CHAPTER 2

# *Square-Free Words and Idempotent Semigroups*

## 2.0. Introduction

The investigation of words includes a series of combinatorial studies with rather surprising conclusions that can be summarized roughly by the following statement: Each sufficiently long word over a finite alphabet behaves locally in a regular fashion. That is to say, an arbitrary word, subject only to the constraint that it be sufficiently long, possesses some regularity. This claim becomes meaningful only if one specifies the kind of regularities that are intended, of course. The discovery and the analysis of these *unavoidable regularities* constitute a major topic in the combinatorics of words. A typical example is furnished by van der Waerden's theorem.

It should not be concluded that any sufficiently long word is globally regular. On the contrary, the existence of unavoidable regularities leads to the dual question of avoidable regularities: properties not automatically shared by all sufficiently long words. For such a property there exist infinitely many words (finiteness of the alphabet is supposed) that do not satisfy it. The present chapter is devoted mainly to the study of one such property.

A *square* is a word of the form $uu$, with $u$ a nonempty word. A word contains a square if one of its factors is a square; otherwise, the word is called *square-free*. For instance, *abcacbacbc* contains the square *acbacb*, and *abcacbabcb* is square-free. The answer to the question of whether every sufficiently long word contains a square is no, provided the alphabet has at least three letters. As will be shown, the existence of infinitely many square-free words is equivalent to the existence of a square-free word that is infinite (on the right). The formalism of infinite words has the advantage of allowing concise descriptions. Furthermore, infinite iteration of a morphism is a natural and simple way to construct infinite words, and this method applies especially to the construction of infinite square-free words.

We start with the investigation of a famous infinite word, called after its discoverers the word of Thue–Morse. This word contains squares, but it is

cube-free and even has stronger properties. Then we turn to the study of infinite square-free words. A simple coding of the Thue–Morse infinite word gives an example of an infinite square-free word. We then establish a general result of Thue that gives other infinite square-free words.

A more algebraic framework can be used for the theory of square-free words. Consider the monoid $M = A^* \cup 0$ obtained by adjoining a zero to the free monoid $A^*$. Next consider the congruence over $M$ generated by the relations

$$uu \approx 0 \qquad (u \in A^+).$$

The fact that there exist infinitely many square-free words can be rephrased: The quotient monoid $M/\approx$ is infinite, provided $A$ has at least three letters. A natural analogue is to consider the free idempotent monoid, that is, the quotient of $A^*$ by the congruence generated by

$$uu \sim u \qquad (u \in A^+).$$

We will show, in contrast to the previous result, that for each finite alphabet $A$, the quotient monoid $A^*/\sim$ is finite.

Many results, extensions, and generalizations concerning the problems just sketched are not included in the text. They are stated as exercises or briefly mentioned in the Notes, which also contain some bibliographic remarks.

## 2.1. Preliminaries

Before defining infinite words, let us fix some notations concerning distinct occurrences of a word as a factor in a given word. Let $A$ be an alphabet, $w \in A^+$. Let $u$ be a nonempty word having two distinct occurrences as a factor in $w$. Then there are words $x, y, x', y' \in A^*$ such that

$$w = xuy = x'uy', \quad x \neq x'.$$

These two occurrences of $u$ either overlap or are consecutive or are disjoint. More precisely, we may suppose $|x| < |x'|$. Then three possibilities arise (see Figure 2.1).

(i) $|x'| > |xu|$. In this case, $x' = xuz$ for some $z \in A^+$, and $w = xuzuy'$. The occurrences of $u$ are *disjoint*.

(ii) $|x'| = |xu|$. This implies that $x' = xu$, and consequently $w = xuuy'$ contains a square. The occurrences of $u$ are *adjacent*.

(iii) $|x'| < |xu|$. The two occurrences of $u$ are said to *overlap*. The following lemma gives a more precise description of this case.

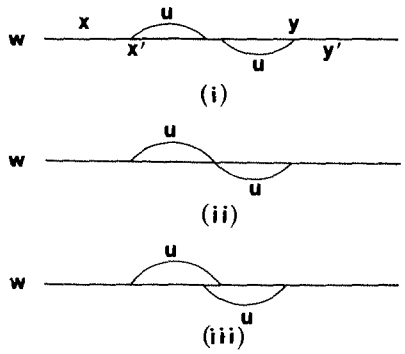



**Figure 2.1.** Two occurrences of $u$ in $w$: (i) disjoint occurrences, (ii) adjacent occurrences, (iii) overlapping occurrences

LEMMA 2.1.1. *Let $w$ be a word; then $w$ contains two overlapping occurrences of a word $u \neq 1$ iff $w$ contains a factor of the form $avava$, with $a$ a letter and $v$ a word.*

*Proof.* Assume first $w = xuy = x'uy'$, where the occurrences of $u$ overlap. Then $|x| < |x'| < |xu| < |x'u|$. Consequently

$$x' = xs, \qquad xu = x'z, \qquad x'u = xut$$

for some nonempty words $s, z, t$, whence

$$u = sz = zt. \tag{2.1.1}$$

Let $a$ be the first letter of $s$, and therefore also of $z$ by Eq. (2.1.1). Set $s = av$, $z = az'$. Then by (2.1.1) $u = avaz'$ and

$$w = xsuy' = xavavaz'y'.$$

Conversely, if $avava$ is a factor of $w$, then $u = ava$ clearly has two overlapping occurrences in $w$. ■

A word of the form $avava$, with $a$ a letter, is said to *overlap*. Thus, according to the lemma, a word has two overlapping occurrences of a word iff it contains an overlapping factor.

