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

3

Abstract numeration systems

Pierre Lecomte, Michel Rigo,

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

The primary motivation for the introduction of the abstract numeration systems stems from the celebrated theorem of Cobham dating back to 1969 about the so-called recognisable sets of integers in any integer base numeration system. Representations of numbers are words over a finite alphabet. There is a one-to-one correspondence between the sets of numbers and the languages made of the corresponding representations. Hence it is natural to consider questions related to formal language theory. In particular, we study sets of integers corresponding to regular languages. The different sections of this chapter are largely independent. However, Section 3.2 presents basic concepts and notation used in all later sections. The main focus is on the representation of integers. Extension to abstract numeration systems of the notion of recognisable sets of integers is studied in Section 3.3. In particular, we present some results about the stability of recognisability after multiplication by a constant. This requires us to discuss the complexity (or counting) function of regular languages. Section 3.4 is about the extension — to any substitutive sequence — of Cobham's theorem from 1972 about the equality of the set of infinite k-automatic words and the set of images under codings of the fixed points of substitutions of constant length k. The notion of an $\mathcal{S}$-automatic sequence is then introduced and various applications to $\mathcal{S}$-recognisability are considered. This chapter ends with a discussion about the representation of real numbers using abstract numeration systems.

3.1 Motivations

The primary role of a numeration system is to replace numbers which by essence are abstract objects by their representations which are words over suitable alphabets. As an example, the k-ary system replaces integers by

their representations in base k. Denote by B_k the set

$$\{0,\dots,k-1\}^* \setminus (0\{0,\dots,k-1\}^*)$$

of words over $\{0,\dots,k-1\}$ not starting with 0. The one-to-one correspondence mapping a non-negative integer n onto its k-ary representation $\mathrm{rep}_k(n) \in B_k$, also denoted $\langle n\rangle_k$ in Chapter 2, can be extended to a one-to-one correspondence between $2^{\mathbb{N}}$ and 2^{B_k}: any set $X \subseteq \mathbb{N}$ is associated with the language $\mathrm{rep}_k(X)$ made up from the k-ary representations of the numbers in X. It is therefore natural to study the relationship existing between the arithmetic properties of integers and the syntactical properties of the corresponding representations in a given numeration system. From the point of view of formal language theory, one can focus on those sets $X \subseteq \mathbb{N}$ for which a (deterministic) finite automaton can be used to decide for any given word w over $\{0,\dots,k-1\}$ whether or not w belongs to $\mathrm{rep}_k(X)$. Sets having such a property are called *k-recognisable* sets. In some sense, a k-recognisable set can be considered as particularly simple because through the k-ary numeration system it has a simple algorithmic description. Recall that in the *Chomsky hierarchy*, see for instance (Sudkamp 2005), (Shallit 2008), deterministic finite automata accepting regular languages are the simplest model of computation. However, dealing with k-recognisable sets has a price. As observed by A. Cobham, see Theorem 1.5.5: k-recognisability depends heavily on the choice of the base and sets which are k-recognisable for all $k \geq 2$ are exactly the eventually periodic sets. For that matter, also see Chapter 2 and in particular Subsection 2.2.4.

First we recall the notion of representation with respect to a U-system. Also see Section 2.3.3.

Definition 3.1.1 Let us extend the notion of k-ary numeration system by replacing the sequence $(k^n)_{n\geq 0}$ with some increasing sequence $U = (U_n)_{n\geq 0}$ of integers such that $U_0 = 1$. Using successive Euclidean divisions, we define the *U-representation* of any positive integer n. Let ℓ be such that $U_\ell \leq n < U_{\ell+1}$. We can greedily decompose n in a unique way as

$$n = \sum_{k=0}^{\ell} c_k\, U_k \ \text{ with } c_\ell \neq 0 \text{ and } \sum_{k=0}^{i} c_k\, U_k < U_{i+1},\ \forall i \in \{0,\dots,\ell\}\ . \quad (3.1)$$

This latter greedy condition implies that, for all $i \in \{0,\dots,\ell\}$, we have

$$c_i \in \{0,\dots,\lceil U_{i+1}/U_i\rceil - 1\}\ .$$

The *U-representation* of n is $c_\ell \cdots c_0$ and is denoted by $\mathrm{rep}_U(n)$. We set

$\mathrm{rep}_U(0) := \varepsilon$. Since we are interested in language theoretic properties related to U-representations, we assume moreover that the set $\{U_{i+1}/U_i \mid i \geq 0\}$ is bounded from above to ensure that $\mathrm{rep}_U(\mathbb{N}) = \{\mathrm{rep}_U(n) \mid n \in \mathbb{N}\}$ is a language over a finite alphabet. We set A_U to be the minimal (or canonical) alphabet of this language, *i.e.*, $A_U = \mathrm{alph}(\mathrm{rep}_U(\mathbb{N}))$. We can similarly to the integer base systems define the notion of U-recognisable sets. A set $X \subseteq \mathbb{N}$ is said to be *U-recognisable*, if $\mathrm{rep}_U(X)$ is accepted by a DFA.

These systems can be referred as *positional numeration systems* and the corresponding sequence U is usually called the *scale* or the *basis* of the system. The greediness of the U-representations implies the next proposition. We recall Definition 1.2.15 for the definition of the genealogical ordering $\prec$.

Proposition 3.1.2 *For all $m, n \in \mathbb{N}$, we have*

$$m < n \Leftrightarrow \mathrm{rep}_U(m) \prec \mathrm{rep}_U(n)$$

where the genealogical ordering $\prec$ is induced by the natural ordering of the alphabet $A_U \subset \mathbb{N}$.

Definition 3.1.3 In what follows, in particular for Propositions 3.1.5 and 3.1.9, when speaking of a *numeration system* $U = (U_n)_{n\geq 0}$ we assume that U is increasing, that $U_0 = 1$ and that the set $\{U_{i+1}/U_i \mid i \geq 0\}$ is bounded.

Amongst the possibly U-recognisable subsets of $\mathbb{N}$, the whole set $\mathbb{N}$ is of special interest. It seems natural to consider numeration systems $U = (U_n)_{n\geq 0}$ for which $\mathrm{rep}_U(\mathbb{N})$ is regular, *i.e.*, for which $\mathbb{N}$ is U-recognisable. In that case, we have an algorithm using a constant amount of memory and working in time proportional to the length of the input — namely a DFA — to check whether or not any given word over A_U is a valid U-representation. Let us investigate a little bit further what is implied by the U-recognisability of $\mathbb{N}$. We recall that a multi-graph is a graph which is permitted to have multiple edges, that is, edges that connect the same pair of vertices. First we start with the following lemma whose proof can be compared with the proof of Proposition 2.6.2.

Lemma 3.1.4 *Let $G = (V, E)$ be a directed finite multi-graph where V is the set of vertices of G, E is its multi-set of arcs in $V \times V$ and let $q, r \in V$. The map $\mathcal{U}_{q,r} : \mathbb{N} \to \mathbb{N}$ counting the number $\mathcal{U}_{q,r}(n)$ of directed paths of length n from q to r satisfies a linear recurrence relation with (constant) integer coefficients.*

Proof Consider the adjacency matrix $\mathbf{M} \in \mathbb{N}^{V\times V}$ of G: for all vertices

$x, y \in V$, $\mathbf{M}_{x,y}$ is the number of arcs from x to y, *i.e.*, paths of length 1. A simple induction shows that, for all $x, y \in V$ and all $n \in \mathbb{N}$, $[\mathbf{M}^n]_{x,y}$ is the number of paths of length n from x to y. By the Cayley-Hamilton theorem, if $C(X) = \det(\mathbf{M} - X\mathbf{I}) = c_k X^k + \cdots + c_1 X + c_0 \in \mathbb{Z}[X]$ is the characteristic polynomial of $\mathbf{M}$ where $\mathbf{I}$ is the identity matrix of size $k = \mathrm{Card}(V)$, then $C(\mathbf{M}) = \mathbf{0}$. Multiplying by $\mathbf{M}^n$, $n \geq 0$, gives $c_k\mathbf{M}^{n+k} + \cdots + c_1\mathbf{M}^{n+1} + c_0\mathbf{M}^n = \mathbf{0}$. To conclude the proof, observe that this latter relation between matrices holds component-wise. □

The next result is a reformulation of Proposition 2.3.47.

Proposition 3.1.5 *Let $U = (U_n)_{n\geq 0}$ be a numeration system as given in Definition 3.1.3. If $\mathbb{N}$ is U-recognisable, then the sequence $(U_n)_{n\geq 0}$ satisfies a linear recurrence relation with (constant) integer coefficients.*

Proof Note that $\mathrm{rep}_U(U_\ell) = 10^\ell$ for all $\ell \geq 0$. Amongst the words of length $\ell + 1$ in $\mathrm{rep}_U(\mathbb{N})$, the smallest one for the genealogical ordering is 10^ℓ. Consequently, for all $\ell \geq 0$, $U_{\ell+1} - U_\ell$ is exactly the number of words of length $\ell + 1$ in $\mathrm{rep}_U(\mathbb{N})$. Since this latter language is regular, it is accepted by a DFA and the number of words of length n in $\mathrm{rep}_U(\mathbb{N})$ is equal to the number of paths of length n from the initial state to the final ones. Using Lemma 3.1.4 we deduce that the sequence $(\mathrm{Card}(\mathrm{rep}_U(\mathbb{N}) \cap A_U^n))_{n\geq 0}$ satisfies a linear recurrence relation with integer coefficients and the conclusion follows easily. □

As sketched by the next two examples, the converse of Proposition 3.1.5 does not hold in general. Sufficient conditions for $\mathbb{N}$ to be U-recognisable are considered in (Loraud 1995), (Hollander 1998). See Theorem 2.3.57. Also see Example 3.1.

