Combinatorics, Automata and Number Theory

CANT

Edited by
Valérie Berthé
LIRMM - Université Montpelier II - CNRS UMR 5506
161 rue Ada, F-34392 Montpellier Cedex 5, France

Michel Rigo
Université de Liège, Institut de Mathématiques
Grande Traverse 12 (B 37), B-4000 Liège, Belgium

1

Preliminaries

V. Berthé
LIRMM - Université Montpelier II - CNRS UMR 5506
161 rue Ada, F-34392 Montpellier cedex 5, France

M. Rigo
Université de Liège, Institut de Mathématiques,
Grande Traverse 12 (B 37), B-4000 Liège, Belgium.

The aim of this chapter is to introduce basic objects that are encountered in the different parts of this book. In the first section, we start with a few conventions. Section 1.2 presents finite and infinite words and fundamental operations that can be applied to them. In particular important concepts like eventually periodic words, substitutive words or factor complexity function are introduced (more material is given in Chapter 4). Sets of words are languages. They are presented in Section 1.3 together with regular languages, finite automata and transducers (more material is presented in Section 2.6). Section 1.4 introduces some matrices naturally associated with automata or morphisms. Section 1.5 presents basic results on numeration systems that will be developed in Chapter 2. Finally, Section 1.6 introduces concepts from symbolic dynamics.

1.1 Conventions

Let us start with some basic notation used along this book. We assume the reader familiar with usual basic set operations like union, intersection or set difference: $\cup$, $\cap$ or $\setminus$. Sets of numbers are of particular interest. The set of non-negative integers (respectively integers, rational numbers, real numbers, complex numbers) is $\mathbb{N}$ (respectively $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$). Let a be a real number and $\mathbb{K} = \mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$. We set

$$\mathbb{K}_{\geq a} := \mathbb{K} \cap [a, +\infty),\ \mathbb{K}_{>a} := \mathbb{K} \cap (a, +\infty)\ ,$$

$$\mathbb{K}_{\leq a} := \mathbb{K} \cap (-\infty, a],\ \mathbb{K}_{<a} := \mathbb{K} \cap (-\infty, a)\ .$$

For instance, $\mathbb{N}_{>0}$ can indifferently be written $\mathbb{N} \setminus \{0\}$ or $\mathbb{N}_{\geq 1}$. Let $i, j \in \mathbb{Z}$ with $i \leq j$. We use the notation $[\![i, j]\!]$ for the set of integers $\{i, i+1, \ldots, j\}$.

Let X, Y be two sets. The notation $X \subseteq Y$ stands for the fact that every

element of X is an element of Y, whereas $X \subset Y$ stands for the strict inclusion, *i.e.*, $X \subseteq Y$ and $X \neq Y$. Let X^Y denote the set of all mappings from Y to X. Therefore the set of sequences indexed by $\mathbb{N}$ (respectively by $\mathbb{Z}$) of elements in X is denoted by $X^{\mathbb{N}}$ (respectively by $X^{\mathbb{Z}}$). As a particular case, 2^X is the power set of X, *i.e.*, the set of all subsets of X. Indeed, 2 can be identified with $\{0,1\}$ and maps from X to $\{0,1\}$ are in one-to-one correspondence with subsets of X. In particular, if X is finite of cardinality $\operatorname{Card} X = n$, then 2^X contains 2^n sets. The Cartesian product of X and Y is denoted by $X \times Y$. It is the set of ordered pairs (x, y) for all $x \in X$ and $y \in Y$. For a subset X of a topological space, $\operatorname{int}(X)$ stands for the *interior* of X, $\overline{X}$ for the closure of X, and ∂X for its *boundary*, that is, $\partial X = \overline{X} \setminus \operatorname{int}(X)$.

The floor of a real number x is $\lfloor x \rfloor = \sup\{z \in \mathbb{Z} \mid z \leq x\}$, whereas $\{x\} = x - \lfloor x \rfloor$ stands for the fractional part of x. For $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we will use the following set of notation for the most usual norms

$$||\mathbf{x}||_1 := \sum_{i=1}^{n} |x_i|, \ ||\mathbf{x}||_\infty = \max_i |x_i|, \ ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \ ,$$

and will denote the corresponding open ball with radius R and center $\mathbf{x}$ as $B_1(\mathbf{x}, R)$, $B_\infty(\mathbf{x}, R)$, $B_2(\mathbf{x}, R)$, respectively. For more on vector norms, see Section 4.7.2.2.

It is a good opportunity to recall here notation about asymptotics. Let $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. The definitions given below can also be applied to functions defined on another domain like $\mathbb{R}_{>a}$, $\mathbb{N}$ or $\mathbb{Z}$. We assume implicitly that the following notions are defined for $x \to +\infty$. We write $f \in \mathcal{O}(g)$, if there exist two constants x_0 and $C > 0$ such that, for all $x \geq x_0$, $|f(x)| \leq C|g(x)|$. We also write $f \ll g$ or $g \gg f$, or else $g \in \Omega(f)$. Note that we can write either $f \in \mathcal{O}(g)$ or $f = \mathcal{O}(g)$. Be aware that in the literature, authors give sometime different meanings to the notation $\Omega(f)$. Here we consider a bound, for all large enough x, but there exist variants where the bound holds only for an increasing sequence $(x_n)_{n \geq 0}$ of reals, *i.e.*, $\limsup_{x \to +\infty} |g(x)|/|f(x)| > 0$.

If g belongs to $\mathcal{O}(f) \cap \Omega(f)$, *i.e.*, there exist constants x_0, C_1, C_2 with $C_1, C_2 > 0$ such that, for all $x \geq x_0$, $C_1|f(x)| \leq |g(x)| \leq C_2|f(x)|$, then we write $g \in \Theta(f)$. As an example, the function $x^2 + \sin 6x$ is in $\Theta(x^2)$ and $x^2|\sin(4x)|$ is in $\mathcal{O}(x^2)$ but not in $\Theta(x^2)$. In Figure 1.1, we have represented the functions $x^2 + \sin 6x$, $x^2|\sin(4x)|$, $4x^2/5$ and $6x^2/5$.

If $\lim_{x \to +\infty} \frac{f(x)}{g(x)} = 0$, we write $f = o(g)$. Finally, if $\lim_{x \to +\infty} \frac{f(x)}{g(x)} = 1$, we write $f \sim g$. For more on asymptotics, see for instance (de Bruijn 1981) or the first chapter of (Hardy and Wright 1985).

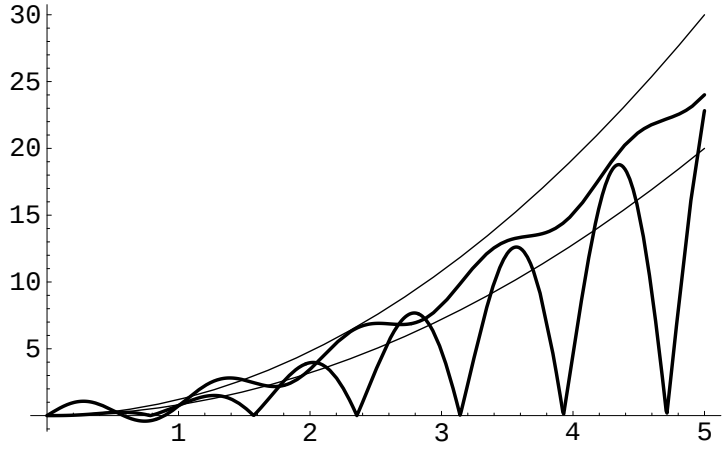

Fig. 1.1. The functions $x^2 + \sin 6x$, $x^2|\sin(4x)|$, $4x^2/5$ and $6x^2/5$.

Lastly, we will use the notation $\log = \log_e$ for the natural logarithm, whereas $\log_2$ will denote the binary logarithm.

1.2 Words

This section is only intended to give basic definitions of concepts developed later on. For material not covered in this book, classical textbooks on finite or infinite words and their properties are (Lothaire 1983), (Lothaire 2002), (Lothaire 2005). See also the chapter (Choffrut and Karhumäki 1997) or the tutorial (Berstel and Karhumäki 2003). The first chapters of the books (Allouche and Shallit 2003) and (Pytheas Fogg 2002) contain also many references for further developments in combinatorics on words.

1.2.1 Finite words

An *alphabet* is a finite set of *symbols* (or *letters*). Usually, alphabets will be denoted using Roman upper case letters, like A or B. The most basic and fundamental objects that we shall deal with are *words*.

Let A be an alphabet. A *finite* word over A (to distinguish with the infinite case that will be considered later on) is a finite sequence of letters in A. In a formal way, a word of length $n \in \mathbb{N}$ is a map u from $[\![0, n-1]\!]$ to A. Instead of a functional notation, it is convenient to write a word as $u = u_0 \cdots u_{n-1}$ to express u as the concatenation of the letters u_i. The

length of u, that is, the size of its domain, is denoted by $|u|$. The unique word of length 0 is the *empty word* denoted by ε.

In order to endow the set of finite words with a suitable algebraic structure, we introduce the following definitions.

Definition 1.2.1 Recall that a *semigroup* is an algebraic structure given by a set R that is equipped with a product operation from $R \times R$ to R which is associative, *i.e.*, for all $a, b, c \in R$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

Moreover, if this associative product on R possesses a (necessarily unique) identity element $1_R \in R$, *i.e.*, for all $a \in R$, $a \cdot 1_R = a = 1_R \cdot a$, then this algebraic structure is said to be a *monoid*. For instance the set $\mathbb{N}^d$, with $d \geq 1$, of d-tuples of non-negative integers with the usual addition component-wise is a monoid with $(0, \ldots, 0)$ as identity element.

Definition 1.2.2 Let $(R, \cdot)$ and $(T, \diamond)$ be monoids with respectively 1_R and 1_T as identity element. A map $f : R \to T$ is a *monoid morphism* (or *homomorphism of monoids*) if $f(1_R) = 1_T$ and for all $a, b \in R$, $f(a \cdot b) = f(a) \diamond f(b)$.

Let $u = u_0 \cdots u_{m-1}$ and $v = v_0 \cdots v_{n-1}$ be two words over A. The *concatenation* of u and v is the word $w = w_0 \cdots w_{m+n-1}$ defined by $w_i = u_i$ if $0 \leq i < m$, and $w_i = v_{i-m}$ otherwise. We write $u \cdot v$ or simply uv to express the concatenation of u and v. Notice that this operation is associative. Let u be a word and $n \in \mathbb{N}$. Naturally, let u^n denote the concatenation of n copies of u and we set $u^0 = \varepsilon$. A *square* is a word of the form uu, where $u \in A^*$.

The set of all (finite) words over A is denoted by A^*. Endowed with the concatenation of words as product operation, A^* is a monoid with ε as identity element. It is the *free* monoid generated by A (freeness means that any element in A^* has a unique factorisation as product of elements in A). Notice that the length map $|\cdot| : (A^*, \cdot) \to (\mathbb{N}, +)$, $w \mapsto |w|$ is a morphism of monoids. Let $A^+ = A^* \setminus \{\varepsilon\}$ denote the free semigroup generated by A. Finally, for $n \in \mathbb{N}$, A^n is the set of words of length n over A and $A^{\leq n}$ is the set of words over A of length less or equal to n.