We now turn to the definition of infinite words. Let $A$ be an alphabet. An *infinite word* on $A$ is a function

$$\mathbf{a}: \mathbb{N} \to A.$$

We use the following notation

$$\mathbf{a} = \mathbf{a}(0)\mathbf{a}(1)\cdots\mathbf{a}(n)\cdots,$$

and also

$$\mathbf{a} = a_0 a_1 \cdots a_n \cdots,$$

where $a_n = \mathbf{a}(n)$ is a letter. The left factor of length $k \geqslant 0$ of a is

$$\mathbf{a}^{[k]} = a_0 a_1 \cdots a_{k-1}.$$

For $u \in A^*$, we write $u < \mathbf{a}$ whenever $u = \mathbf{a}^{[k]}$ for $k = |u|$. Then clearly $\mathbf{a} = u\mathbf{b}$ where $\mathbf{b}(m) = \mathbf{a}(m+k)$ for all $m \geqslant 0$. A *factor* of $u$ is any word in $A^*$ that occurs in a. In the sequel, by a word we always mean a finite word.

Infinite words are useful when one deals with properties $P$ of (finite) words having a special feature, namely that $P(xuy)$ implies $P(u)$ for all words $x, u, y$. In other terms, if $L_P$ is the set of words for which $P$ holds, then $L_P$ contains the factors of its elements. Note that this holds for the set of square-free words. When $P$ satisfies this condition we say that $P$ is *stable for factors*. Given an infinite word a, we say that a has the property $P$ if each factor of **a** satisfies $P$. Thus it is meaningful to speak about infinite square-free words.

LEMMA 2.1.2. *Let $A$ be a finite alphabet and let $P$ be a property of elements of $A^*$ that is stable for factors. Then the two following conditions are equivalent:*

(i) *The set $L_P$ of words $w$ in $A$ such that $P(w)$ is infinite.*
(ii) *There exists an infinite word on $A$ with property $P$.*

A particular case is the assertion mentioned in the introduction, namely that the existence of infinitely many square-free words is equivalent to the existence of an infinite square-free word.

*Proof.* Clearly (ii) implies (i). Conversely, if $L = L_P$ is infinite, the finiteness of $A$ implies that infinitely many words in $L$ start with a same letter, say $a_0$. Set $L_0 = L \cap a_0 A^*$. Assume by induction that there are letters $a_0, a_1, \ldots, a_n$ such that $L_n = L \cap a_0 a_1 \cdots a_n A^*$ is infinite. Then among the sets $(L \cap a_0 a_1 \cdots a_n b A^*)_{b \in A}$ at least one is infinite. Choose one letter $a_{n+1}$ such that $L \cap a_0 a_1 \cdots a_n a_{n+1} A^*$ is infinite.

Thus there exists a sequence $a_0, a_1, \ldots, a_n, \ldots$ of letters in $A$ such that $L \cap a_0 a_1 \cdots a_n A^*$ is infinite for each $n \geqslant 0$. Define $\mathbf{a}$: $\mathbb{N} \to A$ by $\mathbf{a}(n) = a_n$. Then each factor of a is a factor of a word in $L$, thus is itself in $L$. ∎

Sometimes a simpler method can be applied to construct infinite words from finite ones. (Note that the proof of the previous lemma gives such a construction.)

Let $w_0, w_1, \ldots, w_n, \ldots$ be a sequence of words in $A^*$ of unbounded length such that each $w_{n-1}$ is a left factor of $w_n$. Then define an infinite word $\mathbf{a}$ on $A$ by

$$\mathbf{a}^{[k]} = w_n, \qquad k = |w_n|, \qquad n \geqslant 0.$$

The definition is consistent because $\mathbf{a}^{[k]}$ is a left factor of all $w_m$, $m \geqslant n$. The infinite word defined in this way is called the *limit* of $(w_n)_{n \geqslant 0}$ and is denoted by

$$\mathbf{a} = \lim w_n.$$

Consider the following important special case. Let

$$\alpha : A^* \to A^*$$

be a morphism verifying

(i) $\alpha(a) \neq 1$ for $a \in A$, (2.1.2)
(ii) there exists a letter $a_0$ such that

$$\alpha(a_0) = a_0 u \quad \text{for some} \quad u \in A^+. \tag{2.1.3}$$

Then for each $n \geqslant 0$,

$$\alpha^{n+1}(a_0) = \alpha^n(a_0 u) = \alpha^n(a_0)\alpha^n(u).$$

Thus each $\alpha^n(a_0)$ is a proper left factor of $\alpha^{n+1}(a_0)$, and therefore the limit of the sequence $(\alpha^n(a_0))_{n \geqslant 0}$ exists. We denote this limit by $\alpha^\omega(a_0)$:

$$\alpha^\omega(a_0) = \lim \alpha^n(a_0),$$

and we say that it is obtained by *iterating* $\alpha$ *on* $a_0$.

With these notations $\alpha$ can be extended to infinite words by setting, for $\mathbf{b} = b_0 b_1 \cdots b_n \cdots$

$$\alpha(\mathbf{b}) = \alpha(b_0)\alpha(b_1)\cdots\alpha(b_n)\cdots.$$

Condition (i) ensures that $\alpha(\mathbf{b})$ is indeed an infinite word. Observe that

$$\alpha(\mathbf{a}) = \mathbf{a} \quad \text{for} \quad \mathbf{a} = \alpha^\omega(a_0). \tag{2.1.4}$$

In other terms, a is a fixed point for $\alpha$. Indeed set $\mathbf{b} = \alpha(\mathbf{a})$. For each left factor $u$ of $\mathbf{a}$, the word $\alpha(u)$ is a left factor of $\alpha(\mathbf{a})$. Thus each $\alpha^n(a_0)$, $n \geqslant 1$, is a left factor of $\mathbf{b}$, and $\mathbf{b}$ starts with $a_0$ by (ii). Consequently $\mathbf{b} = \lim \alpha^n(a_0) = \mathbf{a}$; this proves Eq. (2.1.4).

## 2.2. The Infinite Words of Thue–Morse

In this section a special infinite word is defined and its properties are studied. The main result is that this infinite word has no overlapping factor.