Example 3.1.6 (Shallit 1994) Such a counterexample is given by the sequence $(U_n)_{n\geq 0}$ defined by $U_n = (n+1)^2$. Then we have $U_0 = 1$, $U_1 = 4$, $U_2 = 9$ and $U_{n+3} = 3\,U_{n+2} - 3\,U_{n+1} + U_n$. In that case, $\mathrm{rep}_U(\mathbb{N}) \cap 10^*10^* = \{10^a10^b \mid b^2 < 2a + 4\}$ showing with the pumping lemma that $\mathbb{N}$ is not U-recognisable.

Example 3.1.7 (Frougny 2002) We sketch another counterexample related to β-expansions, see Example 2.3.62 for details. Let $\beta = (3 + \sqrt{5})/2$. The β-expansion of 1 is 21^ω. Consider the sequence $(U_n)_{n\geq 0}$ satisfying the recurrence relation $U_{n+3} = 4\,U_{n+2} - 4\,U_{n+1} + U_n$, for all $n \geq 0$, with $U_0 = 1$, $U_1 = 2$ and $U_2 = 6$. Proceed by contradiction and assume that $\mathbb{N}$ is U-recognisable. Using the postponed Lemma 3.3.5, the set $X = \{U_n - 1 \mid$

$n \geq 0\}$ is U-recognisable because $\mathrm{rep}_U(X) = \mathsf{Maxlg}\,(\mathrm{rep}_U(\mathbb{N}))$. Due to the β-expansion of 1, one can show that all but a finite number of words in $\mathrm{rep}_U(X)$ are of the kind $21^{i_n}2w_n$ where $i_n \to \infty$ and $|w_n| \to \infty$ as $n \to \infty$. Therefore the pumping lemma shows that $\mathrm{rep}_U(X)$ is not regular.

It is probably worth to recall here a standard result about the general form of linear recurrence sequences, see any standard textbook like (Graham, Knuth, and Patashnik 1989). We assume that all the coefficients and the initial conditions belong to some field extension $\mathbb{K}$ of characteristic zero where the characteristic polynomial of the recurrence factorises as linear factors.

Theorem 3.1.8 *Let $k \geq 1$ and $r_0, \ldots, r_{k-1} \in \mathbb{K}$. Let $(U_n)_{n\geq 0}$ be a sequence satisfying, for all $n \geq 0$,*

$$U_{n+k} = r_{k-1}U_{n+k-1} + \cdots + r_0 U_n \ .$$

If $\alpha_1, \ldots, \alpha_t$ are the roots of the characteristic polynomial $X^k - r_{k-1}X^{k-1} - \cdots - r_0$ of the recurrence with respective multiplicities $m_1, \ldots, m_t$, then there exist polynomials $P_1, \ldots, P_t \in \mathbb{K}[X]$ of degree respectively less than $m_1, \ldots, m_t$ and depending only on the initial conditions $U_0, \ldots, U_{k-1} \in \mathbb{K}$ such that

$$\forall n \geq 0,\ U_n = P_1(n)\,\alpha_1^n + \cdots + P_t(n)\,\alpha_t^n \ .$$

Let $B \subset \mathbb{Z}$ be an alphabet. The function $\mathrm{val}_{B,U} : B^* \to \mathbb{Z}$ maps any word $w = c_\ell \cdots c_0 \in B^*$ onto $\mathrm{val}_{B,U}(w) = \sum_{k=0}^{\ell} c_k\, U_k$. It is clear that, for all $n \in \mathbb{N}$, $\mathrm{val}_{B,U}(\mathrm{rep}_U(n)) = n$. On the other hand, for all $w \in B^*$, such that $\mathrm{val}_{B,U}(w) \geq 0$, the so-called *normalisation* maps w onto $\mathrm{rep}_U(\mathrm{val}_{B,U}(w))$ which is not necessarily equal to w. Indeed, to apply $\mathrm{val}_{B,U}$, it is not required that w is a greedy U-representation. For example, considering the Fibonacci numeration system $F = (1, 2, 3, 5, \ldots)$, $\mathrm{rep}_F(\mathrm{val}_{\{0,1\},F}(11)) = 100$. Note that in general, if the alphabet B contains negative elements, then the normalisation is a partial function whose domain is a strict subset of B^*.

We already know that eventually periodic sets are k-recognisable for all $k \geq 2$. What can be said in a wider framework?

Proposition 3.1.9 *Let $p, r \geq 0$. If $(U_n)_{n\geq 0}$ is a numeration system given as in Definition 3.1.3 and satisfying a linear recurrence relation with integer coefficients, then*

$$\mathrm{val}^{-1}_{A_U,U}(p\,\mathbb{N} + r) = \left\{ c_\ell \cdots c_0 \in A_U^* \mid \sum_{k=0}^{\ell} c_k\, U_k \in p\,\mathbb{N} + r \right\}$$

is accepted by a DFA that can be effectively constructed. In particular, if $\mathbb{N}$ *is* U*-recognisable, then any eventually periodic set is* U*-recognisable.*

Prior to the proof, notice that for any integer $n \geq 0$, $\mathrm{val}_{A_U,U}^{-1}(n) \setminus 0A_U^*$ is a finite set of words $\{x_1, \ldots, x_{t_n}\}$ over A_U such that $\mathrm{val}_{A_U,U}(x_i) = n$ for all $i = 1, \ldots, t_n$. This non-empty set contains in particular $\mathrm{rep}_U(n)$.

Proof Since regular sets are stable under finite modifications, *i.e.*, adding and/or removing a finite number of words to a regular language gives a regular language, we can assume that $p > r \geq 0$. The sequence $(U_n \bmod p)_{n\geq 0}$ is eventually periodic say, with preperiod m and period q, that is, for all $i \geq m$, $U_i \equiv U_{i+q} \pmod{p}$. We build a deterministic finite automaton $\mathcal{A}$ accepting reversal of the words in $\{w \in A_U^* \mid \mathrm{val}_U(w) \in p\,\mathbb{N} + r\}$. The alphabet of the automaton is A_U. States are ordered pairs (t, s) where $0 \leq t < p$ and $0 \leq s < m + q$. The first component of a state handles the value modulo p of the digits that have been read and the second component takes care of the periodicity of $(U_n \bmod p)_{n\geq 0}$. The initial states is $(0, 0)$. Final states are the ones with the first component equal to r. Transitions are defined as follows

$$\forall s < m + q - 1 : \ (t, s) \xrightarrow{j} (jU_s + t \bmod p, \ s + 1)$$

and

$$(t, m + q - 1) \xrightarrow{j} (jU_{m+q-1} + t \bmod p, \ m)$$

for all $j \in A_U$. Note that $\mathcal{A}$ does not check the greediness of the accepted words, the construction only relies on the U-numerical value of the words modulo p. For the particular case, one has to consider the intersection of two regular languages $\mathrm{rep}_U(\mathbb{N}) \cap \mathrm{val}_U^{-1}(p\,\mathbb{N} + r)$. □

Taking into account this latter result, Cobham's theorem and also the above discussion about deciding whether a word is a valid U-representation or not, the recognisability of $\mathbb{N}$ is desirable and can be considered as a natural expectation for any numeration system. In particular in view of the above proposition, $\mathbb{N}$ is U-recognisable if, and only if, all eventually periodic sets are U-recognisable. If this becomes our basic requirement, we can *consider the problem the other way round*. Instead of taking a sequence U of integers and looking for conditions that guarantee the U-recognisability of $\mathbb{N}$, we take an arbitrary infinite regular language L over an alphabet A to build a numeration system, this language L being viewed as the set of valid representations of all the integers. Indeed a numeration system U is characterised by the language of all the representations and its monotonicity. In view of Proposition 3.1.2 about order-preserving representations,

if the alphabet A is totally ordered, say $(A, <)$, we order the words of L by the increasing genealogical order induced by the ordering of A, say $w_0 \prec w_1 \prec w_2 \prec \cdots$. This ordering of L gives a one-to-one correspondence between $\mathbb{N}$ and L: with $n \in \mathbb{N}$ is associated $w_n \in L$. Such a bijection is the essence of a numeration system: associating a representation with any integer. We have therefore the following formal definition.

Definition 3.1.10 An *abstract numeration system* (or *ANS* for short) is a triple $\mathcal{S} = (L, A, <)$ where L is an infinite regular language over a totally ordered alphabet $(A, <)$. The map $\mathrm{rep}_{\mathcal{S}} : \mathbb{N} \to L$ is the one-to-one correspondence mapping $n \in \mathbb{N}$ onto the $(n+1)$th word in the genealogically ordered language L, which is called the $\mathcal{S}$*-representation* of n. The $\mathcal{S}$-representation of 0 is the first word in L. The inverse map is denoted by $\mathrm{val}_{\mathcal{S}} : L \to \mathbb{N}$. If w is a word in L, $\mathrm{val}_{\mathcal{S}}(w)$ is its $\mathcal{S}$*-numerical value*.