The *mirror* (sometimes called *reversal*) of a word $u = u_0 \cdots u_{m-1}$ is the word $\tilde{u} = u_{m-1} \cdots u_0$. It can be defined inductively on the length of the word by $\tilde{\varepsilon} = \varepsilon$ and $\widetilde{au} = \tilde{u}a$ for $a \in A$ and $u \in A^*$. Notice that for $u, v \in A^*$, $\widetilde{uv} = \tilde{v}\tilde{u}$. A *palindrome* is a word u such that $\tilde{u} = u$. For instance, the palindromes of length at most 3 in $\{0, 1\}^*$ are

$$\varepsilon, 0, 1, 00, 11, 000, 010, 101, 111.$$

We end this section about finite words with the notion of code.

Definition 1.2.3 A subset $Y \subset A^+$ is a *code* if, for all $u_1, \ldots, u_m, v_1 \ldots, v_n \in Y$, the equality $u_1 \cdots u_m = v_1 \cdots v_n$ implies $n = m$ and $u_i = v_i$ for $i = 1, \cdots, m$. A code is said to be a prefix code if none of its elements is a prefix of another one.

1.2.2 Infinite words

To define infinite words, we consider maps taking values in an alphabet but defined on an infinite domain. A *(one-sided) infinite word* over an alphabet A is a map from the set $\mathbb{N}$ of non-negative integers to A. Using the same convention as for finite words, we write $x = x_0x_1x_2\cdots$ to represent an infinite word. It is sometimes convenient to use a notation like $x = (x_n)_{n\geq 0}$. If the domain is the set $\mathbb{Z}$ of integers, then we speak of *bi-infinite word* (in the literature, we also find the terminology of *two-sided infinite words*). In this latter situation, a convenient notation is to use a decimal point to determine the position of the image of 0 like $\cdots x_{-2}x_{-1}.x_0x_1x_2\cdots$.

In what follows if no explicit mention is made then we shall be dealing with one-sided infinite words and we will omit reference to it.

The set of infinite words over A is denoted by $A^{\mathbb{N}}$. We can define a concatenation operation from $A^* \times A^{\mathbb{N}}$ to $A^{\mathbb{N}}$ as follows. The concatenation of the finite word $u = u_0 \cdots u_{n-1}$ and the infinite word $x = x_0x_1\cdots$ is the infinite word $y = y_0y_1\cdots$ denoted by ux and defined by $y_i = u_i$ if $0 \leq i \leq n-1$, and $y_i = x_{i-n}$ if $i \geq n$.

Example 1.2.4 Consider the infinite word $x = x_0x_1x_2\cdots$ where the letters $x_i \in \{0, \ldots, 9\}$ are given by the digits appearing in the usual decimal expansion of $\pi - 3$,

$$\pi - 3 = \sum_{i=0}^{+\infty} x_i \, 10^{-i-1} \ ,$$

i.e., $x = 14159265358979323846264338327950288419\cdots$ is an infinite word.

Definition 1.2.5 Any subset X of $\mathbb{N}$ (respectively $\mathbb{Z}$) gives rise to an infinite (respectively bi-infinite) word over $\{0, 1\}$, namely its *characteristic word*. Let x be this word. It is defined as follows

$$x_n = \begin{cases} 1, \text{ if } n \in X, \\ 0, \text{ otherwise.} \end{cases}$$

It also refers to the *indicator function* of the set X, denoted by $\mathbb{1}_X(n)$.

Example 1.2.6 Consider the characteristic sequence of the set of prime numbers $x = x_0x_1\cdots = 0011010100010100010100010000\cdots$.

1.2.3 Factors, topology and orderings

The following notions can be defined for both finite and infinite words. Let us start with the finite case. Let $u = u_0 \cdots u_{n-1}$ be a finite word over A. If u can be factorised as $u = vfw$ with $v, f, w \in A^*$, we say that f is a *factor* of u. If $f = u_i \cdots u_{i+|f|-1}$, then f is said to *occur* at position i in u. For convenience, $u[i, i+\ell-1]$ denotes the factor of u of length $\ell \geq 1$ occurring at position i. The number of occurrences of f in u is denoted by $|u|_f$. In particular, if $a \in A$, then $|u|_a$ denotes the number of letters a occurring in u. If u is a finite or infinite word over A, then $\mathrm{alph}(u)$ is the set of letters which occur in u. If u is the empty word, then $\mathrm{alph}(u)$ is the empty set. One has $\mathrm{alph}(u) \subseteq A$.

Assume that $A = \{a_1 < \cdots < a_n\}$ is totally ordered. The map $\mathbf{P} : A^* \to \mathbb{N}^n,\ w \mapsto {}^t(|w|_{a_1}, \ldots, |w|_{a_n})$ is called the *abelianisation map* . It is trivially a morphism of monoids. Notice that in the literature, this map is also referred to as the *Parikh mapping*. Note that for a matrix $\mathbf{M}$, ${}^t\mathbf{M}$ is the transpose of $\mathbf{M}$.

If $u = fw$ (respectively $u = vf$) then f is a *prefix* (respectively a *suffix*) of u. A word $u = u_0 \cdots u_{n-1}$ of length n has exactly $n+1$ prefixes: ε, u_0, u_0u_1, $\ldots$, $u_0 \cdots u_{n-2}$, u. The same holds for suffixes. A *proper* prefix (respectively *proper* suffix) of u is a prefix (respectively suffix) different from the full word u. Let us observe that a factor of u is obtained as the concatenation of consecutive letters occurring in u. By opposition a *scattered subword* of $u = u_0 \cdots u_{n-1}$ is of the form $u_{i_0}u_{i_1} \cdots u_{i_k}$ with $k < n$ and $0 \leq i_1 < i_2 < \cdots < i_k < n$.

Example 1.2.7 Let $A = \{0, 1\}$ be the binary alphabet consisting of letters 0 and 1. The set A^* contains all the finite words obtained by concatenating 0's and 1's. The concatenation of the words $u = 1001$ and $v = 010$ is the word $w = uv = 1001010 = w_0 \cdots w_6$. The word v occurs twice in w at positions 2 and 4. We have $w[1, 3] = 001$ and the suffix 1010 is a square, *i.e.*, $(10)^2$. To conclude with the example, $|w|_0 = |u|_0 + |v|_0 = 2 + 2 = 4$.

The notions of *factor*, *prefix* or *suffix* as well as the according notation introduced for finite words can be extended to infinite words. Factors and prefixes are finite words, but a suffix of an infinite word is also infinite. Let $x = x_0x_1x_2 \cdots$ be an infinite word over A. For instance, for $\ell \geq 0$, $x[0, \ell-1] = x_0 \cdots x_{\ell-1}$ is the prefix of length ℓ of x. We denote by $x[i, i+\ell-1] = x_i \cdots x_{i+\ell-1}$ the factor of length $\ell \geq 1$ occurring in x at position $i \geq 0$. For $n \geq 0$, the infinite word $x_nx_{n+1} \cdots$ is a suffix of x. See the relationship with the notion of shift introduced in Section 1.6.

Definition 1.2.8 The *language* of the infinite word x is the set of all its factors. It is denoted by $L(x)$. The set of factors of length n occurring in x is denoted by $L_n(x)$.

Definition 1.2.9 An infinite word x is *recurrent* if all its factors occur infinitely often in x. It is *uniformly recurrent* (also called *minimal*), if it is recurrent and for every factor u of x, if $T_x(u) = \{i_1^{(u)} < i_2^{(u)} < i_3^{(u)} < \cdots\}$ is the infinite set of positions where u occurs in x, then there exists a constant C_u such that, for all $j \geq 1$,

$$i_{j+1}^{(u)} - i_j^{(u)} \leq C_u.$$

An infinite set $X \subseteq \mathbb{N}$ of integers having such a property, *i.e.*, where the difference of any two consecutive elements in X is bounded by a constant, is said to be *syndetic* or with *bounded gap*. Otherwise stated, an infinite word x is uniformly recurrent if, and only if, for all factors $u \in L(x)$, the set $T_x(u)$ is infinite and syndetic.

Definition 1.2.10 One can endow $A^{\mathbb{N}}$ with a *distance* d defined as follows. Let x, y be two infinite words over A. Let $x \wedge y$ denote the longest common prefix of x and y. Then the distance d is given by

$$d(x,y) := \begin{cases} 0, & \text{if } x = y, \\ 2^{-|x \wedge y|}, & \text{otherwise.} \end{cases}$$

It is obvious to see that, for all $x, y, z \in A^{\mathbb{N}}$, $d(x,y) = d(y,x)$, $d(x,z) \leq d(x,y) + d(y,z)$ and $d(x,y) \leq \max(d(x,z), d(y,z))$. This last property is not required to have a distance, but when it holds, the distance is said to be *ultrametric*.

This notion of distance extends to $A^{\mathbb{Z}}$. Notice that the topology on $A^{\mathbb{N}}$ is the product topology (of the discrete topology on A). The space $A^{\mathbb{N}}$ is a compact *Cantor set*, that is, a totally disconnected compact space without isolated points. Since $A^{\mathbb{N}}$ is a (complete) metric space, it is therefore relevant to speak of convergent sequences of infinite words. The sequence $(z_n)_{n\geq 0}$ of infinite words over A *converges* to $x \in A^{\mathbb{N}}$, if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, for all $n \geq N$, $d(z_n, x) < \varepsilon$. To express the fact that a sequence of finite words $(w_n)_{n\geq 0}$ over A converges to an infinite word y, it is assumed that A is extended with an extra letter $c \notin A$. Any finite word w_n is replaced with the infinite word $w_n ccc\cdots$ and if the sequence of infinite words $(w_n ccc\cdots)_{n\geq 0}$ converges to y, then the sequence $(w_n)_{n\geq 0}$ is said to converge to y.

Let $(u_n)_{n\geq 0}$ be a sequence of non-empty finite words. If we define, for all

$\ell \geq 0$, the finite word v_ℓ as the concatenation $u_0 u_1 \cdots u_\ell$, then the sequence $(v_\ell)_{\ell \geq 0}$ of finite words converges to an infinite word. This latter word is said to be the concatenation of the elements in the infinite sequence of finite words $(u_n)_{n \geq 0}$. In particular, for a constant sequence $u_n = u$ for all $n \geq 0$, $v_\ell = u^{\ell+1}$ and the concatenation of an infinite number of copies of the finite word u is denoted by u^ω.

Definition 1.2.11 An infinite word $x = x_0 x_1 \cdots$ is *(purely) periodic* if there exists a finite word $u = u_0 \cdots u_{k-1} \neq \varepsilon$ such that $x = u^\omega$, *i.e.*, for all $n \geq 0$, we have $x_n = u_r$ where $n = dk+r$ with $r \in [\![0, k-1]\!]$. An infinite word x is *eventually periodic* if there exist two finite words $u, v \in A^*$, with $v \neq \varepsilon$ such that $x = uvvv \cdots = uv^\omega$. Notice that purely periodic words are special cases of eventually periodic words. For any eventually periodic word x, there exist words u, v of shortest length such that $x = uv^\omega$, then the integer $|u|$ (respectively $|v|$) is referred to as the *preperiod* (respectively *period*) of x. An infinite word is said *non-periodic* if it is not ultimately periodic. A set $X \subseteq \mathbb{N}$ of integers is *eventually periodic* if its characteristic word is eventually periodic. Otherwise stated, X is eventually periodic if, and only if, it is a finite union of arithmetic progressions. Recall that an arithmetic progression is a set of integers of the kind $p\mathbb{N} + q = \{pn + q \mid n \in \mathbb{N}\}$.