In this section $A$ denotes the fixed two-letter alphabet $A = \{a, b\}$. Define a morphism

$$\mu: A^* \to A^*$$

by

$$\mu(a) = ab, \qquad \mu(b) = ba.$$

Then $\mu$ satisfies conditions (2.1.2), (2.1.3) for $a_0 = a$ and also for $a_0 = b$. Consequently, iteration of $\mu$ on $a$ and on $b$ yields two infinite words

$$\mathbf{t} = \mu^\omega(a), \qquad \bar{\mathbf{t}} = \mu^\omega(b).$$

By definition, $\mathbf{t}$ is the *infinite word of Thue–Morse*. Computation gives

$$\begin{aligned} \mu(a) &= ab & \mu(b) &= ba \\ \mu^2(a) &= abba & \mu^2(b) &= baab \\ \mu^3(a) &= abbabaab & \mu^3(b) &= baababba \end{aligned}$$

$$\mathbf{t} = abbabaabbaababbabaababbaabbabaab\cdots$$

$$\bar{\mathbf{t}} = baababbaabbabaababbabaabbaababba\cdots$$

There are several properties relating the words $\mu^n(a), \mu^n(b), n \geqslant 0$. Consider the morphism

$$w \mapsto \bar{w}$$

defined by

$$\bar{a} = b, \quad \bar{b} = a$$

Thus $\bar{w}$ is obtained from $w$ by replacing each $a$ by $b$ and conversely. Of course $\bar{\bar{w}} = w$.

PROPOSITION 2.2.1. *Define* $u_0 = a$, $v_0 = b$ *and for* $n \geqslant 0$

$$u_{n+1} = u_n v_n, \qquad v_{n+1} = v_n u_n.$$

*Then for all* $n \geqslant 0$

(i) $u_n = \mu^n(a), \qquad v_n = \mu^n(b).$

(ii) $v_n = \bar{u}_n, \qquad u_n = \bar{v}_n.$

(iii) $u_{2n}, v_{2n}$ *are palindromes and* $\tilde{u}_{2n+1} = v_{2n+1}$.

*Proof.* The proofs are by induction. The initial step is always clear. Formula (i) follows from

$$u_{n+1} = u_n v_n = \mu^n(a)\mu^n(b) = \mu^{n+1}(a),$$
$$v_{n+1} = v_n u_n = \mu^n(b)\mu^n(a) = \mu^{n+1}(b);$$

next (ii) follows from

$$v_{n+1} = v_n u_n = \bar{u}_n \bar{v}_n = \overline{u_n v_n} = \bar{u}_{n+1}, \bar{v}_{n+1} = \bar{\bar{u}}_{n+1} = u_{n+1};$$

finally for (iii), observe that for $k > 0$

$$\tilde{u}_k = (u_{k-1}v_{k-1})^{\sim} = \tilde{v}_{k-1}\tilde{u}_{k-1}.$$

If $k$ is odd (resp. even) this implies

$$\tilde{u}_k = v_{k-1}u_{k-1} = v_k \text{ (resp. } \tilde{u}_k = u_{k-1}v_{k-1} = u_k). \qquad \blacksquare$$

There exists an interesting definition of

$$\mathbf{t} = t_0 t_1 \cdots t_n \cdots$$

that is independent of the morphism $\mu$. First let, for $n \geqslant 0$, $d_2(n)$ be the number of 1's in the binary expansion of $n$. Then we have the following proposition.

PROPOSITION 2.2.2. *For each* $n \geqslant 0$,

$$t_n = \begin{cases} a & \text{if } d_2(n) \equiv 0 \mod 2 \\ b & \text{if } d_2(n) \equiv 1 \mod 2 \end{cases} \tag{2.2.1}$$

*Proof.* Note that by (2.1.4) we have

$$\mathbf{t} = \mu(\mathbf{t}) = \mu(t_0)\mu(t_1)\cdots\mu(t_n)\cdots$$

and therefore $\mu(t_n) = t_{2n}t_{2n+1}$ for $n \geqslant 0$. By the definition of $\mu$, this implies

$$t_{2n} = t_n, \qquad t_{2n+1} = \tilde{t}_n \quad (n \geqslant 0). \tag{2.2.2}$$

Formula (2.2.1) holds for $n = 0$. Thus let $n > 0$. If $n = 2m$, then $t_n = t_m$ by (2.2.2), and $d_2(n) = d_2(m)$. Thus (2.2.1) holds in this case. If $n = 2m+1$, then $t_n = \tilde{t}_m$ and $d_2(n) \equiv 1 + d_2(m) \bmod 2$. Therefore (2.2.1) holds in this case too. $\blacksquare$

The inspection of **t** shows that **t** is not square-free. However, we will prove the following:

THEOREM 2.2.3. *The infinite word t has no overlapping factor.*

COROLLARY 2.2.4. *The infinite word t is cube-free.*

The proof of the theorem uses two lemmas.

LEMMA 2.2.5. *Let* $X = \{ab, ba\}$; *if* $x \in X^*$, *then* $axa \notin X^*$ *and* $bxb \notin X^*$.

*Proof.* By induction on $|x|$. If $|x| = 0$, then indeed $aa, bb \notin X^*$. Let $x \in X^*$, $x \neq 1$ and suppose $u = axa \in X^*$ (the case $bxb \in X^*$ is similar). Then $u = x_1 x_2 \cdots x_r$, with $x_1, \ldots, x_r \in X$; consequently $x_1 = ab$ and $x_r = ba$. Thus $u = abyba$ with $y = x_2 \cdots x_{r-1} \in X^*$. But now by induction $x = byb$ is not in $X^*$, contrary to the assumption. ■

LEMMA 2.2.6. *Let* $w \in A^+$. *If w has no overlapping factor, then* $\mu(w)$ *has no overlapping factor.*

*Proof.* Assume that $\mu(w)$ has an overlapping factor for some $w \in A^*$. We show that $w$ also has an overlapping factor.

By asumption, there are $x, v, y \in A^*$, $c \in A$ with

$$\mu(w) = xcvcvcy$$

Note that $|cvcvc|$ is odd, but $\mu(w) \in X^*$ with $X = \{ab, ba\}$: therefore $|\mu(w)|$ is even and $|xy|$ is odd. Thus

- Either: $|x|$ is even, and $x, cvcv, cy \in X^*$,
- Or: $|x|$ is odd, and $xc, vcvc, y \in X^*$.