Note that one could relax the assumption about the regularity of L in the definition of an ANS $\mathcal{S} = (L, A, <)$. In that case, we still have to consider words of L in ascending genealogical order. This would give a wider framework to work with, but then we loose the recognisability of $\mathbb{N}$.

Now let us present four examples. Some of them can be related to a suitably chosen sequence $(U_n)_{n \geq 0}$, others can not, showing that the class of ANS is strictly larger than the usual class of numeration systems given by Definition 3.1.1 and for which $\mathbb{N}$ is U-recognisable.

Example 3.1.11 Let U be a numeration system in the sense of Definition 3.1.1 such that $\mathbb{N}$ is U-recognisable. In view of Proposition 3.1.2, this numeration system can be considered as an ANS by enumerating the words of $\mathrm{rep}_U(\mathbb{N})$ by the genealogical order induced by the natural ordering of the digits. As an example, taking the language $1\{0, 01\}^* \cup \{\varepsilon\}$ with the natural ordering $0 < 1$ gives back the Fibonacci system and the language $B_k = \{0, \ldots, k-1\}^* \setminus 0\{0, \ldots, k-1\}^*$ gives the k-ary system.

Example 3.1.12 Consider $L = a^*b^*$ with $a < b$ and the ANS $\mathcal{S} = (L, \{a, b\}, <)$. The first few words in L in ascending genealogical order are

$$\varepsilon \prec a \prec b \prec aa \prec ab \prec bb \prec aaa \prec aab \prec abb \prec bbb \prec \cdots \ .$$

For example, $\mathrm{val}_{\mathcal{S}}(abb) = 8$ and $\mathrm{rep}_{\mathcal{S}}(3) = aa$. If we consider the bijection from L to $\mathbb{N}^2$ mapping the word $a^i b^j$ onto the pair (i, j), $i, j \geq 0$, it is not difficult to see that the genealogical ordering of L corresponds to the

primitive recursive Peano enumeration of $\mathbb{N}^2$, that is

$$\mathrm{val}_{\mathcal{S}}(a^i b^j) = \frac{1}{2}(i+j)(i+j+1) + j. \tag{3.2}$$

Let us pursue this example a little bit further. Assume that we have a map $v : \{a, b\} \to \mathbb{N}$ which assigns some weight to a and b. We show that there exists *no* sequence $U = (U_n)_{n\geq 0}$ defining a numeration system in the sense of Definition 3.1.1 such that, for all words $w_\ell \cdots w_0 \in L$,

$$\mathrm{val}_{\mathcal{S}}(w_\ell \cdots w_0) = \sum_{k=0}^{\ell} v(w_k)\, U_k \ .$$

We proceed by contradiction and we assume that such a sequence exists. Since $U_0 = 1$ and $\mathrm{val}_{\mathcal{S}}(a) = 1$, $\mathrm{val}_{\mathcal{S}}(b) = 2$, we must have $v(a) = 1$ and $v(b) = 2$. Notice that $\mathrm{val}_{\mathcal{S}}(aa) = 3$ and this quantity should be equal to $v(a)U_1 + v(a)U_0$. Consequently, $U_1 = 2$. Therefore $v(b)U_1 + v(b)U_0 = 6$ but $\mathrm{val}_{\mathcal{S}}(bb) = 5$, which gives a contradiction.

This example shows that the family of ANS contains more numeration systems that those of Definition 3.1.1 for which $\mathbb{N}$ is U-recognisable. To contrast with ANS which only depend on the genealogical ordering, recall that the systems associated with Definition 3.1.1 are referred as positional numeration systems. As we shall soon see in Lemma 3.2.2, the general expression of $\mathrm{val}_{\mathcal{S}}(w)$ for an ANS $\mathcal{S} = (L, A, <)$ and a word $w \in L$ involves usually more than a single linear recurrence sequence.

Example 3.1.13 (Allowing leading zeroes) The reader may have noticed that we have defined greedy U-representations as words not starting with zero. It not only makes the definition unambiguous but this choice was made on purpose because in the context of abstract numeration systems, adding leading zeroes to a word changes its length and therefore its position in the genealogically ordered language. As an example, consider the language $\{0,1\}^*$. The first few words in this language are ε, 0, 1, 00, 01, 10, 11, 000. So for the ANS $\mathcal{S} = (\{0,1\}^*, \{0,1\}, 0 < 1)$, we get $\mathrm{val}_{\mathcal{S}}(0) = 1$, $\mathrm{val}_{\mathcal{S}}(00) = 3$ and so on. Actually, if one considers the map v defined as $v(0) = 1$ and $v(1) = 2$, it is not difficult to see that $\mathrm{val}_{\mathcal{S}}(w_\ell \cdots w_0) = \sum_{k=0}^{\ell} v(w_k)\, 2^k$ which corresponds to the so-called 2-*adic numeration system*: any non-negative integer is uniquely represented as a word over $\{1, 2\}$ with the sequence $(2^n)_{n\geq 0}$ being the underlying scale.

Example 3.1.14 (Pisot numeration system) Recall from Chapter 2 that a *Pisot number* is an algebraic integer $\alpha > 1$ whose conjugates have modulus less than 1. Consider a linear recurrence sequence $(U_n)_{n\geq 0}$ whose

characteristic polynomial is the minimal polynomial of a Pisot number α of degree k. If the integer initial conditions are $1 = U_0 < U_1 < \cdots < U_{k-1}$, then there exists some $c > 0$ such that $U_n \sim c\,\alpha^n$ and moreover $|U_n - c\,\alpha^n| \to 0$, as n tends to infinity, because we can apply Theorem 3.1.8 about the general solution of a linear recurrence† and for any other root $\beta \neq \alpha$ of the characteristic polynomial of the recurrence, since $|\beta| < 1$, we have $\beta^n \to 0$ as $n \to \infty$. This sequence can be used to define a numeration system in the sense of Definition 3.1.1. It is well-known that for such a system, $\mathbb{N}$ is U-recognisable. Moreover, all the nice properties of the integer base numeration systems still hold: logical or substitutive characterisations of the U-recognisable sets, stability of U-recognisability under addition and multiplication by a constant, normalisation is computable by finite automata,... see (Bruyère and Hansel 1997), (Frougny 1992). Since $\mathbb{N}$ is U-recognisable, these "state-of-the-art" positional numeration systems are all special cases of ANS.

Example 3.1.15 (Prefix-closed language) In the case of an ANS based on a prefix-closed language, we propose a useful picture of the map $\mathrm{val}_{\mathcal{S}}$. This is simply another expression of the genealogical ordering of L. As an example consider the language $L = \{a, ba\}^*\{\varepsilon, b\}$ and $a < b$. In Figure 3.1 we represent the first three levels of the corresponding *trie*, *i.e.*, a rooted tree where the edges are labelled by letters from A, and the nodes are labelled by prefixes of words in the considered language L. Let $u \in A^*$, $a \in A$. If ua is (a prefix of) a word in L, then there is an edge between u and ua. Note that for a prefix-closed language L, all prefixes of words in L belong to L. In the nodes, we have written the $\mathcal{S}$-numerical value of the corresponding words in L. The root is associated with ε. When considering a prefix-closed

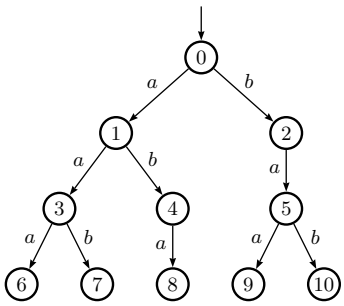

Fig. 3.1. A trie for words of length ≤ 3 in L.

† Remember that all the roots of the minimal polynomial of an algebraic number are simple.

language ordered by genealogical order, the nth level of the trie contains all words of L of length n in lexicographic order from left to right assuming that the sons of a node are also ordered with respect to the ordering of the alphabet.

Definition 3.1.16 For a given ANS $\mathcal{S} = (L, A, <)$, any integer n is mapped onto a word $\mathrm{rep}_{\mathcal{S}}(n)$ and any subset X of $\mathbb{N}$ is mapped onto a language $\mathrm{rep}_{\mathcal{S}}(X) \subseteq L$. We have therefore a one-to-one correspondence between $2^{\mathbb{N}}$ and 2^L. In this general framework of abstract numeration systems, we are interested in sets X of integers such that $\mathrm{rep}_{\mathcal{S}}(X)$ is regular. These sets are called *$\mathcal{S}$-recognisable sets*.

Example 3.1.17 Considering the ANS $\mathcal{S} = (a^*b^*, \{a, b\}, a < b)$ from Example 3.1.12, the set X of triangular numbers

$$X = \{0, 1, 3, 6, 10, \ldots\} = \{n(n+1)/2 \mid n \geq 0\}$$

is $\mathcal{S}$-recognisable. Indeed, it is easy to check that $\mathrm{rep}_{\mathcal{S}}(X) = a^*$ because the number of words of length $n \geq 0$ in a^*b^* is exactly $n+1$. If we consider the ANS $\mathcal{R} = (a^*b^*, \{a, b\}, b < a)$ where the ordering of the alphabet has been reversed, then $\mathrm{rep}_{\mathcal{R}}(X) = b^*$.

