Chapter 1

REPETITIONS

1.1 WORDS AND LANGUAGES

Our first topic is repetitions of words. This is the oldest topic studied in formal language theory, going back to the work of Axel Thue at the beginning of this century. Thue's results have been rediscovered many times in various disguises. Before discussing the results themselves, we have to introduce some notions and terminology needed throughout this book.

An *alphabet* is a finite nonempty set. The elements of an alphabet Σ are called *letters* or *symbols*. A *word* over an alphabet Σ is a finite string consisting of zero or more letters of Σ, whereby the same letter may occur several times. The string consisting of zero letters is called the *empty word*, written λ. Thus, λ, 0, 1, 010, 1111 are words over the alphabet $\Sigma = \{0, 1\}$. The set of all words (resp. all nonempty words) over an alphabet Σ is denoted by Σ^* (resp. Σ^+). The set Σ^* is infinite for any Σ. Algebraically, Σ^* and Σ^+ are the free monoid and free semigroup generated by Σ.

The reader should keep in mind that the basic set Σ, its elements and strings of its elements could equally well be called a vocabulary, words and sentences, respectively. This would reflect an approach aiming at applications mainly in the field of natural languages. In this book, we stick to the standard mathematical terminology introduced above.

If x and y are words over an alphabet Σ, then so is their *catenation* xy. Catenation is an associative operation, and the empty word is an identity with respect to catenation: $x\lambda = \lambda x = x$ holds for all words x. For a word x and a natural number i, the notation x^i means the word obtained by catenating i copies of the word x. By definition, x^0 is the empty word λ.

The *length* of a word x, in symbols $|x|$, is the number of letters in x when each letter is counted as many times as it occurs. Again by definition, $|\lambda| = 0$. The length function possesses some of the formal properties of logarithm:

$$|xy| = |x| + |y|, \; |x^i| = i|x|,$$

for any words x and y and integers $i \geq 0$.

A word x is a *subword* of a word y if there are words x_1 and x_2 such that $y = x_1 x x_2$. Furthermore, if $x_1 = \lambda$ (resp. $x_2 = \lambda$), then x is called an *initial* subword or *prefix* of y (resp. a *final* subword or a *suffix* of y).

Subsets of Σ^* are referred to as (*formal*) *languages* over Σ. Thus,

$$L_1 = \{a, ba, aaba, b^5\} \quad \text{and} \quad L_2 = \{a^p \mid p \text{ prime}\}$$

are languages over the alphabet $\Sigma = \{a, b\}$, the former being finite and the latter infinite. A finite language can, at least in principle, be defined by listing all of its words. Such a procedure is not possible for infinite languages: some finitary specification other than simple listing is needed to define an infinite language. Formal language theory deals mainly with such finitary specifications of infinite languages. We shall meet some classes of them later on in this book.

The reader might find our terminology somewhat unusual: a language should consist of sentences rather than of words, as is the case in our terminology. However, as pointed out above, this is irrelevant and depends merely on the choice of the basic terminology, and anyone who feels strongly about it can rename the latter.

An operation of crucial importance in language theory is the operation of morphism. A mapping $h : \Sigma^* \to \Delta^*$, where Σ and Δ are alphabets, satisfying the condition

$$h(xy) = h(x)h(y), \text{ for all words } x \text{ and } y, \tag{1.1}$$

is called a *morphism*. For languages L over Σ we define

$$h(L) = \{h(w) \mid w \text{ is in } L\}.$$

(Thus, algebraically, a morphism of languages is a monoid morphism linearly extended to subsets of monoids.) Because of the condition (1.1), to define a morphism h, it suffices to list all the words $h(a)$, where a ranges over all the (finitely many) letters of Σ.

Triples $G = (\Sigma, h, w)$, where Σ is an alphabet, $h : \Sigma^* \to \Sigma^*$ is a morphism and w is a word over Σ, are referred to as *DOL systems.* A DOL system G defines the following sequence $S(G)$ of words over Σ:

$$w = h^0(w),\ h(w) = h^1(w),\ h(h(w)) = h^2(w),\ h^3(w),\ \ldots\ .$$

It also defines the following language

$$L(G) = \{h^i(w) \mid i \geq 0\}.$$

Thus, a DOL system constitutes a very simple finitary device for language definition. Languages defined by a DOL system are referred to as *DOL languages.* We shall discuss them later on in Chapter 5, and we will also explain the abbreviation "DOL."

