1

Preliminaries

Valérie Berthé,

Michel Rigo

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 throughout this book. We assume the reader to be 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

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

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

Combinatorics, Automata and Number Theory, ed. Valérie Berthé and Michel Rigo. Published by Cambridge University Press.

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 centre $\mathbf{x}$ 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 sometimes give 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 $6x^2/5$.

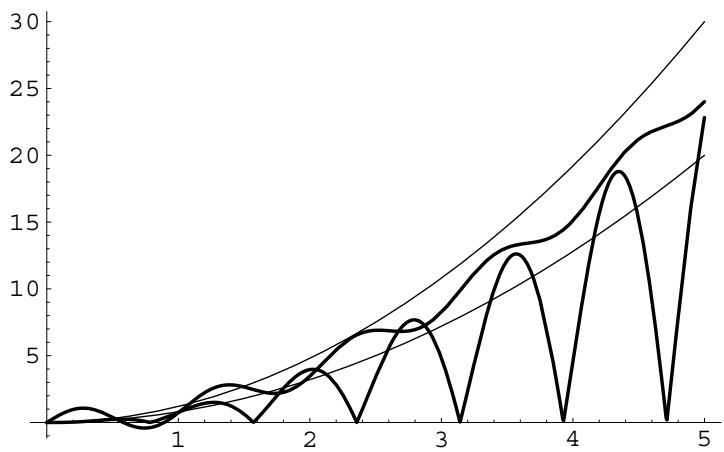

Fig. 1.1 The functions $x^2 + \sin 6x$, $x^2|\sin(4x)|$, $4x^2/5$ and $6x^2/5$.

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).

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) also contain 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 relevant 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 to be *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 detail 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 $p_x(n)$ by one whether 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 $\{a^n b \mid n\ge 0\}$ 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 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 $x = \cdots x_{-2}x_{-1}.x_0x_1x_2 \cdots$ 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 extends 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