3.2 Computing numerical values and $\mathcal{S}$-representations

Let $\mathcal{S} = (L, A, <)$ be an abstract numeration system. Since L is a regular language, we can consider a complete DFA $\mathcal{A} = (Q, A, E, \{q_0\}, T)$ having $\delta_{\mathcal{A}} : Q \times A^* \to Q$ as (extended) transition function. We write $q.w$ as a shorthand for $\delta_{\mathcal{A}}(q, w)$ if the context is clear, $q \in Q$, $w \in A^*$. First we show, as a consequence of the genealogical ordering of L, that the function $\mathrm{val}_{\mathcal{S}}$ can be computed recursively and we obtain a decomposition of any integer using functions $\mathcal{U}$ and $\mathcal{V}$ counting the number of words accepted from the different states of $\mathcal{A}$ and defined below.

For all $q \in Q$, $L_q = \{w \in A^* \mid q.w \in F\}$ is the regular language of words accepted in $\mathcal{A}$ starting from state q. We set

$$\mathcal{U}_q(n) := \mathrm{Card}(L_q \cap A^n) \quad \text{and} \quad \mathcal{V}_q(n) := \sum_{k=0}^{n} \mathcal{U}_q(k) \tag{3.3}$$

being respectively the number of words of length n and, at most n, accepted from q. From Lemma 3.1.4, all the sequences $(\mathcal{U}_q(n))_{n\geq 0}$, $q \in Q$, satisfy the same linear recurrence relation. Indeed, $\mathcal{U}_q(n)$ is the sum over all the final states $f \in T$ of the number of paths of length n from q to f. Moreover, $(\mathcal{V}_q(n))_{n\geq 0}$ satisfies a linear recurrence relation that can be derived from

the one satisfied by $(\mathcal{U}_q(n))_{n\geq 0}$, simply by observing that, for all $n \geq 0$, we have $\mathcal{V}_q(n+1) - \mathcal{V}_q(n) = \mathcal{U}_q(n+1)$. Also we write

$$\mathcal{U}(n) := \mathcal{U}_{q_0}(n) = \mathrm{Card}(L \cap A^n) \quad \text{and} \quad \mathcal{V}(n) := \mathcal{V}_{q_0}(n) = \mathrm{Card}(L \cap A^{\leq n}) \ .$$

Note that these two maps $\mathcal{U}(n)$ and $\mathcal{V}(n)$ are independent of the choice of the DFA accepting L. They only depend on the language L, so if emphasis on L is needed, we also use notation like $\mathcal{U}_L(n)$ and $\mathcal{V}_L(n)$. The map $\mathcal{U} : \mathbb{N} \to \mathbb{N}$ is often called the *counting function* or (*combinatorial*) *complexity function* of L (compare with Definition 1.2.12).

Since, for all $q \in Q$, the language L_q is regular, we can consider the ANS $\mathcal{S}_q = (L_q, A, <)$. The corresponding maps $\mathrm{val}_{\mathcal{S}_q}$ and $\mathrm{rep}_{\mathcal{S}_q}$ are respectively denoted by val_q and rep_q. For some $q \in Q$, L_q can possibly be finite. If this is the case, we extend the definition of an ANS to allow this situation but the domain of rep_q is therefore $\{0, \ldots, \mathrm{Card}\, L_q - 1\}$.

Example 3.2.1 Consider the regular language L accepted by the DFA depicted in Figure 3.2 having states q_0, q_1 and q_2. With notation introduced

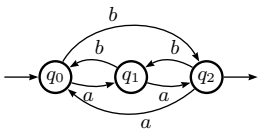

Fig. 3.2. A DFA accepting the language $L \subset \{a, b\}^*$.

above, the first few words in $L = L_{q_0}$, L_{q_1} and L_{q_2} are respectively

$$\begin{aligned} L_{q_0} &= \{b, aa, abb, bab, bba, aaab, aaba, abaa, baaa, bbbb, aaaaa, \ldots\} \\ L_{q_1} &= \{a, bb, aab, aba, baa, aaaa, abbb, babb, bbab, bbba, aaabb, \ldots\} \\ L_{q_2} &= \{\varepsilon, ab, ba, aaa, bbb, aabb, abab, abba, baab, baba, bbaa, aaaab, \ldots\} \end{aligned}$$

and the adjacency matrix of the automaton is

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

Therefore, using the same technique as in the proof of Lemma 3.1.4 (simply compute the characteristic polynomial of $\mathbf{M}$), we get that the sequences $(\mathcal{U}_{q_i}(n))_{n\geq 0}$ satisfy $\mathcal{U}_{q_i}(n+3) = 3\mathcal{U}_{q_i}(n+1) + 2\mathcal{U}_{q_i}(n)$ for all $n \geq 0$. We have computed the first few values of these sequences:

	0	1	2	3	4	5	6	7	8	9	10
$\mathcal{U}_{q_0}(n) = \mathcal{U}_{q_1}(n)$	0	1	1	3	5	11	21	43	85	171	341
$\mathcal{U}_{q_2}(n)$	1	0	2	2	6	10	22	42	86	170	342

.

For instance, $\mathrm{rep}_{\mathcal{S}}(0) = \mathrm{rep}_{q_0}(0) = b$, $\mathrm{rep}_{q_1}(0) = a$ and $\mathrm{rep}_{q_2}(0) = \varepsilon$. In the same way, $\mathrm{val}_{\mathcal{S}}(abb) = \mathrm{val}_{q_0}(abb) = 2$, $\mathrm{val}_{q_1}(aab) = 2 = \mathrm{val}_{q_2}(ba)$.

Now that we have a good knowledge of the different maps val_q, $\mathcal{U}_q$ and $\mathcal{V}_q$, we present a lemma used to compute recursively the $\mathcal{S}$-numerical value of any word in L.

Lemma 3.2.2 *Let $\mathcal{S} = (L, A, <)$ be an ANS where L is accepted by a DFA $\mathcal{A} = (Q, A, E, \{q_0\}, T)$. Let $q \in Q$. If the word xy belongs to L_q where the factor y is non-empty, then*

$$\mathrm{val}_q(xy) = \mathrm{val}_{q.x}(y) + \mathcal{V}_q(|xy| - 1) - \mathcal{V}_{q.x}(|y| - 1) + \sum_{\substack{w<x \\ |w|=|x|}} \mathcal{U}_{q.w}(|y|) \ .$$

Proof We have to compute the number of words belonging to L_q and genealogically less than xy. There are three kinds of such words. The first ones are the words of length less than $|xy|$. We have $\mathcal{V}_q(|xy| - 1)$ such words. Then we have to take into account words in L_q of length $|xy|$ having a prefix w such that $|w| = |x|$ and $w < x$. It is clear that there are

$$\mathrm{Card}\{wz \in L_q \mid w < x,\ |w| = |x|,\ |z| = |y|\} = \sum_{\substack{w<x \\ |w|=|x|}} \mathcal{U}_{q.w}(|y|)$$

words of this kind. Finally, we have words in L_q of length $|xy|$ having x as prefix and lexicographically less than xy. We have to count the number of words in $L_{q.x}$ of length $|y|$ lexicographically less than y. We get $\mathrm{val}_{q.x}(y) - \mathcal{V}_{q.x}(|y| - 1)$ such words because $\mathrm{val}_{q.x}(y)$ is the total number of words less than y in $L_{q.x}$ and we have to subtract words of length less than $|y|$. □

For ANS we have a "multi-scale" analogue to the decomposition (3.1) occurring in positional numeration systems. Let $\mathcal{S} = (L, A, <)$ be an ANS where L is accepted by a DFA $\mathcal{A}$. Instead of having a unique sequence $(U_n)_{n\geq 0}$ to express the numerical value of a word $c_\ell \cdots c_0$ as $\sum_{k=0}^{\ell} c_k U_k$, we are considering the several sequences $(\mathcal{U}_q(n))_{n\geq 0}$, in fact, as many sequences as states in $\mathcal{A}$.

Theorem 3.2.3 *Let $\mathcal{S} = (L, A, <)$ be an ANS where L is accepted by the DFA $\mathcal{A} = (Q, A, E, \{q_0\}, T)$. Let $w = w_1 \cdots w_n \in L$. Then we have*

$$\mathrm{val}_{\mathcal{S}}(w) = \sum_{q\in Q} \sum_{i=1}^{|w|} b_{q,i}(w)\ \mathcal{U}_q(|w| - i) \tag{3.4}$$

where for $i = 1, \ldots, |w|$,

$$b_{q,i}(w) = \mathrm{Card}\{a \in A \mid a < w_i,\ q_0.w_1 \cdots w_{i-1}a = q\} + \mathbf{I}_{q,q_0} \tag{3.5}$$

where $\mathbf{I}$ *is the identity matrix in* $\{0,1\}^{Q\times Q}$*, so* $\mathbf{I}_{q,q_0} = 1$ *if, and only if,* $q = q_0$*. Moreover, these coefficients are bounded:*

$$0 \leq \sum_{q\in Q} b_{q,i}(w) \leq \mathrm{Card}\, A\ .$$

Proof Formula (3.4) can be proved using Lemma 3.2.2 inductively. Also it can be proved by observing that the summand for $(q,i) \in Q \times \{1,\ldots,|w|\}$ with $q \neq q_0$ is the number of words $v = v_1 \cdots v_n$ of length $|w|$ which have prefix $w_1 \cdots w_{i-1}a$ with $a < w_i$, which means that $v \prec w$, the state q is reached after reading the first i letters of v, and the suffix $v_{i+1} \cdots v_n$ is accepted from state q. For $q = q_0$ the summand for (q_0, i) equals the number with the same descriptions as above plus the number of words of length $|w| - i$ which are accepted by the automaton starting from q_0. Summing over all possible pairs (q,i) first gives the number of words $v \prec w$ with $|v| = |w|$, the extra summand for $q = q_0$ equals the number of words v in L with $|v| < |w|$. Altogether this equals $\mathrm{val}_{\mathcal{S}}(w)$. □

The following proposition asserts that the coefficients of the decomposition (3.4) can be obtained almost *automatically*. This is merely the translation of (3.5) but this fact will play an important role when dealing with the representation of real numbers in Section 3.5.