An infinite sequence of elements of an alphabet Σ is called an *ω-word.* Thus, an ω-word can be identified with a mapping of the set of nonnegative integers into Σ. A very convenient way of defining some special ω-words is provided by DOL systems as follows. Consider a DOL system $G = (\Sigma, h, w)$ such that

$$h(w) = wx, \text{ where } x \in \Sigma^+, \tag{1.2}$$

that is, w is a proper prefix of $h(w)$, and furthermore, h is *nonerasing*: $h(a) \neq \lambda$ for every a in Σ. Then by (1.2)

$$h^2(w) = wxh(x), \; h^3(w) = wxh(x)h^2(x)$$

and, in general,

$$h^{i+1}(w) = h^i(w)h^i(x) \text{ for all } i \geq 0. \tag{1.3}$$

The equation (1.3) shows that, for any i, $h^i(w)$ is a proper prefix of $h^{i+1}(w)$. (Observe that $h^i(x) \neq \lambda$ because h is nonerasing.) Consequently, an ω-word α can be defined as the "limit" of the sequence $h^i(w)$, $i = 0, 1, 2, \ldots$. More explicitly, α is the ω-word whose prefix of length $|h^i(w)|$ equals $h^i(w)$, for all i. (The notions of a prefix and a subword are extended to concern ω-words in the natural fashion. A word x is a subword of an ω-word α if α can be written as $x_1 x \alpha_1$, where x_1 is a word and α_1 is an ω-word. If moreover $x_1 = \lambda$, then x is a prefix of α.) The ω-word α obtained in this fashion is said to be *generated* by the DOL system G.

1.2 THUE'S PROBLEM

The problem deals with repetitions occurring in words and ω-words. A word or an ω-word over an alphabet Σ is termed *square-free* (resp. *cube-free*) if it contains no subword of the form x^2 (resp. x^3), where x is a nonempty word. A word or an ω-word is termed *strongly cube-free* if it contains no subword of the form x^2a, where x is a nonempty word and a is the first letter of x. Clearly, every square-free word or ω-word is also strongly cube-free, and every strongly cube-free word or ω-word is also cube-free.

Thue's problem consists of constructing square-free words, as long as possible, over a given alphabet Σ, and preferably square-free ω-words.

Whenever this is not possible, strongly cube-free words or ω-words should be constructed, again as long as possible. Applications of Thue's problem arise in a variety of quite different situations, some of which are mentioned in Exercises 6–9 below.

As an initial observation, it should be noted that Thue's problem becomes easier (in a sense made precise below) if the cardinality of the alphabet Σ increases. Intuitively, this provides more "leeway." In particular, if Σ consists of only one letter, then no word of length ≥ 3 is cube-free. If Σ consists of two letters, then only very short words can be square-free, as seen in the following lemma.

Lemma 1.1. No word of length ≥ 4 over an alphabet Σ with cardinality 2 is square-free. Consequently, no ω-word over Σ is square-free.

Proof. Let Σ consist of the letters a and b. The only square-free words of length 3 are

$$aba \quad \text{and} \quad bab \tag{1.4}$$

because the other words of length 3

$$a^3,\ a^2b,\ ba^2,\ b^3,\ b^2a,\ ab^2$$

all contain either a^2 or b^2 as a subword. On the other hand, no matter in what way another letter is added to one of the words (1.4), the resulting word always contains one of the words a^2, b^2, $(ab)^2$, $(ba)^2$ as a subword and, consequently, is not square-free. □

Let α be a word (resp. an ω-word) over an alphabet Σ. A word α' of the same length as α (resp. an ω-word α') is called an *interpretation* of α if the following condition is satisfied: whenever the ith symbol (counted from the beginning) differs from the jth symbol in α, then also in α' the ith symbol differs from the jth symbol. Thus, both $a_1a_2a_3b_1a_4b_2$ and $a_1a_2a_1a_3a_1a_3$ are interpretations of the word a^3bab, whereas $a_1a_2a_3a_4a_5a_1$ is not an interpretation of a^3bab, because the first and last letter in a^3bab are different. Apart from a possible renaming of the letters, every interpretation of a word or ω-word α is obtained by providing, for each letter a in α, every occurrence of a with some lower index, where only a finite number of indices may be used. Interpretations will be met also in Chapter 7 below.