Definition 1.2.12 The *complexity function* of an infinite word x maps $n \in \mathbb{N}$ onto the number $p_x(n) = \operatorname{Card} L_n(x)$ of distinct factors of length n occurring in x.

This function will be studied in details in Chapter 4.

Definition 1.2.13 An infinite word x is *Sturmian* if $p_x(n) = n + 1$ for all $n \geq 0$. In particular, Sturmian words are over a binary alphabet.

From the developments in Chapter 4 and in particular thanks to the celebrated theorem of Morse and Hedlund, Sturmian words are non-periodic words of smallest complexity.

A survey on Sturmian words by J. Berstel and P. Séébold can be found in (Lothaire 2002), the chapter by P. Arnoux in (Pytheas Fogg 2002) is also of interest.

The complexity function counts the number of different factors of a given length in an infinite word x. Each distinct factor u of length n increments $p_x(n)$ of one whatever it occurs only once in x or conversely occurs many times. So to speak, $p_x(n)$ does not reveal the frequency of occurrences of the different factors. We might need more precise information concerning the frequency of a factor.

Definition 1.2.14 Let x be an infinite word. The *frequency* $f_x(u)$ of a factor u of x is defined as the limit (when n tends towards infinity), if it exists, of the number of occurrences of the factor u in $x_0x_1\cdots x_{n-1}$ divided by n, *i.e.*, provided the limit exists,

$$f_x(u) = \lim_{n\to+\infty} \frac{|x[0,n-1]|_u}{n}.$$

Let us now introduce orders on words. The sets A^* and $A^{\mathbb{N}}$ can be ordered as follows.

Definition 1.2.15 Assume that $(A,<)$ is a totally (or linearly) ordered alphabet. Then the set A^* is totally ordered by the *radix order* (or sometimes called *genealogical order*) defined as follows. Let u,v be two words in A^*. We write $u \prec v$ if either $|u|<|v|$, or if $|u|=|v|$ and there exist $p,q,r\in A^*$, $a,b\in A$ with $u=paq$, $v=pbr$ and $a<b$. By $u\preceq v$, we mean that either $u\prec v$ or $u=v$. The set A^* can also be totally ordered by the *lexicographic order* defined as follows. Let u,v be two words in A^*, we write $u<v$ if u is a proper prefix of v or if there exist $p,q,r\in A^*$, $a,b\in A$ with $u=paq$, $v=pbr$ and $a<b$. By $u\le v$, we mean that either $u<v$ or $u=v$. .

Observe that on a unary (*i.e.*, single letter) alphabet, the two orderings over $\{a\}^*$ coincide but if the cardinality of the alphabet A is at least 2, then the radix order is a well order (*i.e.*, every non-empty subset of A^* has a least element for this order) but the lexicographic order is not. For instance, the set of words $\{a^nb \mid n\ge 0\}$ does not have a least element for the lexicographic order.

Definition 1.2.16 Notice that the lexicographic order introduced on A^* can naturally be extended to $A^{\mathbb{N}}$. Let $x,y\in A^{\mathbb{N}}$. We have $x<y$ if there exist $p\in A^*$, $a,b\in A$ and $w,z\in A^{\mathbb{N}}$ such that $x=paw$, $y=pbz$ and $a<b$.

1.2.4 Morphisms

Particular infinite words of interest can be obtained by iterating morphisms (or homomorphisms of free monoids). A survey on morphisms is given in (Harju and Karhumäki 1997). Again the textbooks like (Queffélec 1987), (Pytheas Fogg 2002), (Lothaire 1983), (Lothaire 2002) or (Berstel, Aaron, Reutenauer, et al. 2008) are worth of reading for topics not considered here.

Let A and B be two alphabets. A *morphism* (also called *substitution*) is a map $\sigma : A^* \to B^*$ such that $\sigma(uv)=\sigma(u)\sigma(v)$ for all $u,v\in A^*$ (see also Definition 1.2.2). Note that the terminology substitution often refers

in the literature to non-erasing endomorphisms. We similarly define the notion of *endomorphism* if $A = B$. Notice that in particular, $\sigma(\varepsilon) = \varepsilon$. Usually morphisms will be denoted by Greek letters. To define completely a morphism, it is enough to know the images of the letters in A, the image of a word $u = u_0 \cdots u_{n-1}$ being the concatenation of the images of its letters, $\sigma(u) = \sigma(u_0) \cdots \sigma(u_{n-1})$. Otherwise stated, any map from A to B^* can be uniquely extended to a morphism from A^* to B^*.

Definition 1.2.17 Let $k \in \mathbb{N}$. A morphism $\sigma : A^* \to B^*$ is *uniform* (or *k-uniform*) if for all $a \in A$, $|\sigma(a)| = k$. A 1-uniform morphism is often called *coding* or *letter-to-letter* morphism. If for some $a \in A$, $\sigma(a) = \varepsilon$, then σ is said to be *erasing*, otherwise it is said to be *non-erasing*.

If $\sigma : A^* \to B^*$ is a non-erasing morphism, it can be extended to a map from $A^{\mathbb{N}}$ to $B^{\mathbb{N}}$ as follows. If $x = x_0 x_1 \cdots$ is an infinite word over A, then the sequence of words $(\sigma(x_0 \cdots x_{n-1}))_{n \geq 0}$ is easily seen to be convergent towards an infinite word over B. Its limit is denoted by $\sigma(x) = \sigma(x_0)\sigma(x_1)\sigma(x_2) \cdots$. We similarly extend σ to a map from $A^{\mathbb{Z}}$ to $B^{\mathbb{Z}}$ as follows. If $x = \cdots x_{-2}x_{-1}.\, x_0 x_1 x_2 \cdots$ is a bi-infinite word over A, then the sequence of words $(\sigma(x_{-n} \cdots x_{-1}.\, x_0 \cdots x_{n-1}))_{n \geq 0}$ is easily seen to be convergent towards a bi-infinite word over B. Its limit is here again denoted by $\sigma(x)$. Consequently, the definition of morphisms extend from A^* to $A^* \cup A^{\mathbb{N}} \cup A^{\mathbb{Z}}$. For the sake of simplicity, we define morphisms on A^*, but we consider implicitly their action on infinite and bi-infinite words. Notice that if σ is erasing, then the image of an infinite or bi-infinite word could be finite.

Let $\sigma : A^* \to A^*$ be a morphism. A finite, infinite or bi-infinite word x such that $\sigma(x) = x$ is said to be a *fixed point* of σ.

Definition 1.2.18 If there exist a letter $a \in A$ and a word $u \in A^+$ such that $\sigma(a) = au$ and moreover, if $\lim_{n \to +\infty} |\sigma^n(a)| = +\infty$, then σ is said to be (right) *prolongable* on a. Let $\sigma : A^* \to A^*$ be a morphism prolongable on a. We have

$$\sigma(a) = a\,u,\ \sigma^2(a) = a\,u\,\sigma(u),\ \sigma^3(a) = a\,u\,\sigma(u)\,\sigma^2(u),\ \ldots\ .$$

Since, for all $n \in \mathbb{N}$, $\sigma^n(a)$ is a prefix of $\sigma^{n+1}(a)$ and because $|\sigma^n(a)|$ tends to infinity when $n \to +\infty$, the sequence $(\sigma^n(a))_{n \geq 0}$ converges to an infinite word denoted by $\sigma^\omega(a)$ and given by

$$\sigma^\omega(a) := \lim_{n \to +\infty} \sigma^n(a) = a\,u\,\sigma(u)\,\sigma^2(u)\,\sigma^3(u) \cdots\ .$$

This infinite word is a fixed point of σ. An infinite word obtained in this

way by iterating a prolongable morphism is said to be *generated by* σ, and more generally, *purely substitutive* or *purely morphic*. In the literature, one also finds the term *pure morphic*. If $x \in A^{\mathbb{N}}$ is purely morphic and if $\tau : A \to B$ is a coding, then the word $y = \tau(x)$ is said to be *morphic* or *substitutive*.

Let $A = \{a, b, c\}$ and $\sigma\colon A^* \to A^*$ be the endomorphism defined by $\sigma(a) = a$, $\sigma(b) = bb$, $\sigma(c) = aab$. The morphism σ is not prolongable on the letter c but the sequence of words $(\sigma^n(c))_{n\geq 0}$ converges to the infinite word aab^ω, which is morphic but not purely morphic. For other examples of morphic words that are not purely morphic, see Example 4.6.5, Exercise 4.12, Exercise 4.13, and Proposition 4.7.2. See also Exercise 10.1.2 in the same vein. For an example of a purely morphic word that is fixed by no non-erasing endomorphism other than the identity, see Exercise 4.11.

We also consider bi-infinite morphic words.

Definition 1.2.19 If there exist a letter $b \in A$ and a word $u \in A^+$ such that $\sigma(b) = ub$ and moreover, if $\lim_{n\to+\infty} |\sigma^n(b)| = +\infty$, then σ is said to be left *prolongable* on b. We have

$$\sigma(b) = u\,b,\ \sigma^2(a) = \sigma(u)\,u\,b,\ \sigma^3(a) = \sigma^2(u)\,\sigma(u)\,u\,b,\ \ldots\ .$$

Let $\sigma : A^* \to A^*$ be a morphism that is both right prolongable on a and left prolongable on b. The sequence $(\sigma^n(b).\sigma^n(a))_{n\geq 0}$ converges to a bi-infinite word denoted by $\sigma^\omega(b).\sigma^\omega(a)$ which is a fixed point of σ. If furthermore, there exist a letter $c \in A$ and $\ell \in \mathbb{N}$ such that ba is a factor of $\sigma^\ell(c)$, then the bi-infinite word $\sigma^\omega(b).\sigma^\omega(a)$ is said to be generated by σ, and more generally *purely substitutive* or *purely morphic*. We similarly define as in Definition 1.2.18 a *substitutive* or *morphic* bi-infinite word.

Definition 1.2.20 For each morphism $\sigma : A^* \to B^*$, we define the *width* of σ, denoted by $\|\sigma\|$, as $\|\sigma\| := \max_{a\in A} |\sigma(a)|$.

It is clear that $|\sigma(w)| \leq \|\sigma\|\,|w|$ for every $w \in A^*$.

Let us note that a morphism $\sigma : A^* \to B^*$ is injective if, and only if, letters of A are mapped to distinct words and the language $\sigma(A)$ is a code. See (Lothaire 2002, Proposition 6.13).