Proposition 3.2.4 *Let* $\mathcal{S} = (L, A, <)$ *be an ANS where* L *is accepted by a DFA* $\mathcal{A} = (Q, A, E, \{q_0\}, T)$*. For any* $q \in Q$*, one can efficiently build a sequential letter-to-letter transducer* $\mathcal{T}_q$ *computing, for all* i*, the coefficients* $b_{q,i}(w)$ *occurring in* (3.4)*,* i.e.*, to any input* $w_1 \cdots w_n \in L$ *is associated the output* $b_{q,1}(w) \cdots b_{q,n}(w)$ *of* $\mathcal{T}_q$*.*

Proof It is a direct consequence of (3.5). Let $q \in Q$. We build a transducer $\mathcal{T}_q$ having the same set of states and the same initial state and final states as $\mathcal{A}$. The input and output alphabets of $\mathcal{T}_q$ are respectively A and $\{0,\ldots,\mathrm{Card}\, A\}$. For any transition $(r,c,s) \in E$ appearing in $\mathcal{A}$, we take for the transducer $\mathcal{T}_q$ the transition $(r,(c,x_{r,c}),s)$ where $x_{r,c} = \mathrm{Card}\{a \in A \mid a < c, r.a = q\} + \mathbf{I}_{q_0,q}$. □

Example 3.2.5 Consider the ANS from Example 3.2.1. The corresponding three transducers are given in Figure 3.3 (one for each state of the DFA). Notice that the output associated with a is always 0 except for q_0 where it

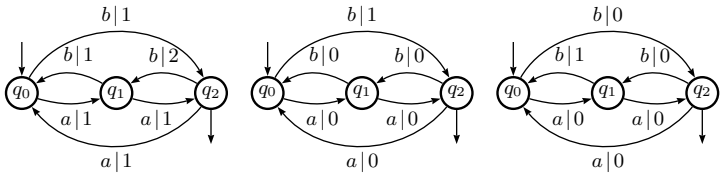

Fig. 3.3. The three transducers $\mathcal{T}_{q_0}$, $\mathcal{T}_{q_1}$ and $\mathcal{T}_{q_2}$.

is 1. Consider for instance the word $abaa$ which is such that $\mathrm{val}_{\mathcal{S}}(abaa) = 7$. Feeding the transducers with this word gives respectively the words 1111, 0000 and 0100. Therefore, we find the expected decomposition of 7:

$$\begin{array}{llllll} & \mathbf{1}\mathcal{U}_{q_0}(3) & +\mathbf{1}\mathcal{U}_{q_0}(2) & +\mathbf{1}\mathcal{U}_{q_0}(1) & +\mathbf{1}\mathcal{U}_{q_0}(0) & \\ + & \mathbf{0}\mathcal{U}_{q_1}(3) & +\mathbf{0}\mathcal{U}_{q_1}(2) & +\mathbf{0}\mathcal{U}_{q_1}(1) & +\mathbf{0}\mathcal{U}_{q_1}(0) & \\ + & \mathbf{0}\mathcal{U}_{q_2}(3) & +\mathbf{1}\mathcal{U}_{q_2}(2) & +\mathbf{0}\mathcal{U}_{q_2}(1) & +\mathbf{0}\mathcal{U}_{q_2}(0) & = 3+1+1+0+2. \end{array}$$

Now let us turn our attention to the computation of $\mathrm{rep}_{\mathcal{S}}(n)$ where $\mathcal{S} = (L, A, <)$ and $A = \{a_1 < \cdots < a_t\}$. We assume that we have at our disposal a DFA $\mathcal{M}$ having q_0 as initial state and accepting L. In particular, $\mathcal{U}_q(n)$ and $\mathcal{V}_q(n)$ can be obtained using the linear recurrence relations derived from $\mathcal{M}$ and its adjacency matrix. As usual, we simply write $q.w$ for the action of w in A^* on q in the set of states of $\mathcal{M}$.

Observe that $|\mathrm{rep}_{\mathcal{S}}(n)| = \ell > 0$ if, and only if, $\mathcal{V}(\ell-1) \leq n < \mathcal{V}(\ell)$. Indeed, if L contains some words of length ℓ, then the first word of length ℓ has position $\mathcal{V}(\ell-1)$ in the genealogically ordered language L and $\mathcal{U}(\ell) > 0$. So, for all $n \geq 0$, we get

$$|\mathrm{rep}_{\mathcal{S}}(n)| = \inf\{m \in \mathbb{N} \mid n < \mathcal{V}(m)\} \ .$$

Let $n \geq 0$ and $\ell = |\mathrm{rep}_{\mathcal{S}}(n)|$. To determine the first letter of the $\mathcal{S}$-representation of n, we compute, for all $s \in \{1, \ldots, t\}$, the number $N[\ell, a_s]$ of words of length ℓ belonging to L and beginning with a_1, a_2, ... or a_s. It is given by $N[\ell, a_s] := \sum_{i=1}^{s} \mathcal{U}_{q_0.a_i}(\ell-1)$. For convenience, we set $N[\ell, a_0] = 0$. There exists a unique r such that $N[\ell, a_{r-1}] \leq n - \mathcal{V}(\ell-1) < N[\ell, a_r]$ and the first letter of the $\mathcal{S}$-representation of n is therefore a_r. We proceed in the same way to determine the other letters of the $\mathcal{S}$-representation. Table 3.1 sketches the structure of the genealogically ordered language L for words of length ℓ with their corresponding position in L. The pseudocode algorithm presented in Table 3.2 computes the $\mathcal{S}$-representation w of n. In the last line of this algorithm, wa_j represents the concatenation of the word w and the letter a_j.

$\mathcal{V}_{\ell-1}$	a_1	a_1	$\cdots$
	$\vdots$	$\vdots$	
$\mathcal{V}_{\ell-1}+\mathcal{U}_{q_0.a_1a_1}(\ell-2)$		a_2	$\cdots$
		$\vdots$	
	$\vdots$		
	a_1	a_p	$\cdots$
$\mathcal{V}_{\ell-1}+\mathcal{U}_{q_0.a_1}(\ell-1)$	a_2	a_1	$\cdots$
	$\vdots$	$\vdots$	
$\mathcal{V}_{\ell-1}+\mathcal{U}_{q_0.a_1}(\ell-1)+\mathcal{U}_{q_0.a_2a_1}(\ell-2)$		a_2	$\cdots$
		$\vdots$	
	$\vdots$		
	a_2	a_p	$\cdots$
	$\vdots$		
$\mathcal{V}_{\ell-1}+\sum_{i=1}^{p-1}\mathcal{U}_{q_0.a_i}(\ell-1)$	a_p	a_1	$\cdots$
	$\vdots$	$\vdots$	
$\mathcal{V}_{\ell-1}+\sum_{i=1}^{p-1}\mathcal{U}_{q_0.a_i}(\ell-1)+\mathcal{U}_{q_0.a_pa_1}(\ell-2)$		a_2	$\cdots$
		$\vdots$	
	$\vdots$		
$\mathcal{V}_{\ell}-1$	a_p	a_p	$\cdots$

Table 3.1. *Words in L of length ℓ in increasing genealogical order and their corresponding $\mathcal{S}$-numerical values.*

3.3 $\mathcal{S}$-recognisable sets

The aim of this section is to present some properties of $\mathcal{S}$-recognisable sets of integers. We know that eventually periodic sets are k-recognisable, for all $k \geq 2$, and by Proposition 3.1.9 also U-recognisable for numeration systems such that $\mathbb{N}$ is U-recognisable. Interestingly this property† still holds for ANS which is somehow encouraging if one thinks about a possible analogue of the Cobham theorem.

Theorem 3.3.1 *Let $\mathcal{S} = (L, A, <)$ be an ANS. Any eventually periodic set is $\mathcal{S}$-recognisable.*

Due to the importance of this result, we provide two different proofs. The first one is direct: we show that the minimal automaton of the set

† It was the very first result we were looking for. Getting it was a true motivation for the study of ANS.

```
Find the unique ℓ be such that V(ℓ − 1) ≤ n < V(ℓ)
q ← q_0
m ← n − V(ℓ − 1)
w ← ε
FOR i = 1 TO ℓ DO
   s ← 1
   WHILE m ≥ U_{q.a_s}(ℓ − i) DO
     m ← m − U_{q.a_s}(ℓ − i)
     s ← s + 1
   END-WHILE
   q ← q.a_s
   w ← w a_s
END-FOR
```

Table 3.2. *An algorithm for computing* $\mathrm{rep}_{\mathcal{S}}(n)$.

of representations of any eventually periodic set is finite. It presents some sharp argument but it does not provide any "constructive feeling" about the machinery behind as does the second proof.