Lemma 1.2. If α (a word or an ω-word) is square-free, strongly cube-free or cube-free, then so is every interpretation α' of α.

Proof. The assertion follows directly by the definition of an interpretation. For instance, if α' has a subword of the form x^2, then the subword occurring in the same position in α must be of the form y^2. □

Lemma 1.2 shows that if α is a square-free, (resp. strongly cube-free, cube-free) ω-word strictly over an alphabet Σ (meaning that all letters of Σ actually occur in α) and Σ_1 is an alphabet of cardinality greater than that of Σ, then a square-free (resp. strongly cube-free, cube-free) ω-word strictly over Σ_1 can be constructed from α.

Returning to Thue's problem, it is obvious that if we are able to construct a square-free (resp. strongly cube-free) ω-word over an alphabet Σ, then we can also construct arbitrarily long square-free (resp. strongly cube-free) words over Σ. (The converse implication is not so obvious; however, it turns out to be true as we shall see below.) Consequently, in view of Lemma 1.1, the best results we can hope for are the solutions to the following two problems.

(i) Construct a strongly cube-free ω-word over an alphabet with cardinality 2. (*Strong cube-freeness problem.*)
(ii) Construct a square-free ω-word over an alphabet with cardinality 3. (*Square-freeness problem.*)

By Lemma 1.2, solutions to problems (i) and (ii) imply that we can construct strongly cube-free (resp. square-free) ω-words, as well as arbitrarily long words with this property, over any alphabet with cardinality ≥ 2 (resp. ≥ 3).

We shall now give a solution to problems (i) and (ii). Readers who are able to find a solution themselves might have been forerunners of formal language theory like Axel Thue.

1.3 SOLUTION TO THE STRONG CUBE-FREENESS PROBLEM

Consider a DOL system $G = (\{a, b\}, h, a)$, where the morphism h is defined by

$$h(a) = ab, \; h(b) = ba.$$

Then the first few words in the sequence $S(G)$ are

$$a, \; ab, \; abba, \; abba\,baab, \; abba\,baab\,baab\,abba, \; \ldots .$$

In general, for any $i \geq 1$, the $(i + 1)$st word w_{i+1} in the sequence $S(G)$ satisfies

$$w_{i+1} = w_i w_i', \tag{1.5}$$

where we denote by x' the word obtained from the word x by interchanging a and b.

We prove (1.5) inductively, observing first that it holds for $i = 1$. Assuming (1.5) for a fixed value i, we complete the inductive step as follows:

$$\begin{aligned} w_{i+2} &= h(w_{i+1}) = h(w_i w_i') = h(w_i)h(w_i') \\ &= w_{i+1}h(w_i') = w_{i+1}w_{i+1}', \end{aligned}$$

where the last equation follows because

$$h(x') = (h(x))'$$

holds for any word x, by the definition of h.

Denote now by α the ω-word generated by G. Our aim is to show that α is strongly cube-free. Observe that (1.5) provides a convenient way of writing down an arbitrarily long prefix of α. Following the grouping due to (1.5), we obtain

$$a \quad b \quad ba \quad baab \quad baababba \quad baababbaabbabaab \quad \ldots . \qquad (1.6)$$

Thus, (1.6) shows the beginning of α. The empty spaces are only used to indicate the position where a new w_i' has been added.

We need some properties of the ω-word α for the proof of the main result. It will be useful to consider α also in the form

$$\alpha = c_1 c_2 c_3 \ldots , \qquad (1.7)$$

where each c_j equals either a or b.

Lemma 1.3. Neither a^3 nor b^3 occurs as a subword in α. Neither *ababa* nor *babab* occurs as a subword in α. Consequently, every subword x of α such that $|x| = 5$ contains either a^2 or b^2 as a subword.

Proof. Consider the first sentence. If either a^3 or b^3 occurs as a subword in α, then it occurs as a subword in some w_i. But this is impossible because $w_i = h(w_{i-1})$ and, consequently, w_i is obtained by catenating words *ab* and *ba* in some order.