Example 1.2.21 (Thue–Morse word) Consider the 2-uniform morphism defined over the alphabet $\{a, b\}$ by $\sigma : a \mapsto ab, b \mapsto ba$. The infinite (purely morphic) word

$$\sigma^\omega(a) = abbabaabbaababbabaababbaabbabaab\cdots$$

is the celebrated *Thue–Morse word*. This word can also be obtained as follows. Consider the morphism $\gamma : a \mapsto b, b \mapsto a$ and define the sequence of finite words $u_0 = a$ and, for all $n \geq 1$, $u_n = u_{n-1}\gamma(u_{n-1})$. It is an exercise to show that the sequence $(u_n)_{n\geq 0}$ converges to the Thue–Morse word. For more details on the Thue–Morse word, see Section 4.10.4.

Many properties of the Thue–Morse word can be found in the paper (Allouche and Shallit 1999). In several chapters of (Pytheas Fogg 2002), the Thue–Morse word or the Fibonacci word introduced below are also discussed in details.

Example 1.2.22 (Fibonacci word) Another consecrated example of purely morphic word is the *Fibonacci word*. It is obtained from the non-uniform morphism defined over the alphabet $\{a, b\}$ by $\sigma : a \mapsto ab, b \mapsto a$,

$$\sigma^\omega(a) = (x_n)_{n\geq 0} = abaababaabaababaababaabaababaabaababaababaa\cdots .$$

It is a Sturmian word and can be obtained as follows. Let $\varphi = (1 + \sqrt{5})/2$ be the Golden Ratio. For all $n \geq 1$, if $\lfloor (n+1)\varphi \rfloor - \lfloor n\varphi \rfloor = 2$, then $x_{n-1} = a$, otherwise $x_{n-1} = b$. For more details, see Section 4.10.3.

Example 1.2.23 (Squares) Consider the alphabet $A = \{a, b, c\}$ and the morphism $\sigma : A^* \to A^*$ defined by $\sigma : a \mapsto abcc, b \mapsto bcc, c \mapsto c$. We get

$$\sigma^\omega(a) = abccbccccbccccccbccccccccbccccccccccbcc\cdots .$$

Using the special form of the images of b and c, it is not difficult to see that the difference between the position of the nth b and the $(n + 1)$st b in $\sigma^\omega(a)$ is $2n+1$. Since the difference between two corresponding consecutive squares $(n+1)^2-n^2$ is also $2n+1$, if we define the coding $\tau : a, b \mapsto 1, c \mapsto 0$, we get exactly

$$\tau(\sigma^\omega(a)) = 1100100001000000100000000100000000000100\cdots$$

which proves that the characteristic sequence of the set of squares is morphic. One can show that this morphic sequence cannot be generated using a uniform morphism, for instance see (Eilenberg 1974) where it is shown that the set of squares is not k-recognisable. Also see Example 1.3.16.

Example 1.2.24 (Powers of 2) Consider the 2-uniform morphism defined over the alphabet $\{a, b, c\}$ by $\sigma : a \mapsto ab, b \mapsto bc, c \mapsto cc$ and the coding $\tau : a, c \mapsto 0, b \mapsto 1$. We have

$$\sigma^\omega(a) = abbcbcccbcccccccbcccccccccccccccbcc\cdots$$

and

$$\tau(\sigma^\omega(a)) = 01101000100000001000000000000000100\cdots.$$

Developing the same kind of arguments as in the previous example, this latter morphic word is easily seen to be the characteristic word of the set of powers of two. For more details on this morphic word, see Section 4.10.2.

For complements on morphisms, see Section 4.6.1. These notions will also be extended to the framework of D0L and HD0L systems in Chapter 10, see also Section 3.4.2, where a *D0L system* is a triple of the form (A, σ, w) where A is an alphabet, σ is a morphism of A^* and w is a word over A. A language D, that is a set of words, is called a *D0L language* if there exists a D0L-system (A, σ, w) such that $D = \{\sigma^k(w) \mid k \in \mathbb{N}\}$. A language H is called a *HD0L language* if there exist two alphabets A and B, a D0L language D over A and a morphism $\tau : A^* \to B^*$ such that $H = \tau(D)$.

1.3 Languages and machines

Formal languages theory is mostly concerned with the study of the mathematical properties of sets of words. For an exhaustive exposition on regular languages and automata theory, see (Sakarovitch 2003), or in the same spirit (Eilenberg 1974). See also the chapter (Yu 1997), or (Sudkamp 2005), (Hopcroft and Ullman 1979) and the updated revision (Hopcroft, Motwani, and Ullman 2006) for general introductory books on formal languages theory. In (Perrin 1990), the relationship of automata with recognisable sets of integers is presented. In this section, we do not present languages of infinite words and the corresponding automata crafted to recognise these languages, a reference is the book (Perrin and Pin 2003), see also (Thomas 1990). Notions presented here have been kept minimal, more definitions and results on finite automata and transducers can be found in Section 2.6.

1.3.1 Languages of finite words

Let A be an alphabet. A subset L of A^* is said to be a *language*. Note for instance that this terminology is consistent with the one of Definition 1.2.8. Since a language is a *set* of words, we can apply all the usual set operations like union, intersection or set difference: $\cup$, $\cap$ or $\setminus$. The concatenation of words can be extended to define an operation on languages. If L, M are languages, LM is the language of the words obtained by concatenation of

a word in L and a word in M, *i.e.*,

$$LM = \{uv \mid u \in L, v \in M\} .$$

We can of course define the concatenation of a language with itself, so it permits to introduce the power of a language. Let $n \in \mathbb{N}$, A be an alphabet and $L \subseteq A^*$ be a language. The language L^n is the set of words obtained by concatenating n words in L. We set $L^0 := \{\varepsilon\}$. In particular, we recall that A^n denotes the set of words of length n over A, *i.e.*, concatenations of n letters in A. The *(Kleene) star* of the language L is defined as

$$L^* = \bigcup_{i \geq 0} L^i .$$

Otherwise stated, L^* contains the words that are obtained as the concatenation of an arbitrary number of words in L. Notice that the definition of Kleene star is compatible with the notation A^* introduced to denote the set of finite words over A. We also write $L^{\leq n}$ as a shorthand for

$$L^{\leq n} = \bigcup_{i=0}^{n} L^i.$$

Note that if the empty word belongs to L, then $L^{\leq n} = L^n$. We recall that $A^{\leq n}$ is the set of words over A of length at most n. If L is a language, then $\mathrm{alph}(L)$ is the set of all letters which occur in the words of L, *i.e.*, $\mathrm{alph}(L) = \cup_{u \in L} \mathrm{alph}(u)$.

Example 1.3.1 Let $L = \{a, ab, aab\}$ and $M = \{a, ab, ba\}$ be two finite languages. We have $L^2 = \{aa, aab, aaab, aba, abab, abaab, aaba, aabab, aabaab\}$ and $M^2 = \{aa, aab, aba, abab, abba, baa, baab, baba\}$. One can notice that $\mathrm{Card}(L^2) = (\mathrm{Card}\, L)^2$ but $\mathrm{Card}(M^2) < (\mathrm{Card}\, M)^2$. This is due to the fact that all words in L^2 have a unique factorisation as concatenation of two elements in L but this is not the case for M, where $(ab)a = a(ba)$. We can notice that

$$L^* = \{a\}^* \cup \{a^{i_1} b a^{i_2} b \cdots a^{i_n} b a^{i_{n+1}} \mid \forall n \geq 1, i_1, \ldots, i_n \geq 1,\ i_{n+1} \geq 0\} .$$

Since languages are sets of (finite) words, a language can be either *finite* or *infinite*. For instance, a language L differs from $\emptyset$ or $\{\varepsilon\}$ if, and only if, the language L^* is infinite. Let L be a language, we set $L^+ = LL^*$. The mirror operation can also be extended from words to languages: $\tilde{L} = \{\tilde{u} \mid u \in L\}$.

Definition 1.3.2 A language is *prefix-closed* (respectively *suffix-closed*) if it contains all prefixes (respectively suffixes) of any of its elements. A language is *factorial* if it contains all factors of any of its elements.

Obviously, any factorial language is prefix-closed and suffix-closed. The converse does not hold. For instance, the language $\{a^n b \mid n > 0\}$ is suffix-closed but not factorial.

Example 1.3.3 The set of words over $\{0,1\}$ containing an even number of 1's is the language

$$\begin{aligned} E &= \{w \in \{0,1\}^* \mid |w|_1 \equiv 0 \pmod 2\} \\ &= \{\varepsilon, 0, 00, 11, 000, 011, 101, 110, 0000, 0011, \ldots\} \ . \end{aligned}$$

This language is closed under mirror, *i.e.*, $\tilde{L} = L$. Notice that the concatenation $E\{1\}E$ is the language of words containing an odd number of 1's and $E \cup E\{1\}E = E(\{\varepsilon\} \cup \{1\}E) = \{0,1\}^*$. Notice that E is neither prefix-closed, since $1001 \in E$ but $100 \notin E$, nor suffix-closed.

Similarly as for infinite words (see Definition 1.2.12), we can count the number of words of a language of a given length. A language L of A^* is said to have to have *bounded growth* if for every n there are less than k words of length n in L, for a fixed integer k. Such languages are also called *slender*. For more on slender languages, see, *e.g.*, Proposition 2.6.3 and Section 3.3.2.

If a language L over A can be obtained by applying to some finite languages a finite number of operations of union, concatenation and Kleene star, then this language is said to be a *regular language*. This generation process leads to *regular expressions* which are well-formed expressions used to describe how a regular language is built in terms of these operations. From the definition of a regular language, the following result is immediate.

Theorem 1.3.4 *The class of regular languages over A is the smallest subset of 2^{A^*} (for inclusion) containing the languages $\emptyset$, $\{a\}$ for all $a \in A$ and closed under union, concatenation and Kleene star.*

Example 1.3.5 For instance, the language L over $\{0,1\}$ whose words do not contain the factor 11 is regular. This language can be described by the regular expression $L = \{0\}^*\{1\}\{0,01\}^* \cup \{0\}^*$. Otherwise stated, it is generated from the finite languages $\{0\}$, $\{0,01\}$ and $\{1\}$ by applying union, concatenation and star operations. Its complement in A^* is also regular and is described by the regular expression $A^*\{11\}A^*$. The language E from Example 1.3.3 is also regular, we have the following regular expression $\{0\}^*(\{1\}\{0\}^*\{1\}\{0\}^*)^*$ describing E.

1.3.2 Automata

As we shall briefly explain in this section, the regular languages are exactly the languages recognised by finite automata.

Definition 1.3.6 A *finite automaton* is a labelled graph given by a 5-tuple $\mathcal{A} = (Q, A, E, I, T)$ where Q is the (finite) *set of states*, $E \subseteq Q \times A^* \times Q$ is the finite set of *edges* defining the *transition relation*, $I \subseteq Q$ is the set of *initial states* and T is the *set of terminal (or final) states*. A *path* in the automaton is a sequence

$$(q_0, u_0, q_1, u_1, \ldots, q_{k-1}, u_{k-1}, q_k)$$

such that, for all $i \in [\![0, k-1]\!]$, $(q_i, u_i, q_{i+1}) \in E$, $u_0 \cdots u_{k-1}$ is the *label* of the path. Such a path is *successful* if $q_0 \in I$ and $q_k \in T$. The language $L(\mathcal{A})$ *recognised* (or *accepted*) by $\mathcal{A}$ is the set of labels of all successful paths in $\mathcal{A}$.