This implies that $|v|$ is odd, since otherwise we get from $cvcv \in X^*$ (resp. $vcvc \in X^*$) that both $v, cvc$ are in $X^*$, which contradicts Lemma 2.2.5.

In the case $|x|$ is even, it follows that $cv$ is in $X^*$ and $w = rsst$ with $\mu(r) = x, \mu(s) = cv, \mu(t) = cy$. But then $s$ and $t$ start with the same letter $c$ and $ssc$ is an overlapping factor in $w$.

In the case $|x|$ is odd, similarly $vc \in X^*$, and $w = rsst$ with $\mu(r) = xc, \mu(s) = vc, \mu(t) = y$. Here $r$ and $s$ end with $c$ and $css$ is an overlapping factor in $w$. ■

*Proof of Theorem 2.2.3.* Assume that **t** has an occurrence of an overlapping factor. Then it occurs in a left factor $\mu^k(a)$ for some $k > 0$. On the other hand, since $a$ has no overlapping factor, by iterated application of Lemma 2.2.6 no $\mu^n(a) (n \geq 0)$ has an overlapping factor. Contradiction. ■

## 2.3. Infinite Square-Free Words

The infinite word of Thue–Morse has square factors. In fact, the only square-free words over two letters $a$ and $b$ are

$$a, b, ab, ba, aba, bab.$$

On the contrary, there exist infinite square-free words over three letters. This will now be demonstrated.

As before let $A = \{a, b\}$, and let $B = \{a, b, c\}$. Define a morphism

$$\delta: B^* \to A^*$$

by setting

$$\delta(c) = a, \qquad \delta(b) = ab, \qquad \delta(a) = abb$$

For any infinite word $\mathbf{b}$ on $B$,

$$\delta(\mathbf{b}) = \delta(b_0)\delta(b_1)\cdots\delta(b_n)\cdots$$

is a well-defined infinite word on $A$ starting with the letter $a$. Conversely, consider an infinite word a on $A$ without overlapping factors and starting with $a$. Then $\mathbf{a}$ can be factored as

$$\mathbf{a} = y_0 y_1 \cdots y_n \cdots \tag{2.3.1}$$

with each $y_n \in \{a, ab, abb\} = \delta(B)$. Indeed, each $a$ in a is followed by at most two $b$ since $bbb$ is overlapping, and then followed by a new $a$. Moreover, the factorization (2.3.1) is unique. Thus there exists a unique infinite word $\mathbf{b}$ on $B$ such that $\delta(\mathbf{b}) = \mathbf{a}$.

THEOREM 2.3.1. *Let* $\mathbf{a}$ *be an infinite word on* $A$ *starting with* $a$, *and without overlapping factor, and let* $\mathbf{b}$ *be the infinite word over* $B$ *such that* $\delta(\mathbf{b}) = \mathbf{a}$; *then* $\mathbf{b}$ *is square-free.*

*Proof.* Assume the contrary. Then $\mathbf{b}$ contains a square, say $uu$. Let $d$ be the letter following $uu$ in one of its occurrences in $\mathbf{b}$. Then $\delta(uud)$ is a factor of $\mathbf{a}$. Since $\delta(u) = av$ for some $v \in A^*$ and $\delta(d)$ starts with $a$, $\mathbf{a}$ contains the factor $avava$. Contradiction. ∎

By applying the theorem to the Thue–Morse word $\mathbf{t}$, we obtain an infinite square-free word $\mathbf{m}$ over the three letter alphabet $B$ such that $\delta(\mathbf{m}) = \mathbf{t}$. This infinite word is

$$\mathbf{m} = abcacbabcbacabcacbacabcbabcacbabcbacabcbabc\cdots$$

Note that the converse of Theorem 2.3.1 is false: There are square-free infinite words $\mathbf{b}$ over $B$ such that $\delta(\mathbf{b})$ has overlapping factors (see Problem 2.3.7). There are several alternative ways to obtain the word $\mathbf{m}$. We quote just one.

PROPOSITION 2.3.2. *Define a morphism* $\varphi$: $B^* \to B^*$ *(with* $B = \{a, b, c\}$*) by* $\varphi(a) = abc$, $\varphi(b) = ac$, $\varphi(c) = b$. *Then* $\mathbf{m} = \varphi^\omega(a)$.

The proof is left as an exercise.

There exist other constructions that allow one to obtain more systematically infinite square-free words. We now present one of them. In the sequel of this paragraph, $A, B \cdots$ are again arbitrary alphabets.

First we introduce a new notion. A morphism $\alpha$: $A^* \to B^*$ is *square-free* if $\alpha(A) \neq \{1\}$ and if $\alpha(w)$ is a square-free word for each square-free word $w$. Thus a square-free morphism preserves square-free words. The first condition is present simply to avoid uninteresting discussions on the square-freeness of the empty word. A square-free morphism $\alpha$ from $A^*$ into itself produces by iteration only square-free words, when one starts with a square-free word, or simply with a letter. Thus a square-free morphism usually gives an infinite set of square-free words. Note that the morphism $\varphi$ of Proposition 2.3.2 is *not* square-free since

$$\varphi(aba) = abcacabc$$

contains a square. The following theorem gives sufficient conditions for a morphism to be square-free.

THEOREM 2.3.3. *Let* $\alpha$: $A^* \to B^*$ *be a morphism with* $\alpha(A) \neq \{1\}$ *such that*

(i) $\alpha(u)$ *is square-free for each square-free word of length* $\leqslant 3$,
(ii) *No* $\alpha(a)$ *is a proper factor of an* $\alpha(b)$ $(a, b$ *in* $A)$.

*Then* $\alpha$ *is a square-free morphism.*

*Proof.* First we note that $\alpha(a) \neq 1$ for each $a \in A$; otherwise if $\alpha(a) = 1$ let $b \in A$ be a letter with $x = \alpha(b) \neq 1$. Then $bab$ is square-free, but $\alpha(bab) = xx$ violates condition (i). Next $\alpha$ is injective on $A$: if $\alpha(a) = \alpha(b)$, then $\alpha(ab)$ is a square, consequently $a = b$ by (i). Furthermore, $X = \alpha(A)$ is a biprefix code by (ii). Now we prove the following claim.