Consider the second sentence. Assume that *ababa* occurs as a subword of α, starting with the jth letter of α. Thus, in the notation of (1.7),

$$c_j c_{j+1} c_{j+2} c_{j+3} c_{j+4} = ababa. \tag{1.8}$$

We choose an i large enough such that

$$|w_i| \geq j + 4.$$

Thus, the occurrence (1.8) is already in w_i. We use again the relation $w_i = h(w_{i-1})$ and conclude that either a^3 or b^3 occurs as a subword in w_{i-1}, depending on whether j in (1.8) is odd or even. But this cannot happen because of the already established first sentence of our lemma. In the same way (or arguing by symmetry) we see that *babab* does not occur as a subword of α.

Finally, the last sentence is a consequence of the second sentence because, apart from the words *ababa* and *babab*, every word of length 5 over $\{a, b\}$ contains a^2 or b^2 as a subword. □

Lemma 1.4. Assume that a^2 or b^2 occurs as a subword of α, starting with the jth letter of α. Then j is even.

Proof. We use the notation of (1.7), assuming that $c_j c_{j+1}$ equals either a^2 or b^2. We choose again an i large enough such that

$$|w_i| \geq j + 1,$$

and make use of the relation $w_i = h(w_{i-1})$. Because of this relation, j being odd implies that $c_j c_{j+1}$ equals either $h(a)$ or $h(b)$. Since neither one of the latter equals a^2 or b^2, we obtain the lemma. □

We are now in the position to establish our main result.

Theorem 1.5. The ω-word α is strongly cube-free.

Proof. Arguing again indirectly, we assume that xxc, where c is the first letter of x, is a subword of α and, furthermore, no word yyd, where d is the first letter of y and $|y| < |x|$, is a subword of α. (In other words, xxc provides the shortest possible counterexample to Theorem 1.5.) If $|x|$ equals 1 or 2, one of the words a^3, b^3, *ababa*, *babab* occurs as a subword of α. Since this is impossible by Lemma 1.3, we conclude that

$$|x| = t \geq 3. \tag{1.9}$$

Assume that the occurrence of xxc we are considering starts with the jth letter of α. Hence, in the notation of (1.7),

$$c_j \cdots c_{j+2t} = xxc. \tag{1.10}$$

By Lemma 1.3 and (1.9), either a^2 or b^2 occurs as a subword in xx. This implies that either a^2 or b^2 occurs twice as a subword in xxc. Indeed, a^2 or b^2 must occur as a subword either in x or else in xc. In both cases, it occurs twice as a subword in xxc.

We can now conclude by Lemma 1.4 that $|x| = t$ is even. For if t is odd then at least one of the occurrences of a^2 or b^2 in xxc must start with the kth letter of α, for some odd k. But this is impossible by Lemma 1.4.

Consequently, $t = 2u$ for some natural number u. We consider (1.10), assuming first that j is even and, hence, $j \geq 2$. We now apply our standard technique, choosing a large enough i (such that $|w_i| \geq j + 2t$) and making use of the relation $w_i = h(w_{i-1})$. We conclude that

$$c_{j-1}c_j = ab \text{ or } c_{j-1}c_j = ba. \tag{1.11}$$

This implies that also

$$c_{j-1+t}c_{j+t} = ab \text{ or } c_{j-1+t}c_{j+t} = ba \tag{1.12}$$

because $j + t$ is even. By (1.10), $c_j = c_{j+t}$ which together with (1.11) and (1.12) gives the result $c_{j-1} = c_{j-1+t}$. Consequently,

$$c_n = c_{n+t} \text{ for every } n \text{ with } j - 1 \leq n \leq j + t. \tag{1.13}$$

For $0 \leq n \leq t$, the word $c_{j-1+2n}c_{j+2n}$ equals either $h(a)$ or $h(b)$. This follows because of the relation $w_i = h(w_{i-1})$ and because $j - 1$ is odd. We now infer by (1.13) that w_{i-1} contains a subword yyd, where d is the first letter of y and $|y| = t/2 = u$. Hence, also α contains yyd as a subword. But this contradicts the choice of x.