Any finite automaton $\mathcal{A}$ gives a partition of A^* into $L(\mathcal{A})$ and $A^* \setminus L(\mathcal{A})$. When depicting an automaton, initial states are marked with an incoming arrow and terminal states are marked with an outgoing arrow. A transition like (q, u, r) is represented by a directed edge from q to r with label u, $q \xrightarrow{u} r$.

Example 1.3.7 In Figure 1.2 the automaton has two initial states p and r, three terminal states q, r and s. For instance, the word ba is recognised by the automaton. There are two successful paths corresponding to the label ba: (p, b, q, a, s) and (p, b, p, a, s). For this latter path, we can write $p \xrightarrow{b} p \xrightarrow{a} s$. On the other hand, the word $baab$ is not recognised by the automaton.

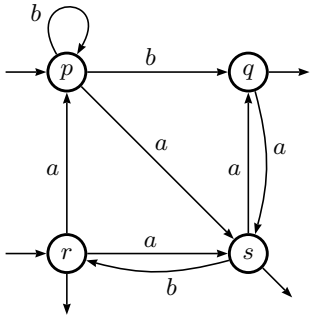

Fig. 1.2. A finite automaton.

Example 1.3.8 The automaton in Figure 1.3 recognises exactly the language E of the words having an even number of 1 from Example 1.3.3.

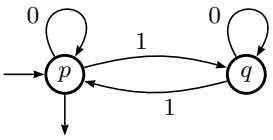

Fig. 1.3. An automaton recognising words with an even number of 1.

Definition 1.3.9 Let $\mathcal{A} = (Q, A, E, I, T)$ be a finite automaton. A state $q \in Q$ is *accessible* (respectively *co-accessible*) if there exists a path from an initial state to q (respectively from q to some terminal state). If all states of $\mathcal{A}$ are both accessible and co-accessible, then $\mathcal{A}$ is said to be *trim*.

Definition 1.3.10 A finite automaton $\mathcal{A} = (Q, A, E, I, T)$ is said to be *deterministic* (*DFA*) if it has only one initial state q_0, if E is a subset of $Q \times A \times Q$ and for each $(q, a) \in Q \times A$ there is at most one state $r \in Q$ such that $(q, a, r) \in E$. In that case, E defines a partial function $\delta_{\mathcal{A}} : Q \times A \to Q$ that is called the *transition function* of $\mathcal{A}$. The adjective *partial* means that the domain of $\delta_{\mathcal{A}}$ can be a strict subset of $Q \times A$. To express that the partial transition function is total, the DFA can be said to be *complete*. To get a total function, one can add to Q a new "sink state" s and, for all $(q, a) \in Q \times A$ such that δ_A is not defined, set $\delta_{\mathcal{A}}(q, a) := s$. This operation does not alter the language recognised by $\mathcal{A}$. We can extend $\delta_{\mathcal{A}}$ to be defined on $Q \times A^*$ by $\delta_{\mathcal{A}}(q, \varepsilon) = q$ and, for all $q \in Q$, $a \in A$ and $u \in A^*$, $\delta_{\mathcal{A}}(q, au) = \delta_{\mathcal{A}}(\delta_{\mathcal{A}}(q, a), u)$. Otherwise stated, the language recognised by $\mathcal{A}$ is $L(\mathcal{A}) = \{u \in A^* \mid \delta_{\mathcal{A}}(q_0, u) \in F\}$ where q_0 is the initial state of $\mathcal{A}$. If the automaton is deterministic, it is sometimes convenient to refer to the 5-tuple $\mathcal{A} = (Q, A, \delta_{\mathcal{A}}, I, T)$.

As explained by the following result, for languages of finite words, finite automata and deterministic finite automata recognise exactly the same languages.

Theorem 1.3.11 (Rabin and Scott 1959) *If L is recognised by a finite automaton $\mathcal{A}$, there exists a DFA which can be effectively computed from $\mathcal{A}$ and recognising the same language L.*

A proof and more details about classical results in automata theory can be found in textbooks like (Hopcroft, Motwani, and Ullman 2006), (Sakarovitch 2003) or (Shallit 2008). For standard material in automata theory we shall not refer again to these references below.

One important result is that the set of regular languages coincides with the set of languages recognised by finite automata.

Theorem 1.3.12 (Kleene 1956) *A language is regular if, and only if, it is recognised by a (deterministic) finite automaton.*

Observe that if L, M are two regular languages over A, then $L \cap M$, $L \cup M$, LM and $L \setminus M$ are also regular languages. In particular, a language over A is regular if, and only if, its complement in A^* is regular.

Example 1.3.13 The regular language $L = \{0\}^*\{1\}\{0,01\}^* \cup \{0\}^*$ from Example 1.3.5 is recognised by the DFA depicted in Figure 1.4. Notice that the state s is a *sink*: non-terminal state and all transitions remain in s.

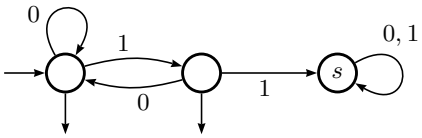

Fig. 1.4. A DFA accepting words without factor 11.

The following result is often useful to prove that a given language is not regular. It appeared first in (Bar-Hillel, Perles, and Shamir 1961).

Lemma 1.3.14 (Pumping lemma) *Let $L \subseteq A^*$ be a regular language accepted by a DFA with ℓ states. If $t \in L$ is a word of length $|t| \geq \ell$, then there exist $u, v, w \in A^*$ such that $t = uvw$, $|uv| \leq \ell$, $v \neq \varepsilon$, and $uv^*w \subseteq L$.*

The idea of the proof (pigeonhole principle) is that any path of length at least ℓ must contain a cycle. Let us first consider a simple example showing an application of the pumping lemma.

Example 1.3.15 Let us show that the language P of all the palindromes over an alphabet A of cardinality at least 2 is not regular. Assume that P is regular and accepted by a DFA with ℓ states. Consider the word $a^\ell b a^\ell \in P$, with a, b letters in A. With notation of Lemma 1.3.14, there exist i, j with $i \geq 0$, $j > 0$ and $i + j \leq \ell$, such that $u = a^i$, $v = a^j$, $w = a^{\ell-i-j}ba^\ell$ and for all $n \in \mathbb{N}$, $uv^nw = a^i a^{nj} a^{\ell-i-j} b a^\ell \in P$ which is a contradiction.

Example 1.3.16 The language of the decimal representations of squares $R = \{1, 4, 9, 16, 25, 36, \ldots\} \subset \{0, \ldots, 9\}^*$ is not regular. Assume to the contrary that R is regular. Notice that the square of the integer with 10^n1 as decimal representation has 10^n20^n1 as decimal representation. Since regular languages are closed under intersection, $R' = R \cap \{1\}\{0\}^*\{2\}\{0\}^*\{1\}$ must be regular. A careful inspection shows that

$$R' = \{10^n20^n1 \mid n \geq 0\} \ .$$

Indeed, it is left as an exercise to show that 10^i20^j1 is the decimal representation of a square if, and only if, $i = j$. Assume that R' is accepted by a DFA with ℓ states. Let us apply the pumping lemma to R'. The word $10^\ell20^\ell1$ belonging to R' can be factored as uvw with $|uv| \leq \ell$. But uv^kw does not belong to R' for $k > 1$ which leads to a contradiction.

The special case of morphic words obtained by q-uniform morphism were introduced by A. Cobham in his seminal paper (Cobham 1972). These infinite words are usually referred to as *q-automatic sequences*. We conclude this section on automata by explaining where does the term "automatic" come from. See also (Hopcroft and Ullman 1979) and the surveys (Allouche 1987), (Allouche and Mendès France 1995).

Definition 1.3.17 A *deterministic finite automaton with output* (*DFAO*) over an alphabet A is a 6-tuple $\mathcal{A} = (Q, A, \delta, \{q_0\}, B, \tau)$ where Q, δ and $\{q_0\}$ are defined as for DFA, $\delta : Q \times A \to Q$ being a total function, B is a finite alphabet and $\tau : Q \to B$ is the *output function*.

A DFAO acts like a map from A^* to B. With any word $w \in A^*$ is associated the output $\tau(\delta(q_0, w))$. If $B = \{b_1, \ldots, b_t\}$ then the DFAO $\mathcal{A}$ corresponds to a partition of A^* into t (regular) languages

$$L_i = \{w \in A^* \mid \tau(\delta(q_0, w)) = b_i\}, \quad i = 1, \ldots, t \ .$$

An infinite word $x = (x_n)_{n\geq 0} \in B^{\mathbb{N}}$ is said to be *k-automatic* if there exists a DFAO $(Q, \{0, \ldots, k-1\}, \delta, \{q_0\}, B, \tau)$ such that, for all n,

$$x_n = \tau(\delta(q_0, \mathrm{rep}_k(n)))$$

where $\mathrm{rep}_k(n)$ denotes the k-ary representation of n (see Section 1.6). Roughly speaking, the nth term of the sequence is obtained by feeding a DFAO with the k-ary representation of n. For a complete and comprehensive exposition on k-automatic sequences and their applications see the book (Allouche and Shallit 2003).

Theorem 1.3.18 (Cobham 1972) *Let $k \geq 2$. An infinite word $x \in A^{\mathbb{N}}$ is*

k-automatic if, and only if, there exist a k-uniform morphism $\sigma : B^* \to B^*$ prolongable on a letter $b \in B$ and a coding $\tau : B \to A$ such that $x = \tau(\sigma^\omega(b))$.

Example 1.3.19 We have seen in Example 1.2.21 that the Thue–Morse word is generated using a 2-uniform morphism. This word is also 2-automatic. Indeed, we can consider the automaton in Figure 1.3 as a DFAO where the output of the states p and q are respectively 0 and 1.

1.3.3 Transducers

Let A, B be two alphabets. A *transducer* is an automaton given by a 6-tuple $\mathcal{T} = (Q, A, B, E, I, T)$ whose transitions are labelled by elements in $A^* \times B^*$ instead of considering a unique alphabet A. We can therefore use the terminology introduced for automata. Notice that to obtain the label of a path, if $(u, v), (w, x) \in A^* \times B^*$, then the product in $A^* \times B^*$ is the concatenation component-wise, *i.e.*, $(u, v)(w, x) = (uw, vx)$. The *language accepted* by $\mathcal{T}$ is a subset of $A^* \times B^*$, *i.e.*, a relation from A^* into B^*.

It is common to encounter special cases of transducers. First, if labels of transitions belong to $A \times B$, then the transducer is said to be *letter-to-letter*. From a given transducer $\mathcal{T} = (Q, A, B, E, I, T)$, we get an automaton $\mathcal{T}' = (Q, A, E', I, T)$, the *underlying input automaton* of $\mathcal{T}$, where $(q, u, q') \in E'$ if, and only if, there exists $v \in B^*$ such that $(q, (u, v), q') \in E$. If $\mathcal{T}'$ is deterministic, the transducer $\mathcal{T}$ is said to be *sequential*.

