and J. Karhumäki. Morphisms. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 439–510. Springer-Verlag, 1997.

[18] T. Harju and D. Nowotka. Cyclically repetition-free words on small alphabets. *Inform. Process. Lett.*, 110:591–595, 2010.

[19] S. Horváth, J. Karhumäki, and J. Kleijn. Results concerning palindromicity. *J. Inf. Process. Cybern. EIK*, 23:441–451, 1987.

[20] L. Ilie, P. Ochem, and J. Shallit. A generalization of repetition threshold. *Theoret. Comput. Sci.*, 345:359–269, 2005.

[21] M. Ito, M. Katsura, H. J. Shyr, and S. S. Yu. Automata accepting primitive words. *Semigroup Forum*, 37:45–52, 1988.

[22] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.*, 9:289–298, 1962.

[23] M. Mohammad-Noori and J. D. Currie. Dejean's conjecture and sturmian words. *European Journal of Combinatorics*, 28(3):876–890, 2007.

[24] J. Moulin-Ollagnier. Proof of Dejean's conjecture for alphabets with $5, 6, 7, 8, 9, 10$ and 11 letters. *Theoret. Comput. Sci.*, 95:187–205, 1992.

[25] H. Mousavi and J. Shallit. Repetition avoidance in circular factors. In M. P. Bal and O. Carton, editors, *Developments in Language Theory*, volume 7907 of *Lecture Notes in Computer Science*, pages 384–395. Springer-Verlag, 2013.

[26] H. Mousavi and J. Shallit. Shortest repetition-free words accepted by automata. In H. Jurgensen and R. Reis, editors, *Descriptional Complexity of Formal Systems*, volume 8031 of *Lecture Notes in Computer Science*, pages 182–193. Springer-Verlag, 2013.

[27] J. J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots. *Discrete Applied Mathematics*, 7:297–311, 1984.

[28] M. Rao. Last cases of Dejean's conjecture. *Theoret. Comput. Sci.*, 412:3010–3018, 2011.

[29] A. Restivo and S. Salemi. Overlap free words on two symbols. In M. Nivat and D. Perrin, editors, *Automata on Infinite Words*, volume 192 of *Lecture Notes in Computer Science*, pages 198–206. Springer-Verlag, 1985.

[30] J. Shallit. *A Second Course in Formal Languages and Automata Theory*. Cambridge University Press, New York, NY, USA, 2008.

[31] M. Sipser. *Introduction to the Theory of Computation*. Cengage Learning, 3rd edition, 2012.

[32] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 7:1–22, 1906. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[33] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 1:1–67, 1912. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

[34] S. Yu. Regular languages. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 41–110. Springer-Verlag, 1997.