Thus, there remains the case that in (1.10), j is odd. We argue as before, extending now (1.10) to the letter c_{j+2t+1} instead of c_{j-1}. Since both $j + t$ and $j + 2t$ are odd, the words $c_{j+t}c_{j+t+1}$ and $c_{j+2t}c_{j+2t+1}$ are of the form $h(a)$ or $h(b)$, analogously with (1.11) and (1.12). Hence, the equation $c_{j+t} = c_{j+2t}$ gives us the equation $c_{j+t+1} = c_{j+2t+1}$. Consequently,

$$c_n = c_{n+t} \text{ for every } n \text{ with } j \leq n \leq j + t + 1.$$

From this we obtain exactly as above the result that α contains a subword yyd, where d is the first letter of y and $|y| = u$, contradicting again the choice of x. We have shown that α is a strongly cube-free ω-word over the alphabet $\{a, b\}$, completing the proof of Theorem 1.5. □

1.4 SOLUTION TO THE SQUARE-FREENESS PROBLEM

We now turn to the discussion of the problem of constructing a square-free ω-word over an alphabet with cardinality 3. In fact, we are able to reduce the entire matter to the already established Theorem 1.5. This becomes possible by applying a technique very common and useful in formal language theory. The technique consists of grouping several letters into one. By this technique, we obtain first the following lemma.

Lemma 1.6. There exists a square-free ω-word β over an alphabet with four letters.

Proof. Consider the ω-word α of Theorem 1.5. We define a new alphabet Σ_1 by

$$\Sigma_1 = \{[aa], [ab], [ba], [bb]\}.$$

Using the expression (1.7) for α, we now define an ω-word

$$\beta = d_1 d_2 d_3 \dots \tag{1.14}$$

over the alphabet Σ_1 by the condition

$$d_j = [c_j c_{j+1}] \text{ for every } j \geq 1.$$

Assume y^2 occurs as a subword in β where

$$y = d_{j+1} \dots d_{j+t} = d_{j+t+1} \dots d_{j+2t},\ t \geq 1.$$

Consequently,

$$[c_{j+1}c_{j+2}] \dots [c_{j+t}c_{j+t+1}] = [c_{j+t+1}c_{j+t+2}] \dots [c_{j+2t}c_{j+2t+1}].$$

This means that $c_{j+1} = c_{j+t+1} = c_{j+2t+1}$ and that $c_{j+n} = c_{j+t+n}$ for $1 \leq n \leq t$. Consequently,

$$(c_{j+1} \dots c_{j+t})^2 c_{j+1}$$

occurs as a subword in α, contradicting Theorem 1.5. Hence, β is square-free. □

We now strengthen Lemma 1.6 to the result we are looking for. For this purpose it will be convenient to abbreviate the letters of Σ_1 as follows:

$$[aa] = 1,\ [ab] = 2,\ [ba] = 3,\ [bb] = 4.$$

In this notation, the beginning of β is

$$\beta = 243231243123243231232431243231 2\ldots$$

(cf. 1.6). By the definition of β, the letter 1 must be preceded by 1 or 3. However, if 1 is preceded by 1 then a^3 occurs as a subword in α, which is impossible by Lemma 1.3. Hence, 1 is always preceded by 3 in β. The other parts of the following lemma are established in exactly the same way.

Lemma 1.7. Every occurrence of the letter 1 in β is preceded by an occurrence of 3 and followed by an occurrence of 2. Every occurrence of the letter 4 in β is preceded by an occurrence of 2 and followed by an occurrence of 3.

We are now ready for the main result.

Theorem 1.8. There exists a square-free ω-word γ over an alphabet with three letters.

Proof. Consider the alphabet $\Sigma_2 = \{1, 2, 3\}$. The ω-word γ is obtained from β by replacing 4 with 1. Thus, the beginning of γ is

$$\gamma = 2132312131232132312321312132312\ldots\ .$$

We will show that γ is square-free.

Assume the contrary: xx occurs as a subword in γ, where x is a nonempty word. This implies that β contains a subword y_1y_2 such that

$$|y_1| = |y_2| = |x| = t$$

and, furthermore, y_1 and y_2 become identical when every occurrence of the letter 4 is replaced by the letter 1.

We observe first that $t \geq 2$ because, by Lemma 1.7, none of the words 11, 14, 41, 44 occurs as a subword in β.