*Claim: If* $\alpha(a_1 a_2 \cdots a_n) = x\alpha(a)y$ *for* $a, a_i \in A$, $x, y \in B^*$, *then* $a = a_j$ *for some* $j$, $x = \alpha(a_1 \cdots a_{j-1})$, $y = \alpha(a_{j+1} \cdots a_n)$.

The claim is clear for $n=1$ by (ii). Arguing by induction on $n$, assume $n>1$. If

$$|x\alpha(a)| \leqslant |\alpha(a_1 a_2 \cdots a_{n-1})|$$

or

$$|\alpha(a)y| \leqslant |\alpha(a_2 \cdots a_n)|,$$

the claim follows by the induction hypothesis. Thus, we may assume that both

$$|x\alpha(a)| > |\alpha(a_1 a_2 \cdots a_{n-1})|$$

and

$$|\alpha(a)y| > |\alpha(a_2 \cdots a_n)|.$$

Consequently, $y$ is a proper right factor of $\alpha(a_n)$, and $x$ is a proper left factor of $\alpha(a_1)$:

$$\alpha(a_1) = xu, \qquad \alpha(a_n) = vy$$

for some $u, v$ in $B^+$, and

$$\alpha(a) = u\alpha(a_2)\cdots\alpha(a_{n-1})v.$$

By (ii), this implies $n=2$ and $\alpha(a)=uv$.

The words $\alpha(a_1 a) = xuuv$ and $\alpha(aa_n) = uvvy$ are not square-free. According to (i), $a_1 = a = a_n$, whence

$$xu = uv = vy.$$

The first equation shows that $|x| = |v|$. In view of $xu = vy$, it follows that $x = v$. Consequently $vu = uv$. By a result of Chapter 1, $\alpha(a) = uv$ is not a primitive word and thus is not square-free. This contradicts condition (i) and proves the claim.

Now we prove the theorem. Assume the conclusion is false. Then there is a shortest square-free word $w \in A^+$ such that $\alpha(w)$ contains a square, say

$$\alpha(w) = yuuz \qquad \text{with } u \neq 1.$$

Set $w = a_1 a_2 \cdots a_n$, $v_i = \alpha(a_i)$ $(a_i \in A)$. By condition (i), one has $n \geqslant 4$. Next $y$ is a proper left factor of $v_1$ and $z$ is a proper right factor of $v_n$ since $w$ was chosen shortest. Also $yu$ is not a left factor of $v_1$ since otherwise $v_2 v_3$ is a

factor of $u$, hence of $v_1$, violating condition (ii). For the same reason, $uz$ is not a right factor of $v_n$. Thus there is an index $j$ $(1<j<n)$ and a factorization

$$v_j = st$$

such that (see Figure 2.2(i))

$$yu = v_1 \cdots v_{j-1}s, \qquad uz = tv_{j+1} \cdots v_n.$$

We may assume $s \neq 1$, since otherwise $j-1 \neq 1$ and we can replace $v_j$ by $v_{j-1}$. Next, define $y'$ and $z'$ by

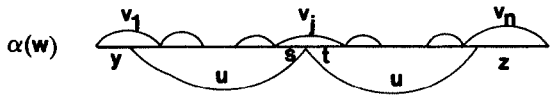
$$v_1 = yy', \qquad v_n = z'z.$$

As mentioned before, $y'$ and $z'$ are nonempty. Further (see Figure 2.2(ii))

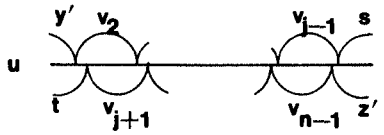
$$\begin{aligned} u &= y'v_2 \cdots v_{j-1}s, \\ u &= tv_{j+1} \cdots v_{n-1}z', \end{aligned} \tag{2.3.2}$$

Now, we derive a contradiction by showing that $w$ contains a square. Consider first the case where $yt = 1$. In this case, $v_1 = y'$, $v_j = s$, whence by Eqs. (2.3.2)

$$u = v_1v_2 \cdots v_{j-1}v_j = v_{j+1} \cdots v_{n-1}z'.$$




(i)




(ii)

**Figure 2.2.** Occurrence of $uu$ in $\alpha(w)$: (i) localization of $uu$, (ii) double factorization of $u$.

Since $\alpha(A)$ is a prefix code, this implies $v_1 = v_{j+1}, \ldots, v_{j-1} = v_{n-1}$, $v_j = z'$, and since $v_j = z' \leq v_n$, we have $v_n = v_j$. Thus $w = (a_1 \cdots a_j)^2$ is a square.

Next consider the case $yt \neq 1$. Multiplying (2.3.2) by $s$ and $z$, and by $y$ and $t$, gives

$$suz = (sy')v_2 \cdots v_{j-1}(sz) = v_j v_{j+1} \cdots v_{n-1} v_n \tag{2.3.3}$$

$$yut = v_1 v_2 \cdots v_{j-1} v_j = (yt) v_{j+1} \cdots v_{n-1}(z't) \tag{2.3.4}$$