Prior to these proofs we can make the following observation. It is well-known that taking in a regular language the smallest (respectively largest) word of every length for the genealogical ordering gives again a regular language, see Proposition 2.6.4 and Lemma 3.3.5. We can reformulate Theorem 3.3.1 to obtain some *decimation* operation preserving the regularity of languages.

Theorem 3.3.2 *Let* $(A, <)$ *be a totally ordered alphabet. If we order the words of a regular language* $L \subseteq A^*$ *in the genealogical order induced by* $<$*, say* $w_0 \prec w_1 \prec w_2 \prec \cdots$*, then for all* $p > r \geq 0$ *the language* $\{w_{np+r} \in L \mid n \geq 0\}$ *is regular.*

Let us present a first proof of Theorem 3.3.1 or equivalently of the above theorem.

Proof It is well-known that a language $M \subseteq A^*$ is regular if, and only if, its minimal automaton $\mathcal{A}_M$ is finite. The set of states of $\mathcal{A}_M$ is $\{w^{-1}M \mid w \in A^*\}$ where $w^{-1}M = \{u \mid wu \in M\}$. See any standard textbook about automata theory like (Eilenberg 1974) or (Sakarovitch 2003).

Since a finite union of regular languages is regular and since adding or removing a finite number of words in a regular language does not change its regularity, it is enough to show that the minimal automaton $\mathcal{A}_P$ of the language $P = \mathrm{rep}_{\mathcal{S}}(p\mathbb{N} + r) \subseteq A^*$ is finite, with $p > r \geq 0$. The states of $\mathcal{A}_P$ are the sets

$$w^{-1}P = \{x \in A^* \mid \mathrm{val}_{\mathcal{S}}(wx) \equiv r \pmod p\}, w \in A^* \ .$$

Consider the regular language L on which the ANS $\mathcal{S}$ is built and its corresponding minimal automaton $\mathcal{A}_L$. In fact we could consider any DFA accepting L, the arguments remain unchanged. The reader should be careful, we are considering two different minimal automata: $\mathcal{A}_L$ which we know is finite and $\mathcal{A}_P$ which we would like to prove to be finite. We know that, for all states q of $\mathcal{A}_L$, the sequences $(\mathcal{U}_q(n))_{n\geq 0}$ and $(\mathcal{V}_q(n))_{n\geq 0}$ introduced in (3.3) satisfy a linear recurrence equation and are therefore eventually periodic mod p. Let q_0 be the initial state of $\mathcal{A}_L$. Assume that the period of the sequence $(\mathcal{V}_{q_0}(n) \bmod p)_{n\geq 0}$ is t and its preperiod is s. By Lemma 3.2.2, we have

$$\mathrm{val}_{\mathcal{S}}(wx) = \mathrm{val}_{q_0.w}(x) + \mathcal{V}_{q_0}(|wx|-1) - \mathcal{V}_{q_0.w}(|x|-1) + \sum_{\substack{v<w \\ |v|=|w|}} \mathcal{U}_{q_0.v}(|x|).$$

Since $\mathcal{A}_L$ is finite, $q_0.w$ can only take a finite number of values in Q, the set of states of $\mathcal{A}_L$. Working modulo p, for $|w| > s$, the term $\mathcal{V}_{q_0}(|wx|-1)$ can be written as $\mathcal{V}_{q_0}(|x|+i)$ for some $i \in \{0,\dots,t-1\}$ because $(\mathcal{V}_{q_0}(n) \bmod p)_{n\geq 0}$ is eventually periodic. Modulo p, for all $w \in A^*$, there exist coefficients $j_q \in \{0,\dots,p-1\}$ such that

$$\sum_{\substack{v<w \\ |w|=|v|}} \mathcal{U}_{q_0.v}(|x|) \equiv \sum_{q\in Q} j_q\,\mathcal{U}_q(|x|) \pmod p\ .$$

Note that the number of maps $n \mapsto \sum_{q\in Q} j_q \mathcal{U}_q(n)$ is finite and bounded by $p^{\mathrm{Card}\,Q}$. Consequently, for any $w \in A^*$ such that $|w| > s$, the set $w^{-1}P$ is of the form

$$\Big\{x \mid \mathrm{val}_k(x) + \mathcal{V}_{q_0}(|x|+i) - \mathcal{V}_k(|x|-1) + \sum_{q\in Q} j_q\,\mathcal{U}_q(|x|) \equiv r \pmod p\Big\}$$

for some $k \in Q$, $j_q \in \{0,\dots,p-1\}$ and $i \in \{0,\dots,t-1\}$. So, there are finitely many sets of this kind and the set $\{w^{-1}P \mid w \in A^*\}$ of states of the minimal automaton of P is finite. □

Now let us consider an alternative proof followed by an example.

Idea of the proof. We notice that all the sequences occurring in Lemma 3.2.2 are eventually periodic modulo p with some common period M and they are all periodic after at most K terms. We build an NFA which reads entries from the left, say leading letter first, and which computes the numerical value of entries modulo p. In order to apply Lemma 3.2.2, we have to keep track of the state the DFA accepting L is in when reading such an entry. Also we have to deal with the common period M: when we enter a new word w, the value of $|w| \bmod M$ is guessed non-deterministically.

Only a correct guess can lead to the unique final state. The last K letters are treated separately because we cannot rely anymore on the periodic structure.

Proof Let $\mathcal{A} = (Q, A, E, \{q_0\}, T)$ be a DFA accepting the language L with $\delta_{\mathcal{A}} : Q \times A \to Q$ as the transition function. We give a method to construct an NFA accepting $\mathrm{rep}_{\mathcal{S}}(p\,\mathbb{N} + r)$. The key argument is again that, for all $q \in Q$, the sequences $(\mathcal{U}_q(n) \bmod p)_{n\geq 0}$ and $(\mathcal{V}_q(n) \bmod p)_{n\geq 0}$ are eventually periodic. Therefore, for each $q \in Q$, there exist g_q, h_q, s_q and t_q belonging to $\mathbb{N}$ such that $h_q, t_q \geq 1$,

$$\forall n \geq g_q,\ \mathcal{U}_n(q) \equiv \mathcal{U}_{n+h_q}(q) \pmod{p}$$

and

$$\forall n \geq s_q,\ \mathcal{V}_n(q) \equiv \mathcal{V}_{n+t_q}(q) \pmod{p}\ .$$

Set M to be the least common multiple of the constants h_q and t_q and

$$K = \max\left\{\sup_{q\in Q} g_q, \sup_{q\in Q} s_q + 1\right\}\ .$$

Taking $s_q + 1$ instead of s_q is due to the term $\mathcal{V}_{q.a}(|y|-1)$ in the expression of $\mathrm{val}_q(ay)$ given by Lemma 3.2.2: for all $a \in A$ and all $y \in A^+$ such that $ay \in L_q$, we have

$$\mathrm{val}_q(ay) = \mathrm{val}_{q.a}(y) + \overbrace{\mathcal{V}_q(|y|) - \mathcal{V}_{q.a}(|y|-1) + \sum_{\substack{b<a\\ b\in A}} \mathcal{U}_{q.b}(|y|)}^{=:R(q,a,|y|)}\ . \tag{3.6}$$

This shows that for $|y| \geq K$, $\mathrm{val}_q(ay)$ is congruent to $\mathrm{val}_{q.a}(y)$ modulo p but a quantity $R(q, a, |y|) \bmod p$ depending only on q, a and $|y| \bmod M$ has to be added. Hence, for $n \geq 1$ and letters $a_1, \ldots, a_{K+n} \in A$, we obtain inductively that $\mathrm{val}_q(a_1 \cdots a_n a_{n+1} \cdots a_{n+K})$ is equal to

$$\begin{aligned} & R(q, a_1, K+n-1) + R(q.a_1, a_2, K+n-2) + \cdots \\ + \quad & R(q.a_1 \cdots a_{n-1}, a_n, K) + \mathrm{val}_{q.a_1\cdots a_n}(a_{n+1} \cdots a_{n+K})\ . \end{aligned}$$