Let

$$y_1 = d_{j+1} \ldots d_{j+t}, y_2 = d_{j+t+1} \ldots d_{j+2t},$$

where the notation of (1.14) is used. Thus, for every n satisfying $1 \le n \le t$, $d_{j+n} = d_{j+n+t}$ with the possible exception of the case where one of the numbers d_{j+n} and d_{j+n+t} equals 1 and the other equals 4. We shall prove that this exceptional case is, in fact, impossible.

Consider first a fixed value of n satisfying $1 \le n < t$. By Lemma 1.7, if d_{j+n} equals 1 (resp. 4), then d_{j+n+1} equals 2 (resp. 3). Hence, also $d_{j+n+1+t}$ equals 2 (resp. 3), by our assumption concerning y_1 and y_2. Thus, another application of Lemma 1.7 gives us the result

$$d_{j+n} = d_{j+n+t} \text{ whenever } 1 \le n < t. \tag{1.15}$$

Second, consider the letters d_{j+t} and d_{j+2t}. Instead of successors, we use now predecessors in our argument based on Lemma 1.7. If d_{j+t} equals 1 (resp. 4), then d_{j+t-1} equals 3 (resp. 2). Consequently, also d_{j+2t-1} equals 3 (resp. 2). Hence, $d_{j+t} = d_{j+2t}$, which combined with (1.15) shows that $(d_{j+1} \ldots d_{j+t})^2$ occurs as a subword in β, contradicting Lemma 1.6. Our assumption about xx occurring as a subword in γ is wrong, whence Theorem 1.8 follows. □

We still summarize our complete solution of Thue's problem in the next theorem.

Theorem 1.9. If Σ is of cardinality ≥ 3 then there exists a square-free ω-word over Σ. If Σ is of cardinality 2 then there exists a strongly cube-free ω-word over Σ but no square-free words over Σ with length exceeding 3.

We want to emphasize that the expression "there exists" in the statement of the previous theorem means, in fact, that the required ω-word can be effectively constructed. By (1.5), we are able to compute any "digit" in α, β or γ. With very few exceptions, this is generally true in formal language theory: proofs of theorems give a method of effectively constructing the objects involved.

1.5 OVERLAPPING

By definition, a word or an ω-word w is square-free if it does not contain a subword of the form xx, where x is nonempty. It is still conceivable that w

could contain two "overlapping" occurrences of x, i.e., a subword $xy = zx$, where

$$1 \leq |y| = |z| < |x|. \tag{1.16}$$

If this does not happen, we say that w is *overlap-free*.

The following characterization result shows, among other things, that the square-free and strongly cube-free ω-words considered above are also overlap-free.

Lemma 1.10. A word or an ω-word w is overlap-free if and only if it is strongly cube-free.

Proof. Assume w is not overlap-free. Hence, w possesses a subword $xy = zx$ such that (1.16) holds. Let a be the first letter of z. By our assumptions, $x = zx_1$ where the first letter of x_1 is also a. Hence, zza is a subword of w. Consequently, w is not strongly cube-free.

Conversely, assume that w is not strongly cube-free. Hence, w possesses a subword z_1z_1a, where a is the first letter of z_1. Denoting $z_1 = az_2$, we see that az_2az_2a is a subword of w. We now choose

$$x = az_2a, \; y = z_2a, \; z = az_2.$$

Then $xy = zx$ is a subword of w and, moreover, (1.16) is satisfied. This implies that w is not overlap-free. □

Thue also considered a notion stronger than square-freeness. He called an ω-word over Σ with cardinality n "irreducible" if, whenever it contains a subword xyx where x is nonempty, then $|y| \geq n - 2$. Thus, for $n = 3$, this notion coincides with the notion of square-freeness. Given any alphabet Σ, an irreducible ω-word over Σ can be constructed. (Observe that for $n \leq 2$ every ω-word is trivially irreducible.)

The condition of two occurrences of x lying apart can be further strengthened by requiring that the length of the word y separating the occurrences is bounded from below by $|x|$. Along these lines, the following result can be obtained.

Theorem 1.11. If Σ is of cardinality ≥ 3, there is an ω-word w over Σ such that, whenever xyx with $x \neq \lambda$ is a subword of w, then $|y| \geq \frac{1}{3}|x|$.