When depicting an automaton recall that terminal states are marked by an outgoing arrow. We consider here transducers where the outgoing arrows, *i.e.*, the edges designating terminal states, can be labelled with pairs of the form (ε, w). The convention to define the relation realised by $\mathcal{T}$ is that if a path from $i \in I$ to $t \in T$ is labelled by (u, v) and if t has an outgoing arrow labelled by (ε, w), then (u, vw) belongs to the relation realised by $\mathcal{T}$. This can be seen as a shortcut to describe the following construction. Note that the labels of outgoing arrows are not involved in the possible sequentiality of the transducer as defined above. First, add a new terminal state t' that does not belong to the set of states of $\mathcal{T}$ and having only incident edges. For all terminal states $t \in T$ having a labelled outgoing arrow, add an edge with that label from t to t'. The new set of terminal states is the subset of T made of the terminal states with unlabelled outgoing arrow and t'. An example is given below, compare Figures 1.5 and 1.6. Another way of performing this is to to consider terminal states as functions, as in Section 2.6.

Finally, it is implicitly understood that concatenation of labels of transi-

tions is read from left to right. But if words are read from right to left, we speak of *right transducers*.

We recall that a primer on finite automata and transducers is given in Section 2.6.

Example 1.3.20 Consider the sequential right transducer depicted in Figure 1.5. Recall that the adjective "right" means that entries are read from right to left. For instance $(101, 1000)$ and $(1001, 1010)$ belong to the relation

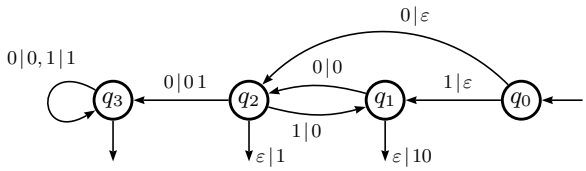

Fig. 1.5. A transducer.

realised by the right transducer. Indeed, we have the successful paths

$$q_0 \xrightarrow{1|\varepsilon} q_1 \xrightarrow{0|0} q_2 \xrightarrow{1|0} q_1 \xrightarrow{\varepsilon|10}$$

and

$$q_0 \xrightarrow{1|\varepsilon} q_1 \xrightarrow{0|0} q_2 \xrightarrow{0|01} q_3 \xrightarrow{1|1} q_4$$

where consecutive labels $(u_1, v_1), \ldots, (u_k, v_k)$, $u_i, v_i \in \{0, 1\}^*$ are concatenated from the right: $(u_k \cdots u_1, v_k \cdots v_1)$. An equivalent right transducer realising the same relation is given in Figure 1.6.

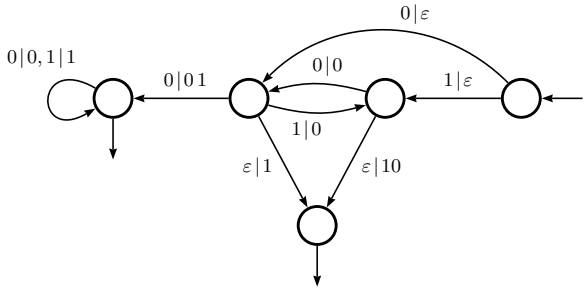

Fig. 1.6. Another transducer realising the same relation.

1.4 Associated matrices

Let $\mathbb{K}$ be a field. The set of matrices with r rows and c columns having entries in $\mathbb{K}$ is denoted by $\mathbb{K}^{r \times c}$. If entries are indexed by elements belonging

to two finite sets S and T, we write $\mathbb{K}^{S\times T}$. Let $\mathcal{A} = (Q, A, E, I, T)$ be an automaton such that $E \subseteq Q \times A \times Q$, *i.e.*, labels of edges are letters. An automaton being a directed graph, we can define its *adjacency matrix* $\mathbf{M} \in \mathbb{N}^{Q\times Q}$ indexed by $Q \times Q$ by

$$\mathbf{M}_{q,r} = \mathrm{Card}\{a \in A \mid (q, a, r) \in E\} \ .$$

If we are dealing with more than one automaton, we use notation like $\mathbf{M}(\mathcal{A})$ to specify the considered automaton. Using classical arguments from graph theory, one can shows that, for all $n \geq 0$, $(\mathbf{M}^n)_{q,r}$ counts the number of paths of length n from q to r. In particular, if $\mathcal{A}$ is deterministic, the element $(\mathbf{M}^n)_{q,r}$ is the number of words of length n which are labels of paths from q to r.

Example 1.4.1 Consider the automaton given in Figure 1.4. If states are ordered from left to right, we get the adjacency matrix

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 2 \end{pmatrix} .$$

It is easy to see that, for all $n \geq 1$,

$$\mathbf{M}^n = \begin{pmatrix} F_n & F_{n-1} & * \\ F_{n-1} & F_{n-2} & * \\ 0 & 0 & 2^n \end{pmatrix} .$$

where $F_{-1} = 0$, $F_0 = 1$ and $F_j = F_{j-1} + F_{j-2}$ for all $j \geq 1$. In particular, the number of words of length $n \geq 0$ over $\{0, 1\}$ not containing the factor "11" is $F_n + F_{n-1} = F_{n+1}$. Indeed, one has to count paths of length n from the initial state to one of the two terminal states. So we sum up the first two entries on the first row.

The same kind of idea can be applied to morphisms. Let $\sigma : A^* \to A^*$ be an endomorphism. The matrix $\mathbf{M}_\sigma \in \mathbb{N}^{A\times A}$ associated with σ is called the *incidence matrix* of σ and is defined by

$$\forall a, b \in A,\ (\mathbf{M}_\sigma)_{a,b} = |\sigma(b)|_a \ .$$

Let us recall that $\mathbf{P}$ stands for the abelianisation map. If $A = \{a_1, \ldots, a_d\}$, then the matrix $\mathbf{M}_\sigma$ can be defined by its columns:

$$\mathbf{M}_\sigma = \begin{pmatrix} \mathbf{P}(\sigma(a_1)) & \cdots & \mathbf{P}(\sigma(a_d)) \end{pmatrix}$$

and it satisfies:

$$\forall\, w \in A^*,\ \mathbf{P}(\sigma(w)) = \mathbf{M}_\sigma \mathbf{P}(w) \ .$$

A square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with entries in $\mathbb{R}_{\geq 0}$ is *irreducible* if, for all i, j, there exists k such that $(\mathbf{M}^k)_{i,j} > 0$. A square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with entries in $\mathbb{R}_{\geq 0}$ is *primitive* if there exists k such that, for all i, j, we have $(\mathbf{M}^k)_{i,j} > 0$. Similarly, a morphism over the alphabet A is *irreducible* if its incidence matrix is irreducible. A substitution is *primitive* if its incidence matrix is primitive.

The terminology irreducible comes from the fact that a matrix $\mathbf{M}$ is irreducible if, and only if, if it has no non-trivial invariant space of coordinates. Primitive matrices are also called *irreducible and aperiodic matrices*. See (Gantmacher 1960) or (Seneta 1981) for details on matrices with non-negative entries.

One checks that a primitive morphism always admits a power that is (left and right) prolongable, and which thus generates both an infinite word and a bi-infinite word which are fixed points of σ. See (Queffélec 1987, Proposition V.1).

Theorem 1.4.2 (Perron–Frobenius' theorem) *Let* $\mathbf{M}$ *be an irreducible matrix with non-negative entries. Then* $\mathbf{M}$ *admits a positive eigenvalue* α *which is larger than or equal in modulus to the other eigenvalues* λ*:* $\alpha \geq |\lambda|$*. The eigenvalue* α *and its algebraic conjugates (that is, the roots of the minimal polynomial of* α*) are simple roots of the characteristic polynomial of* $\mathbf{M}$ *and thus are simple eigenvalues. Furthermore, there exists an eigenvector with positive entries associated with* α*. The eigenvalue* α *is called the* Perron–Frobenius eigenvalue *of* $\mathbf{M}$*.*

Furthermore, if $\mathbf{M}$ *is* primitive*, then the eigenvalue* α *dominates (strictly) in modulus the other eigenvalues* λ*:* $\alpha > |\lambda|$*.*

Definition 1.4.3 Let σ be a morphism. If its incidence matrix $\mathbf{M}_\sigma$ is irreducible, then the Perron–Frobenius eigenvalue of $\mathbf{M}_\sigma$ is called the *inflation factor* of the morphism σ.

Lemma 1.4.4 *(Gantmacher 1960) Let* $\mathbf{M}$ *be an irreducible matrix with non-negative entries, and let* α *be its Perron–Frobenius eigenvalue. The inequality* $\alpha\mathbf{v} \leq \mathbf{M}\mathbf{v}$ *(considered as a component-wise inequality) either implies that* $\mathbf{v}$ *is an eigenvector associated with the Perron–Frobenius eigenvalue, or that* $\mathbf{v} = \mathbf{0}$*. In either case, we have* $\mathbf{M}\mathbf{v} = \alpha\mathbf{v}$*.*

As an application of Perron–Frobenius' theorem, we deduce the existence of frequencies for every factor of an infinite word generated by a primitive morphism. For a proof, see (Queffélec 1987) or Chapter 5 in (Pytheas Fogg 2002). For general results on frequencies of factors, see Chapter 7.

Theorem 1.4.5 *Let σ be a primitive prolongable morphism. Let u be an infinite word generated by σ. Then every factor of u has a frequency. Furthermore, all the frequencies of factors are positive. The frequencies of the letters are given by the coordinates of the positive eigenvector associated with the Perron–Frobenius eigenvalue, renormalised in such a way that the sum of its coordinates equals* 1.

The positive eigenvector associated with the Perron–Frobenius eigenvalue, renormalised in such a way that the sum of its coordinates equals 1 is usually called the *normalised Perron–Frobenius eigenvector* or *Perron–Frobenius eigenvector*. This is the normalisation choice that will be used in Chapter 10.

Furthermore, words generated by primitive morphisms are uniformly recurrent:

Proposition 1.4.6 *Let σ be a primitive morphism. For every $k \in \mathbb{N}$, any fixed point of σ^k is uniformly recurrent. Let σ be a morphism prolongable on the letter $a \in A$. We assume furthermore that all the letters in A actually occur in $\sigma^\omega(a)$ and that $\lim_{n\to+\infty} |\sigma^n(b)| = +\infty$ for all $b \in A$. The morphism σ is primitive if, and only if, the fixed point of σ beginning by a is uniformly recurrent.*

For an example of a morphism which is not primitive with a uniformly recurrent fixed point, consider $\sigma : 0 \mapsto 0010,\ 1 \mapsto 1$. The infinite word $\sigma^\omega(0)$ is called the *Chacon word* (see Exercise 1.8). One has $\lim_{n\to+\infty} |\sigma^n(1)| = 1$. Also see connections with Section 6.5.

The case where the Perron–Frobenius eigenvalue of the incidence matrix of a primitive morphism is a Pisot number is of particular interest.

Definition 1.4.7 An algebraic integer $\alpha > 1$, *i.e.*, a root of a monic polynomial with integer coefficients, is a *Pisot-Vijayaraghavan number* or a *Pisot number* if all its algebraic conjugates λ other than α itself satisfy $|\lambda| < 1$. An algebraic integer is a *unit* if its norm equals 1, *i.e.*, if the constant term of its minimal polynomial equals 1 in absolute value.