Consider Eq. (2.3.3) first. Then the claim can be applied to each of the $v_2, \ldots, v_{j-1}$. Consequently $a_2 \cdots a_{j-1}$ is a factor of $a_j a_{j+1} \cdots a_{n-1} a_n$. Since $s \neq 1$, $a_2 \cdots a_{j-1}$ is neither a left nor a right factor of $a_j, \ldots, a_n$; thus $a_2 \cdots a_{j-1}$ is a factor of $a_{j+1} \cdots a_{n-1}$ and

$$pa_2 \cdots a_{j-1} q = a_{j+1} \cdots a_{n-1} \tag{2.3.5}$$

for some $p, q \in A^*$. Now consider Eq. (2.3.4). As before, $a_{j+1} \cdots a_{n-1}$ is a factor of $a_1 \cdots a_j$, and since neither $yt$ nor $z'$ is the empty word, $a_{j-1} \cdots a_{n-1}$ is a factor of $a_2 \cdots a_{j-1}$. Thus

$$\bar{p} a_{j+1} \cdots a_{n-1} \bar{q} = a_2 \cdots a_{j-1} \tag{2.3.6}$$

for some $\bar{p}, \bar{q} \in A^*$. By (2.3.5) and (2.3.6),

$$\bar{p}pa_2 \cdots a_{j-1} q\bar{q} = \bar{p} a_{j+1} \cdots a_{n-1} \bar{q} = a_2 \cdots a_{j-1}$$

showing that $p = \bar{p} = q = \bar{q} = 1$. Thus setting

$$x = a_2 \cdots a_{j-1} = a_{j+1} \cdots a_{n-1}$$

we have

$$w = a_1 x a_j x a_n \tag{2.3.7}$$

whence by (2.3.3) and (2.3.4)

$$st = v_j = sy', \qquad z'z = v_n = sz, \qquad yy' = v_1 = yt$$

Thus the word

$$(a_1 a_j a_n) = v_1 v_j v_n = yt\, st\, sz$$

is not square-free. By condition (i), $a_1 a_j a_n$ is not square-free. Therefore $a_1 = a_j$ or $a_j = a_n$. In view of (2.3.7), $w$ contains a square. This yields the contradiction. ■

*Example.* A tedious but finite computation shows that the morphism $\alpha$: $A^* \to A^*$ with $A = \{a, b, c\}$ defined by

$$\alpha(a) = abcab, \qquad \alpha(b) = acabcb, \qquad \alpha(c) = acbcacb$$

fulfills the two conditions of Theorem 2.3.3 and therefore is a square-free morphism.

## 2.4. Idempotent Semigroups

Let $A$ be an alphabet having at least three letters. Then there are infinitely many square-free words in $A^*$. As already mentioned in the introduction, this fact can be rephrased as follows. Let $A^* \cup 0$ be the monoid obtained by adjoining a zero to $A$, and consider the congruence $\approx$ generated by

$$uu \approx 0, \quad u \in A^+.$$

Each square-free word constitutes an equivalence class modulo this congruence. Consequently the quotient monoid $A^* \cup 0/\approx$ is infinite.

There is another situation where square-free words can be used. Let $m, n \geqslant 2$ be fixed integers and consider the congruence $\equiv$ over $A^*$ generated by

$$u^m \equiv u^n, \quad u \in A^*. \tag{2.4.1}$$

Once more, each square-free word defines an equivalence class, and thus the monoid $A^*/\equiv$ is infinite. In fact, this result also holds for a two-letter alphabet (Brzozowski, Culik II, and Gabrielian 1971).

These considerations can be placed in the framework of the classical Burnside problem (originally, the Burnside problem was formulated for groups only, but it is easy to state for semigroups also): *Is every finitely generated torsion semigroup finite?* (A torsion semigroup is a semigroup such that each element generates a finite subsemigroup.) We have just seen that the answer is negative in general, and this is due to the existence of infinitely many square-free words. For groups, the answer also is negative (see Chapter 8 in Herstein 1968). Moreover, the groups of exponent $n$—that is, groups where each element has exponent $n$—are in general infinite (see Adjan 1979). The proof uses the fact that there are infinitely many square-free words. For another result on the Burnside problem, see Chapter 7, Section 7.3.

In one special case, surprisingly, the answer is positive. Let $A$ be an arbitrary finite alphabet, and consider the congruence $\sim$ generated by the

relations

$$ww \sim w, \qquad w \in A^*. \tag{2.4.2}$$

The quotient monoid

$$M = A^*/\sim$$

is called the *free idempotent monoid* on $A$; indeed, any element in $M$ is idempotent ($mm = m$), and any finitely generated idempotent monoid is easily seen to be a quotient of a free idempotent monoid.

THEOREM 2.4.1 (Green–Rees). *The free idempotent monoid on $A$ is finite and has exactly*

$$\sum_{k=0}^{n} \binom{n}{k} \prod_{1 \leqslant i \leqslant k} (k-i+1)^{2^i} \tag{2.4.3}$$

*elements, where $n = Card(A)$.*

The numbers (2.4.3) are growing very rapidly. For $n = 0, 1, 2, 3, 4$, they are $1, 2, 7, 160, 332381$.

Before starting the proof, it will be interesting to note the difference between the relations (2.4.1) and (2.4.2). For the congruence defined by (2.4.1), two distinct words can be congruent only if both contain at least one $p$th power, for $p = \min(m, n)$. On the contrary, two distinct square-free words may be congruent for $\sim$. Indeed, the defining relations allow introduction of squares and then dropping of other ones. We give now a nontrivial illustration of this situation by verifying that $x \sim y$ with $x = bacbcabc$ and $y = bacabc$. Both $x$ and $y$ are square-free words, and they are also equivalent. Indeed, note first that with $u = abcaca$, we have (boldfaced factors are those to be reduced) $uy = abc\mathbf{acabacabc} \sim abc\mathbf{aca}bc \sim \mathbf{abcabc} \sim abc$ whence $x = (bacbc)abc \sim bacbcuy = vy$ for $v = bacbcu$.

Next, for $r = bcabacbcacbcbac$, we have

$$\begin{aligned} xr &= ba\mathbf{cbcabcbcab}acbcacbcbac \\ &\sim \mathbf{bacbcabacbca}cbcbac \\ &\sim b\mathbf{acbcacbc}bac \sim ba\mathbf{cbcb}ac \sim \mathbf{bacbac} \sim bac \end{aligned}$$

whence

$$y = bacabc \sim xrabc \sim xs$$

with $s = rabc$. Finally,

$$x \sim vy \sim vyy \sim xy \sim xxs \sim xs \sim y,$$

which proves the claim.