The proof of Theorem 1.11 is given in [De]. A w as required is generated by the DOL system $(\{a, b, c\}, h, a)$, where

$$\begin{aligned} h(a) &= abc\ \ acb\ \ cab\ \ c\ \ bac\ \ bca\ \ cba,\\ h(b) &= bca\ \ bac\ \ abc\ \ a\ \ cba\ \ cab\ \ acb,\\ h(c) &= cab\ \ cba\ \ bca\ \ b\ \ acb\ \ abc\ \ bac. \end{aligned}$$

It is also discussed in [De] why $|h(a)|$ cannot be smaller and shown that the constant $1/3$ is the best possible in the following sense. Assume that the cardinality of Σ equals 3. Then every word over Σ with length ≥ 39 contains a subword xyx with the properties $x \neq \lambda$ and $|y| \leq 1/3\,|x|$.

1.6 DOL SYSTEMS AND ω-WORDS

We conclude this chapter with some further examples of square-free ω-words. At the same time, some general remarks about ω-words generated by DOL systems will be made.

Observe first that the ω-word α of Theorem 1.5 is also generated by the DOL system

$$G_1 = (\{a, b\}, h, abba), \text{where } h(a) = ab, h(b) = ba,$$

as well as by the DOL system

$$G_2 = (\{a, b\}, h_1, a), \text{ where } h_1(a) = abba, h_1(b) = baab.$$

This is a special case of the following more general result, the proof of which is immediate by the definitions.

Lemma 1.12. Assume that δ is the ω-word generated by the DOL system (Σ, h, w) and that $i \geq 1$ and $j \geq 0$ are integers. Then δ is generated also by the DOL system $(\Sigma, h^i, h^j(w))$.

It may also happen that the original DOL system does not generate an ω-word, but there still exist numbers i and j such that the DOL system $(\Sigma, h^i, h^j(w))$ generates an ω-word. An example is given in Exercise 14. In Exercise 15 the problem of finding out whether or not such numbers exist is discussed.

In general, it is not easy to tell whether or not two DOL systems G_1 and G_2 (satisfying the additional condition for the generation of ω-words) generate the *same* ω-word. This problem is closely linked with the DOL sequence equivalence problem considered below in Chapter 5. If G_1 and G_2 are sequence equivalent (i.e., $S(G_1) = S(G_2)$), then they generate the same ω-word. The converse is not necessarily true, as exemplified by Lemma 1.12.

It is difficult to decide whether or not an ω-word defined by some other effective method can also be defined by a DOL system. This can be stated as a precise decision problem for each class of effective methods defining ω-words.

For instance, the ω-word β of Lemma 1.6 was not originally defined by a DOL system. However, β is generated by the DOL system $(\{1, 2, 3, 4\}, h, 2)$, where

$$h(1) = 2431, h(2) = 2432,$$
$$h(3) = 3123, h(4) = 3124.$$

No general method is known for deciding whether or not the ω-word generated by a DOL system G is square-free. In [Be1] such a method is given for the case where the alphabet of G consists of three letters.

We say that a morphism *h preserves square-freeness* if $h(x)$ is square-free whenever x is square-free. Clearly, the ω-word δ generated by the DOL system (Σ, h, w), where w is square-free and h preserves square-freeness, is itself square-free (providing, of course, that the conditions for generating an ω-word are satisfied). For instance, Thue shows in [T2] that the morphism h_1 defined by

$$h_1(a) = abcab, h_1(b) = acabcb, h_1(c) = acbcacb$$

is square-free. Thus, the DOL system $(\{a, b, c\}, h_1, a)$ generates a square-free ω-word.

However, a DOL system (Σ, h, w) may generate a square-free ω-word although h does not preserve square-freeness. An example is provided by the DOL system $(\{a, b, c\}, h_2, a)$, where h_2 is defined by

$$h_2(a) = abc, h_2(b) = ac, h_2(c) = b.$$

It can be shown that the generated ω-word is square-free. However, h_2 does not preserve square-freeness because

$$h_2(aba) = abcacabc.$$

EXERCISES

1. What is the number of subwords of a word w of length n, provided w contains no two occurrences of the same letter? Determine an upper bound, as sharp as possible, for the number of subwords in case the letters of w are not all distinct.

2. Consider the DOL system $G = (\{a, b\}, h, a)$, where $h(a) = b$ and $h(b) = ab$. Show that the lengths of the words in the sequence $S(G)$ constitute the Fibonacci sequence.