We recall that the algebraic conjugates of an algebraic integer are the roots of its minimal polynomial.

A primitive morphism σ is said to be *Pisot* if its Perron–Frobenius eigenvalue is a Pisot number.

A Pisot morphism σ is said to be *unit* if its Perron–Frobenius eigenvalue is a unit Pisot number.

Example 1.4.8 Consider the Fibonacci morphism σ introduced in Example 1.2.22. The incidence matrix of σ is

$$\mathbf{M}_\sigma = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Since $\mathbf{M}_\sigma^2$ contains only positive entries, the morphism is primitive. The Perron–Frobenius eigenvalue of $\mathbf{M}_\sigma$ is the Golden Ratio $\varphi = (1+\sqrt{5})/2$ satisfying $\varphi^2 - \varphi - 1 = 0$. This algebraic integer has $(1-\sqrt{5})/2$ as Galois conjugate which is of modulus less than 1. Consequently, we have a unit Pisot morphism.

A Pisot morphism σ is said to be a *Pisot irreducible substitution* if the algebraic degree of the Perron–Frobenius eigenvalue of its incidence matrix is equal to the size of the alphabet. This is equivalent to the fact that the characteristic polynomial of its incidence matrix is irreducible. A Pisot morphism which is not a Pisot irreducible morphism is called a *Pisot reducible morphism*. Examples of Pisot reducible morphisms are $1 \to 12$, $2 \to 3$, $3 \to 4$, $4 \to 5$, $5 \to 1$ and the Thue–Morse morphism $1 \to 12$, $2 \to 21$. Indeed, the characteristic polynomial of the incidence matrix is respectively equal to $X^5 - X^4 - 1 = (X^2 - X + 1)(X^3 - X - 1)$ and to $X^2 - 2X = X(X-2)$.

Theorem 1.4.9 *Let σ be a morphism such that its incidence matrix $\mathbf{M}_\sigma$ is irreducible. If its Perron–Frobenius eigenvalue α is such that for every other eigenvalue λ of $\mathbf{M}_\sigma$ one has $\alpha > 1 > |\lambda| > 0$, then σ is primitive and Pisot irreducible.*

For a proof, see (Canterini and Siegel 2001b) and (Pytheas Fogg 2002, Chapter 1).

We end this section with the notion of spectral radius that will de developed in Chapter 11, also see Section 4.7.2.2. Let $\|\cdot\|$ be a submultiplicative matrix norm. That is a vector norm that satisfies for all square matrices $\mathbf{A}$, $\mathbf{B}$, $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$. Note that some authors use the terminology matrix norm only for those norms which are submultiplicative. The *spectral radius* of the complex square matrix $\mathbf{A}$ is defined as the largest modulus of its eigenvalues. It is proved to represent the asymptotic growth rate of the norm of the successive powers of A:

$$\rho(\mathbf{A}) = \lim_{t\to\infty} \|\mathbf{A}^t\|^{1/t}.$$

This quantity does provably not depend on the used norm.

1.5 A glimpse at numeration systems

Various numeration systems will be considered in details in this book, see mainly Chapters 2 and 3: integer base, real base, rational base, canonical number systems, abstract numeration systems. As a short appetiser, we merely recall in this section how to write down non-negative integers in the usual p-ary numeration system, $p \geq 2$ being an integer. More details are given in Section 2.2.1.

For any positive integer n, there exist $\ell \geq 0$ such that $p^\ell \leq n < p^{\ell+1}$ and unique coefficients $c_0, \ldots, c_\ell \in \{0, \ldots, p-1\}$ such that

$$n = \sum_{i=0}^{\ell} c_i \, p^i \text{ and } c_\ell \neq 0 \ .$$

The coefficients $c_\ell, \ldots, c_0$ can be computed by successive Euclidean divisions. Set $n_0 := n$. We have $n_0 = c_\ell \, p^\ell + n_1$ with $n_1 < p^\ell$ and for $i = 1, \ldots, \ell$, $n_i = c_{\ell-i} \, p^{\ell-i} + n_{i+1}$ with $n_{i+1} < p^{\ell-i}$. The word $c_\ell \cdots c_0$ is said to be the *p-ary representation* or *p-expansion* of n (sometimes called *greedy* representation) and we write

$$\mathrm{rep}_p(n) = c_\ell \cdots c_0 \ .$$

Longer developments are given Section 2.2.1 of Chapter 2 where $\mathrm{rep}_p(n)$ is denoted by $\langle n \rangle_p$. We set $\mathrm{rep}_p(0) = \varepsilon$. So rep_p is a one-to-one correspondence between $\mathbb{N}$ and $\{\varepsilon\} \cup \{1, \ldots, p-1\}\{0, \ldots, p-1\}^*$.

Let $A \subset \mathbb{Z}$ be a finite alphabet and $u = a_0 \cdots a_\ell$ be a word over A. We set

$$\mathrm{val}_p(a_0 \cdots a_\ell) = \sum_{i=0}^{\ell} a_{\ell-i} \, p^i \ .$$

We say that $\mathrm{val}_p(u)$ is the *numerical value* or *evaluation* of u, also denoted by $\pi_p(u)$ (see, *e.g.*, Chapter 2).

The restriction of $\mathrm{rep}_p \circ \mathrm{val}_p$ to the set of words over A having a non-negative numerical value is the *normalisation*:

$$\nu_{A,p} : u \in A^* \mapsto \mathrm{rep}_p(\mathrm{val}_p(u)) \ .$$

Again, reference to the alphabet A can be omitted if the context is clear.

Example 1.5.1 (Signed digits) Let $A = \{\overline{1}, 0, 1\}$ where $\overline{1}$ stands for -1. We have $\mathrm{val}_2(100\overline{1}) = 7$ and $\mathrm{val}_2(\overline{1}01) = -3$. In particular, $\mathrm{rep}_2(\mathrm{val}_2(100\overline{1})) = 111$, *i.e.*, $\nu_2(100\overline{1}) = 111$.

Definition 1.5.2 A set $X \subseteq \mathbb{N}$ of integers is *p-recognisable* if the language

$$\mathrm{rep}_p(X) = \{\mathrm{rep}_p(n) \mid n \in X\}$$

is regular. Observe that a set X is p-recognisable if, and only if, its characteristic word is p-automatic.

Proposition 1.5.3 *Let $p \geq 2$. Any eventually periodic set of integers is p-recognisable.*

It is an easy exercise. See for instance (Sakarovitch 2003, Prologue) for a proof. The realisation of division by finite automata (together with arithmetic operations modulo q) is discussed in Chapter 2. Also see Proposition 3.1.9 for similar considerations.

Definition 1.5.4 Two integers $p, q \geq 2$ are *multiplicatively independent* if the only integers m, n satisfying $p^m = q^n$ are $m = n = 0$. Otherwise, p and q are said *multiplicatively dependent*. In other words, p and q are multiplicatively dependent if, and only if, $\log p / \log q$ is rational.

Theorem 1.5.5 (Cobham 1969) *Let $p, q \geq 2$ be two multiplicatively independent integers. If $X \subseteq \mathbb{N}$ is both p-recognisable and q-recognisable, then X is eventually periodic.*

Many efforts have been made to get a simpler presentation of Cobham's theorem, as G. Hansel did in (Hansel 1982). Also see (Perrin 1990), (Allouche and Shallit 2003) and (Rigo and Waxweiler 2006).

Several aspects of numeration systems are treated in Chapter 2. Various numeration systems for the representation of integers are discussed in (Fraenkel 1985). The chapter by Ch. Frougny in (Lothaire 2002) presents non-standard numeration systems for the representations of integers as well as β-numeration systems. The surveys (Barat, Berthé, Liardet, et al. 2006) and (Bruyère, Hansel, Michaux, et al. 1994) are also of interest and contain many pointers to the existing literature. The latter one develops also a logical characterisation of p-recognisable sets in terms of an extension of the Presburger arithmetic $\langle \mathbb{N}, + \rangle$ and extension of Cobham's theorem on the base dependence to the multidimensional case.

1.6 Symbolic dynamics

Let us introduce some basic notions in symbolic dynamics. For expository books on the subject, see (Cornfeld, Fomin, and Sinaĭ 1982), (Kitchens 1998), (Lind and Marcus 1995), (Perrin 1995b), (Queffélec 1987) and (Kůrka 2003).

1.6.1 Subshifts

Let S denote the following map defined on $A^{\mathbb{N}}$, called the *one-sided shift*:

$$S((x_n)_{n\geq 0}) = (x_{n+1})_{n\geq 0} \ .$$

In particular, if $x = x_0x_1x_2\cdots$ is an infinite word over A, then, for all $n \geq 0$, its suffix $x_nx_{n+1}\cdots$ is simply $S^n(x)$. Note that for convenience, the shift is sometimes denoted by σ, when no misunderstanding with morphisms on words can be made. This latter convention is used in Chapter 2. The map S is uniformly continuous, onto but not one-to-one on $A^{\mathbb{N}}$. This notion extends in a natural way to $A^{\mathbb{Z}}$. In this latter case, the shift S is one-to-one. The definitions given below correspond to the *one-sided shift*, but they extend to the *two-sided shift*.

Definition 1.6.1 Let x be an infinite word over the alphabet A. The *orbit* of x under the action of the shift S is defined as the set

$$\mathcal{O}(x) = \{S^n x \mid n \in \mathbb{N}\} \ .$$

The *symbolic dynamical system* associated with x is then defined as $(\overline{\mathcal{O}(x)}, S)$, where $\overline{\mathcal{O}(x)} \subseteq A^{\mathbb{N}}$ is the closure of the orbit of x.

In the case of bi-infinite words we similarly define $\mathcal{O}(x) = \{S^n x \mid n \in \mathbb{Z}\}$ where the (two-sided) shift map is defined on $A^{\mathbb{Z}}$. The set $X_x := \overline{\mathcal{O}(x)}$ is a closed subset of the compact set $A^{\mathbb{N}}$, hence it is a compact space and S is a continuous map acting on it. One checks that, for every infinite word $y \in A^{\mathbb{N}}$, the word y belongs to X_x if, and only if, $L(y) \subseteq L(x)$. For a proof, see (Queffélec 1987) or Chapter 1 of (Pytheas Fogg 2002). Note that $\overline{\mathcal{O}(x)}$ is finite if, and only if, x is eventually periodic.

More generally, let Y be a closed subset of $A^{\mathbb{N}}$ that is stable under the action of the shift S. The system (Y, S) is called a *subshift*. The *full shift* is defined as $(A^{\mathbb{N}}, S)$.

A subshift (X, S) is said *periodic* if there exist $x \in X$ and an integer k such that $X = \{x, Sx, \ldots, S^k x = x\}$. Otherwise it is said to be *aperiodic*.

If (Y, S) is a subshift, then there exists a set $X \subseteq A^*$ such that for every $u \in A^{\mathbb{N}}$, the word u belongs to Y if, and only if, $L(u) \cap X = \emptyset$. A subshift Y is said to be of *finite type* if the set $X \subseteq A^*$ is finite. A subshift is said to be *sofic* if the set X is a regular language.

Example 1.6.2 The set of infinite words over $\{0, 1\}$ which do not contain the factor 11 is a subshift of finite type, whereas the set of infinite words over $\{0, 1\}$ having an even number of 1's between two occurrences of the letter 0 is a sofic subshift which is not of finite type.