*Proof of Theorem 2.4.1.* Recall from Chapter 1 that for $w \in A^*$,

$$\text{alph}(w) = \{a \in A \mid |w|_a \neq 0\}$$

It is clear that $x \sim y$ implies $\text{alph}(x) = \text{alph}(y)$. First we prove the following claim:

*Claim* (i). *If* $\text{alph}(y) \subset \text{alph}(x)$, *there exists* $u$ *such that* $x \sim xyu$.

This is indeed clear if $y = 1$. Assume $|y| \geqslant 1$, and let $y = y'a$ with $a \in A$. By induction, there is a word $u'$ such that $x \sim xy'u'$. Furthermore, $a \in \text{alph}(x)$, whence $x = zaz'$. Thus for $u = z'y'u'$

$$xyu = zaz'y'az'y'u' \sim zaz'y'u' = xy'u' \sim x.$$

This proves Claim (i).

For $x \in A^+$, let $x'$ be the shortest left factor of $x$ such that $\text{alph}(x') = \text{alph}(x)$. Setting $x' = pa$ for some $p \in A^*$, $a \in A$, we have $\text{alph}(p) = \text{alph}(x) - \{a\}$. Symmetrically, the shortest right factor $x''$ of $x$ with $\text{alph}(x'') = \text{alph}(x)$ has the form $x'' = bq$ for some $b \in A$, $q \in A^*$ and $\text{alph}(q) = \text{alph}(x) - \{b\}$. Thus to $x$ there is associated a quadruple $(p, a, b, q)$. We write this fact $x \triangleq (p, a, b, q)$, and prove:

*Claim* (ii). *If* $x \triangleq (p, a, b, q)$, *then* $x \sim pabq$.

Indeed let, $x = pay = zbq$. Since $\text{alph}(y) \subset \text{alph}(x) = \text{alph}(pa)$, there is by (i) a word $u$ such that $pa \sim payu = xu$. Since $\text{alph}(pa) \subset \text{alph}(bq)$, the dual of (i) shows that there is a word $v$ with $bq \sim vpabq = v\hat{x}$, where $\hat{x} = pabq$. This implies that

$$\hat{x} = pabq \sim xubq = xw$$

for $w = ubq$ and

$$x = zbq \sim zv\hat{x} = t\hat{x}$$

for $t = zv$. Whence

$$x \sim t\hat{x} \sim t\hat{x}\hat{x} \sim x\hat{x} \sim xxw \sim xw \sim \hat{x}.$$

This proves (ii).

In view of Claim (ii), we can show that $M$ is finite as follows. Assume that the finiteness holds for alphabets that have fewer elements than $A$. If $x \triangleq (p, a, b, q)$, then $\text{Card}(\text{alph}(p)) < \text{Card}(A)$ and $\text{Card}(\text{alph}(q)) < \text{Card}(A)$, thus there are only finitely many $p$s and $q$s modulo $\sim$. Since there are only finitely many letters, $M$ itself is finite. In order to compute the number of elements in $M$, we prove the following equivalence.

*Claim* (iii). *Let* $x \triangleq (p, a, b, q)$ *and* $x' \triangleq (a', a', b', q')$; *then* $x \sim x'$ *iff* $p \sim p'$, $a = a'$, $b = b'$, $q \sim q'$.

Suppose first that $p \sim p'$, $a = a'$, $b = b'$, $q \sim q'$. Then $pabq \sim p'a'b'q'$ and $x \sim x'$ by (ii). Suppose now $x \sim x'$. One can assume that $x = \alpha\beta\gamma$, $x' = \alpha\beta^2\gamma$ for some words $\alpha, \beta, \gamma \in A^*$. We distinguish two cases.

*Case 1.* $|\alpha\beta| > |p|$. Setting $x = pay$, we have

$$\alpha\beta = pat, \qquad z = t\gamma$$

for some $t$ in $A^+$. Then $x' = pat\beta\gamma$ and $\mathrm{alph}(p) = \mathrm{alph}(x) - \{a\} = \mathrm{alph}(x') - \{a\}$. Thus by definition $p' = p$ and $a' = a$.

*Case 2.* $|\alpha\beta| \leqslant |p|$. Setting $x = pay$, there is a word $s \in A^*$ such that

$$p = \alpha\beta s, \qquad \gamma = sag.$$

Then $x' = \alpha\beta^2 sag$ and $\mathrm{alph}(\alpha\beta^2 s) = \mathrm{alph}(\alpha\beta s) = \mathrm{alph}(x) - \{a\} = \mathrm{alph}(x') - \{a\}$. Thus by definition $p' = \alpha\beta^2 s$ whence $p' \sim p$, $a = a'$.

The relations $b = b'$, $q \sim q'$ are proved in a symmetric manner.

We now are ready to compute the number of elements in $M = A^*/\sim$. Let $\pi\colon A^* \to M$ be the canonical morphism and let, for $B \subset A$,

$$\overline{B} = \{x \in A^* \mid \mathrm{alph}(x) = B\}.$$

Then $A^*$ is the disjoint union of the sets $\overline{B}$, $B \subset A$. Since $x \sim x'$ implies $\mathrm{alph}(x) = \mathrm{alph}(x')$, each $\overline{B}$ is a union of equivalence classes mod $\sim$, whence $M$ is the disjoint union of the sets $\pi(\overline{B})$, $B \subset A$.