We will mimic this latter decomposition using the following NFA. Consider the NFA $\mathcal{B} = (Q' \cup \{f\}, A, E', I, F)$ where $Q' = Q \times \{0, \ldots, p-1\} \times \{0, \ldots, M-1\}$, $I = \{(q_0, 0, j) \mid j = 0, \ldots, M-1\}$ and $F = \{f\}$ where $f \notin Q'$. Let us show that this NFA accepts the language $\mathrm{rep}_{\mathcal{S}}(p\,\mathbb{N} + r) \cap A^{\geq K}$. The language $\mathrm{rep}_{\mathcal{S}}(p\,\mathbb{N} + r) \cap A^{<K}$ is finite and can be handled separately. The first component of any state of $\mathcal{B}$ is used to store and mimic

the behaviour of $\mathcal{A}$. The estimated numerical value modulo p resulting from the letters that have already been read, is stored in the second component (starting from zero, first we add $R(q, a_1, K+n-1)$, then $R(q.a_1, a_2, K+n-2)$, *etc.*). The length modulo M of the remaining part of the word to be read is stored in the last component of the state, this length is unknown at the beginning and will non-deterministically be guessed by $\mathcal{B}$. Now we will explain the details. The transition relation of $\mathcal{B}$ is such that

$$\begin{array}{ll} ((q,i,j),a,(\delta_{\mathcal{A}}(q,a),k,j-1)) \in E', & \text{if } j \in \{1,\ldots,M-1\}; \\ ((q,i,j),a,(\delta_{\mathcal{A}}(q,a),k,M-1)) \in E', & \text{if } j = 0 \end{array} \tag{3.7}$$

where the unique k, depending on q, a, i and j, is easily computed using (3.6). Actually $k = i+R(q,a,j) \bmod p$. If $x \in L_q \cap A^K$ and $i \in \{0,\ldots,p-1\}$ are such that $\mathrm{val}_q(x) + i \equiv r \pmod p$ then we also add

$$((q, i, K \bmod M), x, f) \in E'$$

and note that these are the only relations leading to the final state. The reading of a word w of length at least K could *a priori* be started from any of the M initial states of $\mathcal{B}$. But note that only one of these states has to be chosen with respect to $|w|$ to reach the unique final state f at the end of the reading of w. □

Example 3.3.3 We apply the above construction to obtain an NFA recognising $\mathrm{rep}_{\mathcal{S}}(3\,\mathbb{N}+1)$ where $\mathcal{S}$ is the ANS based on the language L of the words over $\{a,b\}$ having an even number of b. We assume that $a < b$. The minimal automaton of L is depicted in Figure 3.4. We have

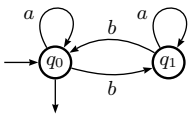

Fig. 3.4. DFA accepting words with an even number of b.

$$\begin{cases} \mathcal{U}_{q_0}(n) = 2^{n-1},\ \forall n \geq 1 \\ \mathcal{U}_{q_0}(0) = 1 \end{cases} \quad \text{and} \quad \begin{cases} \mathcal{U}_{q_1}(n) = 2^{n-1},\ \forall n \geq 1 \\ \mathcal{U}_{q_1}(0) = 0\ . \end{cases}$$

For all $n \geq 1$, $\mathcal{U}_{q_0}(n) = \mathcal{U}_{q_1}(n) \equiv (-1)^{n-1} \pmod 3$ and, for all $n \in \mathbb{N}$, $\mathcal{V}_{q_0}(n) \equiv (-1)^n \pmod 3$ and $\mathcal{V}_{q_1}(n) \equiv (-1)^n - 1 \pmod 3$. Using the notation given in the previous proof, we set $K = 1$ and $M = 2$. From (3.6),

we get the following relations modulo 3. If $|w| \geq 1$,

$$\begin{array}{lcllll} \mathrm{val}_{q_0}(aw) & \equiv & \mathrm{val}_{q_0}(w) & + & (-1)^{|w|+1} & \pmod 3 \\ \mathrm{val}_{q_0}(bw) & \equiv & \mathrm{val}_{q_1}(w) & + & (-1)^{|w|}+1 & \pmod 3 \\ \mathrm{val}_{q_1}(aw) & \equiv & \mathrm{val}_{q_1}(w) & + & (-1)^{|w|+1} & \pmod 3 \\ \mathrm{val}_{q_1}(bw) & \equiv & \mathrm{val}_{q_0}(w) & + & (-1)^{|w|}-1 & \pmod 3 \end{array}$$

where we can notice that the last term depends only on $|w| \pmod 2$. Taking these relations into account, we define as in (3.7) the main part of the transition relation:

	$(q_0,0,0)$	$(q_0,1,0)$	$(q_0,2,0)$	$(q_0,0,1)$	$(q_0,1,1)$	$(q_0,2,1)$
a	$(q_0,2,1)$	$(q_0,0,1)$	$(q_0,1,1)$	$(q_0,1,0)$	$(q_0,2,0)$	$(q_0,0,0)$
b	$(q_1,2,1)$	$(q_1,0,1)$	$(q_1,1,1)$	$(q_1,0,0)$	$(q_1,1,0)$	$(q_1,2,0)$

	$(q_1,0,0)$	$(q_1,1,0)$	$(q_1,2,0)$	$(q_1,0,1)$	$(q_1,1,1)$	$(q_1,2,1)$
a	$(q_1,2,1)$	$(q_1,0,1)$	$(q_1,1,1)$	$(q_1,1,0)$	$(q_1,2,0)$	$(q_1,0,0)$
b	$(q_0,0,1)$	$(q_0,1,1)$	$(q_0,2,1)$	$(q_0,1,0)$	$(q_0,2,0)$	$(q_0,0,0)$

For instance, $((q_0,1,0),b,(q_1,0,1)) \in E$ because in the minimal automaton of L, $q_0.b = q_1$ and if $|w| \equiv 0 \pmod 2$, then $1+(-1)^{|w|}+1 \equiv 0 \pmod 3$. To conclude, observe that $\mathrm{val}_{q_0}(a) = 1$, $b \notin L_{q_0}$, $a \notin L_{q_1}$ and $\mathrm{val}_{q_1}(b) = 0$. So, $((q_0,0,1),a,f)$ and $((q_1,1,1),b,f)$ also belong to the relation defining the NFA.

To reach the final state, the words of even, respectively odd, length have to be read starting from the initial state $(q_0,0,0)$, respectively the second initial state $(q_0,0,1)$. If the reading of a word begins in the wrong initial state with respect to the parity of its length, then no path can reach the final state.

In the general framework of abstract numeration systems, we can consider several kinds of questions about $\mathcal{S}$-recognisability of sets of integers. They are natural extensions of those considered in the classical context of positional numeration systems.

- For a given set $X \subseteq \mathbb{N}$, can we build an ANS $\mathcal{S}$ such that X is $\mathcal{S}$-recognisable?
- For a given ANS $\mathcal{S}$, what kind of arithmetic operations on sets of integers do preserve $\mathcal{S}$-recognisability?
- For a given ANS $\mathcal{S}$, what can be said about the $\mathcal{S}$-recognisable subsets of $\mathbb{N}$?
- In particular, can we obtain some characterisation (logical, arithmetic, whatever...) of the $\mathcal{S}$-recognisable subsets of $\mathbb{N}$?
- How $\mathcal{S}$-recognisability is dependent on the ANS?

As the reader may observe many challenging questions can be considered in this context, also see the bibliographic notes at the end of the chapter for other related questions. We are far from being able to answer all of them but in the next pages, we will develop some of these topics. Also the use of ANS casts some new light on well-known results occurring in the classical context. Let us start with the following result.

Proposition 3.3.4 (Translation by a constant) *Let $\mathcal{S} = (L, A, <)$ be an ANS. If $X \subseteq \mathbb{N}$ is $\mathcal{S}$-recognisable, then also $X + t$ is $\mathcal{S}$-recognisable for all $t \in \mathbb{N}$.*

Proof See for instance (Lecomte and Rigo 2001). Taking into account the theory of synchronised relations (Frougny and Sakarovitch 1993), the *successor* map defined on L by $w \mapsto \mathrm{rep}_{\mathcal{S}}(\mathrm{val}_{\mathcal{S}}(w) + 1)$ is shown to be realised by a left letter-to-letter finite transducer (see Corollary 2.6.11) and the conclusion follows. Also the paper (Angrand and Sakarovitch) is relevant in that context, see Proposition 2.6.14. □

The following result is proved in (Shallit 1994), also see Proposition 2.6.4.

Lemma 3.3.5 *Let L be a regular language over the totally ordered alphabet $(A, <)$. The following languages are regular:*

$$\mathsf{minlg}\,(L) = \{u \in L \mid w \in L, w \neq u, |w| = |u| \Rightarrow u \prec w\}\ ,$$

$$\mathsf{Maxlg}\,(L) = \{u \in L \mid w \in L, w \neq u, |w| = |u| \Rightarrow w \prec u\}\ .$$

The following observation is an immediate reformulation of the above result. Because it will be used quite often we state it as a lemma.

Lemma 3.3.6 *Let $\mathcal{S} = (L, A, <)$ be an ANS. The set $\{\mathcal{V}_L(n) \mid n \geq 0\} = \{\mathrm{Card}(L \cap A^{\leq n}) \mid n \geq 0\}$ is $\mathcal{S}$-recognisable.*

Example 3.1.17 about the set of triangular numbers $\{P(n) \mid n \geq 0\}$ where $P(n) = n(n+1)/2$ can be revisited in light of this result.

3.3.1 Building ANS to recognise specific sets

Considering an infinite set $X \subseteq \mathbb{N}$ we can under particular circumstances look for an ANS $\mathcal{S} = (L, A, <)$ such that $X = \{\mathcal{V}_L(n) \mid n \geq 0\}$. This is the case when X has the form given in the following result whose proof is the main goal of this subsection.