Definition 1.6.3 Let $x \in A^{\mathbb{N}}$. For a word $w = w_0 \cdots w_r$, the *cylinder set* $[w]_x$ is the set $\{y \in X_x \mid y_0 = w_0, \cdots, y_r = w_r\}$. If the context is clear, the subscript x will be omitted.

The cylinder sets are *clopen* (open and closed) sets and form a basis of open sets for the topology of X_x. Furthermore, one checks that a clopen set is a finite union of cylinders. In the bi-infinite case the cylinders are the sets $[u.v]_x = \{y \in X_x | y_i = u_i, y_j = v_j,\ -|u| \leq i \leq -1,\ 0 \leq j \leq |v| - 1\}$ and the same remarks hold.

1.6.2 Dynamical systems

We have introduced the notions of a symbolic dynamical system and of a subshift. Such discrete systems belong to the larger class of topological dynamical systems, which have been intensively studied in topological dynamics. See for instance (Cornfeld, Fomin, and Sinaĭ 1982). For references on ergodic theory, see *e.g.* (Walters 1982) or (Silva 2008).

Definition 1.6.4 A *topological dynamical system* (X, T) is defined as a compact metric space X together with a continuous map T defined onto the set X.

A topological dynamical system (X, T) is *minimal* if, for all x in X, the orbit of x, *i.e.*, the set $\{T^n x \mid n \in \mathbb{N}\}$, is dense in X.

Let us note that if (X, S) is a subshift, and if X is furthermore assumed to be minimal, then X is periodic if, and only if, X is finite.

The symbolic dynamical system (X_x, S) associated with the infinite word x is minimal if, and only if, for every $y \in X_x$, $L(y) = L(x)$. More generally, properties of symbolic dynamical systems associated with an infinite word are strongly related to its combinatorial properties. For a proof of Theorem 1.6.5 below, see for instance Chapter 5 of (Pytheas Fogg 2002).

Theorem 1.6.5 *Let x be an infinite word. If x is recurrent, then the shift $S\colon X_x \to X_x$ is onto. Furthermore, (X_x, S) is minimal if, and only if, x is uniformly recurrent.*

In other words, x is uniformly recurrent if, and only if, $L(y) = L(x)$ for every y such that $L(y) \subseteq L(x)$. The idea of the proof of the equivalence in Theorem 1.6.5 can be sketched as follows: if w is a factor of x, we write

$$\overline{\mathcal{O}(x)} = \bigcup_{n \in \mathbb{N}} S^{-n}[w] \ ,$$

and we conclude by a compactness argument.

Two dynamical systems (X_1, T_1) and (X_2, T_2) are said to be *topologically conjugate* (or *topologically isomorphic*) if there exists an homeomorphism f from X_1 onto X_2 which conjugates T_1 and T_2, that is:

$$f \circ T_1 = T_2 \circ f \ .$$

Let (X, T) be a topological dynamical system. Let $\mathcal{M}(X)$ stand for the set of Borel probability measures on X. A Borel measure μ defined over X is said T-*invariant* if $\mu(T^{-1}(B)) = \mu(B)$, for every Borel set B. The map T is said to preserve the measure μ. This is equivalent to the fact that for any continuous function $f \in \mathcal{C}(X)$, then $\int f(Tx)\, d\mu(x) = \int f(x)\, d\mu(x)$. A topological system (X, T) always has an invariant probability measure. For more details, see Proposition 7.2.4.

The case where there exists only one T-invariant measure is of particular interest. A topological dynamical system (X, T) is said to be *uniquely ergodic* if there exists one and only one T-invariant Borel probability measure over X.

We have considered here the notion of dynamical system, that is, a map acting on a given set, in a topological context. This notion can be extended to measurable spaces: we thus get measure-theoretic dynamical systems. For more details about all of the notions defined in this section, one can refer to (Walters 1982).

Definition 1.6.6 A *measure-theoretic dynamical system* is defined as a system $(X, T, \mu, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra, μ a probability measure defined on $\mathcal{B}$, and $T : X \to X$ is a measurable map which preserves the measure μ, *i.e.*, for all $B \in \mathcal{B}$, $\mu(T^{-1}(B)) = \mu(B)$.

A measure-theoretic dynamical system $(X, T, \mu, \mathcal{B})$ is *ergodic* if for every $B \in \mathcal{B}$ such that $T^{-1}(B) = B$, then B has either zero measure or full measure.

In particular, a uniquely ergodic topological dynamical system yields an ergodic measure-theoretic dynamical system.

A measure-theoretic ergodic dynamical system satisfies the *Birkhoff ergodic theorem*, also called *individual ergodic theorem*. Let us recall that the abbreviation a.e. stands for "almost everywhere": a property holds almost everywhere if the set of elements for which the property does not hold is contained in a set of zero measure.

Theorem 1.6.7 (Birkhoff Ergodic Theorem) *Let $(X, T, \mu, \mathcal{B})$ be a measure-theoretic dynamical system. Let $f \in L^1(X, \mathbb{R})$. Then the sequence $(\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k)_{n\geq 0}$ converges a.e. to a function $f^* \in L^1(X, \mathbb{R})$. One has*

$f^* \circ T = f^*$ a.e. and $\int_X f^* \, d\mu = \int_X f \, d\mu$. Furthermore, if T is ergodic, since f^* is a.e. constant, one has:

$$\forall f \in L^1(X, \mathbb{R}) \ , \quad \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \xrightarrow[n \to \infty]{\mu - a.e.} \int_X f \, d\mu \ .$$

The notion of conjugacy between two topological dynamical systems extends in a natural way to this context. Two measure-theoretic dynamical systems $(X_1, T_1, \mu_1, \mathcal{B}_1)$ and $(X_2, T_2, \mu_2, \mathcal{B}_2)$ are said to be *measure-theoretically isomorphic* if there exist two sets of full measure $B_1 \in \mathcal{B}_1$, $B_2 \in \mathcal{B}_2$, a measurable map $f : B_1 \to B_2$ called *conjugacy map* such that

- the map f is one-to-one and onto,
- the reciprocal map of f is measurable,
- $f \circ T_1(x) = T_2 \circ f(x)$ for every $x \in B_1 \cap T_1^{-1}(B_1)$,
- μ_2 is the image of the measure μ_1 with respect to f, that is,

$$\forall B \in \mathcal{B}_2, \ \mu_1(f^{-1}(B \cap B_2)) = \mu_2(B \cap B_2) \ .$$

If the map is f is only onto, then $(X_2, T_2, \mu_2, \mathcal{B}_2)$ is said to be a *measure-theoretic factor* of $(X_1, T_1, \mu_1, \mathcal{B}_1)$.

1.6.3 Substitutive dynamical systems

As a class of examples, let us consider symbolic dynamical systems associated with purely substitutive words. Note that for such symbolic dynamical systems, one rather uses the terminology "substitution" than the terminology "morphism".

First we recall that if σ is a primitive substitution, then there exists a power of σ that is prolongable, and thus, which generates an infinite word (see Definition 1.2.18). Similarly, there exists a power of σ which generates a bi-infinite word which is purely morphic in the sense of Definition 1.2.19. For more details, see (Queffélec 1987, Proposition V.1). We then deduce from Proposition 1.4.6 and Theorem 1.6.5 that if σ is a primitive prolongable substitution, then all the (infinite or bi-infinite) words generated by σ are uniformly recurrent, and thus have the same language. In other words, all the symbolic dynamical systems associated with any of the words generated by one of the powers of σ do coincide. Hence, we can associate in a natural way with a primitive substitution a symbolic dynamical system.

Definition 1.6.8 Let σ be a primitive substitution. The *symbolic dynamical system associated with* σ is the system associated with any of the (infinite or bi-infinite) words generated by one of the powers of σ, according respectively to Definitions 1.2.18 and 1.2.19. We denote it by (X_σ, S).

Let us quote an interesting property of substitutive dynamical systems associated with a primitive substitution. For more details, see (Queffélec 1987) and (Pytheas Fogg 2002).

Theorem 1.6.9 *Let σ be a primitive substitution. The system (X_σ, S) is* uniquely ergodic.

The corresponding invariant measure is uniquely defined by its values on the cylinders: the measure of the cylinder $[w]$ is defined as the frequency of the finite word w in any element of X_σ, which does exist and does not depend on the choice of the fixed point, according to Theorem 1.4.5.

1.7 Exercises

Exercise 1.1 Show that over a binary alphabet, any word of length ≥ 4 contains a square as factor.

Exercise 1.2 Show that over a ternary alphabet A, it is possible to build an infinite word avoiding squares. See for instance (Lothaire 1983). As a by-product, show that the set of (finite) words over A having a non-empty factor which is a square, is not regular.

Exercise 1.3 Show that a word u of even (respectively odd) length is a palindrome if, and only if, there exists a word v such that $u = v\tilde{v}$ (respectively there exist a word v and a letter a such that $u = va\tilde{v}$).

Exercise 1.4 Show that if L and M are regular languages then $L \cap M$ is also regular.

Exercise 1.5 Let L be a language over the unary alphabet $\{a\}$. Show that L is regular if, and only if, there exists an eventually periodic set $X \subseteq \mathbb{N}$ such that $L = \{a^i \mid i \in X\}$.

Exercise 1.6 Let L be a regular language over A. Show that $|L| = \{|u| : u \in L\} \subseteq \mathbb{N}$ is eventually periodic. Give a counter-example illustrating that the converse does not hold.

Exercise 1.7 Prove that a Sturmian word is recurrent. Give an example of a bi-infinite word x with complexity function satisfying $p_x(n) = n+1$ for all n that is not recurrent.

Exercise 1.8 (Chacon word) We recall that the Chacon morphism σ is defined over the alphabet $\{0,1\}$ by $\sigma : 0 \mapsto 0010,\ 1 \mapsto 1$. Prove that the Chacon word $\sigma^\omega(0)$ begins with the following sequence of words $(b_n)_{n\geq 0}$:

$$b_0 = 0, \text{ and } \forall n \in \mathbb{N},\ b_{n+1} = b_n b_n 1 b_n.$$

Deduce that the Chacon word is uniformly recurrent.

Exercise 1.9 Give an example of a morphism that is irreducible but not primitive.

1.8 Notes

The study of combinatorics on words can be traced back to the work of A. Thue (Thue 1906) in 1906 and later on (Thue 1912) where he investigated repetitions in words, also see (Berstel 1995), then rooted in the papers (Morse and Hedlund 1938), (Morse and Hedlund 1940). Later on, impulsion was given on the one hand by M.-P. Schützenberger in France and on the other hand by P. S. Novikov and S. I. Adjan in former Russia. Now *combinatorics on words* is considered as a research topic by itself and has received classification subject `68R15` by the American Mathematical Society. For a comprehensive survey on the origins of combinatorics on words, see (Berstel and Perrin 2007).

For a nice account on the history of automata theory, see (Perrin 1995a). A first reference to automata can be traced back to (McCulloch and Pitts 1943). The notion of regular expressions goes back to (Kleene 1956) and non-deterministic automata were introduced in (Rabin and Scott 1959).