In view of Claim (iii), if $B \neq \varnothing$, there is a bijection

$$\pi(B) \to \bigcup_{a, b \in A} \pi\big(\overline{B - \{a\}}\big) \times \{a\} \times \{b\} \times \pi\big(\overline{B - \{b\}}\big)$$

Thus if $\mathrm{Card}(B) = k \geqslant 1$, and setting $c_k = \mathrm{Card}(\pi(\overline{B}))$, we have

$$c_k = k^2 c_{k-1}^2.$$

Clearly $c_0 = 1$, whence

$$c_k = \prod_{i=1}^{k} (k - i + 1)^{2^i}.$$

Consequently, $M$ being the disjoint union of the $\pi(\overline{B})$,

$$\mathrm{Card}\, M = \sum_{k=0}^{n} \binom{n}{k} c_k.$$

This completes the proof. ∎

## Notes

Axel Thue was the first author to investigate avoidable regularities, especially words without overlapping factors and square-free words. His two papers (Thue 1906, 1912) on this topic contain the definitions of the words **t** and **m**, and the proofs of Theorems 2.2.3 and 2.3.1 as reported here. Theorem 2.3.3 is a slight improvement, due to Bean, Ehrenfeucht, and McNulty (1979), of a result of Thue. The infinite word **t** was discovered independently by Morse (1921, 1938), the square-freeness of **m** was proved by Morse and Hedlund in 1944, Braunholtz in 1963, and Istrail in 1977. Many other papers have been written on infinite square-free words or related topics (Arson 1937; Dean 1965; Gottschalk and Hedlund 1964; Hawkins and Mientka 1956; Leech 1957; Li 1976; Pleasants 1970; Shepherdson 1958; Zech 1958; Dekking 1976; Entringer, Jackson, and Schatz 1974; Ehrenfeucht and Rozenberg 1981; Main and Lorentz 1979; Crochemore 1981). As noted by Hedlund in 1967, some of the work done later is already contained in Thue's papers, which were forgotten for a long time.

One of the problems raised in Thue's 1912 paper that has been significantly developed concerns the distance between two occurrences of a factor in a word. Indeed, an infinite word **a** is square-free iff whenever $xyx$ is a factor of a with $x \neq 1$, then $y \neq 1$. Thus one may define the number

$$e_{\mathbf{a}}(x) = \min\{|y| : xyx \text{ is a factor of a}\}$$

and look for lower bounds for $e_{\mathbf{a}}(x)$. Thue gives an infinite word a over $k$ letters (for each $k \geq 3$) such that $e_{\mathbf{a}}(x) \geq k-2$ for all $x$ occurring twice in a. F. Dejean (1972) improves this inequality. She constructs an infinite word a over three letters such that

$$e_{\mathbf{a}}(x) \geq \tfrac{1}{3}|x|$$

for all factors $x$ occurring twice in $x$. She also shows that this lower bound is optimal. Pansiot, in a forthcoming paper, handles the case of four letters. For more than four letters, the sharp value of the lower bound remains unknown.

Square-free morphisms and more generally $k$th-power-free morphisms are investigated in Bean, Ehrenfeucht and McNulty 1979. Characterizations of square-free morphisms are given in Berstel 1979 and Crochemore 1982. Bean et al. introduce the very interesting concept of so-called avoidable patterns, which are described as follows:

Let $E$ and $A$ be two alphabets. For easier understanding, $E$ will be called the pattern alphabet, a word in $E^+$ is a *pattern*. Let $w = e_1 e_2 \cdots e_n \quad (e_i \in E)$ be a pattern. A word $u$ in $A^+$ is a *substitution instance* of $w$ iff there is a nonerasing morphism $\lambda$: $E^* \to A^*$ such that $u = \lambda(w)$. Equivalently, $u = x_1 x_2 \cdots x_n$ with $x_1, \ldots, x_n \in A^+$ and with $x_1 = x_j$ whenever $e_i = e_j$. Setting

for example $E = \{e\}$, $A = \{a, b, c\}$, the word $u = abcabc$ is a substitution instance of $ee$.

A word $u$ in $A^-$ *avoids* the pattern $w$ in $E^+$ iff no factor of $u$ is a substitution instance of $w$. Thus for example $u \in A^+$ avoids the pattern $ee$ iff $u$ is square-free, and $u$ avoids $ee'ee'e$ iff $u$ has no overlapping factor. Given a pattern $w$ in $E^+$, $w$ is called *avoidable on* $A$ if there exist infinitely many words $u$ in $A^+$ that avoid $w$. The existence of infinite square-free words, and infinite words without overlapping factor can be rephrased as follows: The word $ee$ is avoidable on a three-letter alphabet, the word $ee'ee'e$ is avoidable on a two-letter alphabet. This formulation, of course, raises the question of the structure of avoidable patterns. Among the results of the paper of Bean et al., we report the following: Let $n = \operatorname{Card} E$; then there is a finite alphabet $A$ such that every pattern $w$ with $|w| \geqslant 2^n$ is avoidable on $A$.

Another interesting extension of square-freeness is abelian square-freeness, also called strong nonrepetitivity. An abelian square is a word $uu'$, such that $u'$ is a rearrangement of $u$, that is $|u|_a = |u'|_a$ for each letter $a$. A word is strongly nonrepetitive if it contains no factor that is an abelian square. Calculation shows that over three letters, every word of length $\geqslant 8$ has an abelian square. On the other hand, Pleasants (1970) has shown that there is an infinite strongly nonrepetitive word over five letters. This improves considerable the previously known bound of twenty five letters given by Evdokomov in 1968. The case of four letters is still open. For related results, see Justin 1972, T. C. Brown 1971, and Dekking 1979.

Concerning idempotent semigroups, Theorem 2.4.1 is a special case of a more general result also due to Green and Rees (1952). Let $r \geqslant 1$ be an integer. Then the two following conditions are equivalent:

(i) Any finitely generated group $G$ such that $x^r = 1$ for all $x$ in $G$ is finite
(ii) Any finitely generated monoid $M$ such that $x^{r+1} = x$ for all $x$ in $M$ is finite.

The case considered in Theorem 2.4.1 is $r = 1$, and in this case the group $G$ is trivially finite. For a proof of the theorem, see Green and Rees 1952 or Lallement 1979. Note that there are integers $r$ such that condition (i), and consequently (ii), does not hold; $r = 665$ is such an integer (see Adian 1979). Moreover, Theorem 2.4.1 was generalized by Simon (1980) who proved the result that for a finitely generated semigroup $S$ the following three conditions are equivalent:

(i) $S$ is finite.
(ii) $S$ has only finitely many nonidempotent elements.
(iii) There exists an integer $m$ such that for each sequence $(s_1, \ldots, s_m)$ in $S$ there exist $i$, $j$ $(i < j)$ such that $s_i \cdots s_j$ is idempotent.