Theorem 3.3.7 *Let $m \geq 1$. For $i = 1, \ldots, m$, let P_i be polynomials belonging to $\mathbb{Q}[X]$ such that $P_i(\mathbb{N}) \subseteq \mathbb{N}$ and let c_i be non-negative integers. Set*

$$f : \mathbb{N} \to \mathbb{N},\ n \mapsto \sum_{i=1}^{m} P_i(n)\, c_i^n\ .$$

The range $f(\mathbb{N})$ is $\mathcal{S}$-recognisable, for some ANS $\mathcal{S}$ which can be effectively constructed.

The idea of the proof is to build a suitable regular language L having the "right" counting function *i.e.*, such that $f(\mathbb{N}) = \{\mathcal{V}_L(n) \mid n \geq 0\}$ or $\mathcal{U}_L(n) = f(n+1) - f(n)$ for large enough n. In view of Lemma 3.1.4 and Theorem 3.1.8, it seems reasonable to build such a regular language. Then the conclusion will trivially follow from Lemma 3.3.6. Note that this result can also be related to the work of (Carton and Thomas 2002) and this connection will be discussed in Section 3.4.1.

Example 3.3.8 It is a classical result that, for all integer bases $k \geq 2$, the set of squares is never k-recognisable, see again (Eilenberg 1974). See Example 1.3.16 for the base 10 case. Nevertheless, one can observe that $(n+1)^2 - n^2 = 2n+1$ and the language $L = a^*b^* \cup a^*c^*$ has exactly $2n+1$ words of length n for all $n \geq 0$. Hence $\{n^2 \mid n \geq 0\}$ is $\mathcal{S}$-recognisable for any ANS based on L whatever is the total ordering on $\{a, b, c\}$.

Remark 3.3.9 In the above discussion, the $\mathcal{S}$-recognisability of the considered set does not depend on the ordering of the alphabet. What only matters is to apply Lemma 3.3.5 to the function $\mathcal{V}_L$ which remains unaffected when reordering the alphabet.

Definition 3.3.10 Let x and y be two words in A^*. The *shuffle* of x and y is the finite language $x \sqcup\!\sqcup y$ defined by

$$\{x_1 y_1 \cdots x_n y_n \mid x = x_1 \cdots x_n, y = y_1 \cdots y_n,\ n \geq 1,\ x_i, y_i \in A^*\}\ .$$

The *shuffle* of two languages $L_1, L_2 \subseteq A^*$ is the language

$$L_1 \sqcup\!\sqcup L_2 = \{w \mid \exists x \in L_1, y \in L_2 : w \in x \sqcup\!\sqcup y\} = \bigcup_{\substack{x \in L_1,\\ y \in L_2}} x \sqcup\!\sqcup y\ .$$

If L_1, L_2 are regular then also $L_1 \sqcup\!\sqcup L_2$ is regular, see for instance (Eilenberg 1974, Proposition 3.5).

For each $k \in \mathbb{N}$, we build recursively a regular language $L[n \mapsto n^k]$ such that $\mathcal{U}_{L[n \mapsto n^k]}(n) = n^k$ for all $n \in \mathbb{N}$. The first two languages $L[n \mapsto 1]$ and

$L[n \mapsto n]$ are defined by $L[n \mapsto 1] = a^*$ and $L[n \mapsto n] = a^+b^*$. Let $k \geq 2$ and assume that we have $L[n \mapsto n^0], \ldots, L[n \mapsto n^{k-1}]$ at our disposal. The induction step relies on the fact that if, for all $n \geq 0$, $\mathcal{U}_M(n) = (n+1)^{k-1}$ then $\mathcal{U}_{M \sqcup\!\sqcup \{c\}}(n) = n^k$ provided that c is not a letter in $\mathrm{alph}(M)$. Indeed, for each of the $(n+1)^{k-1}$ words w of length n in M, $w \sqcup\!\sqcup c$ contains $n+1$ words of length $n+1$. So there are exactly $(n+1)^k$ words of length $n+1$ in $M \sqcup\!\sqcup \{c\}$. Due to

$$(n+1)^{k-1} = \sum_{j=0}^{k-1} \binom{k-1}{j} n^j$$

we build M as a finite union of the languages $L[n \mapsto n^0], \ldots, L[n \mapsto n^{k-1}]$ written *over pairwise disjoint alphabets* $A_{i,j}$, *i.e.*, if $(i,j) \neq (i',j')$, then $A_{i,j} \cap A_{i',j'} = \emptyset$:

$$M = \bigcup_{j=0}^{k-1} \bigcup_{i=1}^{\binom{k-1}{j}} L_{i,j}$$

where $L_{i,j} \subseteq A_{i,j}^*$ is a copy of $L[n \mapsto n^j]$.

Proposition 3.3.11 *Let $P \in \mathbb{N}[X]$. There exists an ANS $\mathcal{S} = (L, A, <)$ such that $P(\mathbb{N})$ is $\mathcal{S}$-recognisable.*

Proof The case where P is constant, is trivial. By Proposition 3.3.4 we may assume that $P(0) = 0$. Since the polynomial $P(n+1) - P(n)$ only contains powers of n with non-negative integer coefficients, by a union of copies of languages $L[n \mapsto n^k]$ over disjoint alphabets we can build a regular language $L \subseteq A^*$ such that $\mathcal{U}_L(n) = P(n+1) - P(n)$. Fix a total ordering $<$ on A and let $\mathcal{S} = (L, A, <)$.

To conclude the proof, we still need to find some integer ℓ such that the first word of length ℓ in L has $P(\ell)$ as numerical value. From the above discussion, this will imply that, for all $n \geq \ell$, $P(n)$ is the numerical value of the first word of length n in L. This is the aim of the next paragraph.

We can assume that $\varepsilon \in L$ and that the first word w of length 2 in the genealogically ordered language L is such that $\mathrm{val}_{\mathcal{S}}(w) = P(2)$. Indeed, adding or removing a finite number of words of length 1 in a regular language does not alter its regularity. We can add new letters to the alphabet to increase at will the number of words of length 1. Note that we have to consider words of length 2 and not words of length 1 because $P(1)$ is not necessarily equal to one and therefore cannot possibly be represented by the first word of length 1. Contrarily to words of length 1, there is a single

word of length 0 so we have no freedom to modify the number of words of length 0.

Let $n \geq 2$. Since $\mathcal{U}_L(n) = P(n+1) - P(n)$, if the numerical value of the first word of length n is $P(n)$ then the numerical value of the first word of length $n+1$ is $P(n+1)$. Consequently, we have

$$\mathrm{rep}_{\mathcal{S}}(P(\mathbb{N}) \setminus \{P(1)\}) = Min_{\prec}(L \setminus A) \ .$$

By Lemma 3.3.5, $P(\mathbb{N})$ is $\mathcal{S}$-recognisable. A single word should possibly be added to take into account the $\mathcal{S}$-representation of $P(1)$. □

Lemma 3.3.12 *Let k and t be two positive integers. There exists a regular language $L[n \mapsto n^k - t\, n^{k-1}]$ such that*

$$\mathcal{U}_{L[n \mapsto n^k - t\, n^{k-1}]}(n) = \begin{cases} n^k - t\, n^{k-1}, & \text{if } n \geq t, \\ 0, & \text{otherwise.} \end{cases}$$

Proof For $k = 1$ take the language $L[n \mapsto n^1 - t\, n^0] = a^{t+1}a^*b^*$. Now assume that $k \geq 2$. From the above discussion, we have $L[n \mapsto n^k] = M \sqcup\!\sqcup \{a\}$ where $L[n \mapsto n^k] \subseteq A^*$ and a not belonging to $\mathrm{alph}(M)$. Let $n \geq 1$. For $i = 1, \ldots, n$, $L[n \mapsto n^k]$ has exactly n^{k-1} words of length n with a occurring at position i (say, counted from the right). The language

$$L[n \mapsto n^k - t\, n^{k-1}] = L[n \mapsto n^k] \setminus \bigcup_{i=0}^{t-1} A^*\, a\, A^i \tag{3.8}$$

has $n^k - t\, n^{k-1}$ words of length n for $n \geq t$. □

Proposition 3.3.13 *Let $P \in \mathbb{Z}[X]$ be such that $P(\mathbb{N}) \subseteq \mathbb{N}$. There exists an ANS $\mathcal{S} = (L, A, <)$ such that $P(\mathbb{N})$ is $\mathcal{S}$-recognisable.*

Proof Without loss of generality, we may assume that $\deg(P) = d + 1 \geq 1$. We proceed as in the proof of Proposition 3.3.11 and consider the polynomial $Q(n) = P(n+1) - P(n)$. Since $P(\mathbb{N}) \subseteq \mathbb{N}$, the leading coefficients of P and Q are positive. By possibly adding extra terms of the form $X^j - X^j$, if $\deg(Q) = d$ then to take advantage of the previous lemma, $Q(X)$ can be written as

$$\sum_{\ell=0}^{d} c_\ell\, X^\ell + X^{i_1+1} - t_1\, X^{i_1} + \cdots + X^{i_r+1} - t_r\, X^{i_r} \tag{3.9}$$

for some $c_0, \ldots, c_d \in \mathbb{N}$, $i_1, \ldots, i_r \in \{0, \ldots, d-1\}$ and $t_1, \ldots, t_r \in \mathbb{N} \setminus \{0\}$. Let $t = \sup\{t_1, \ldots, t_r, 2, m\}$ where m is the least integer such that $P(n) < P(n+1)$ for all $n \geq m$. Making the union of regular languages over

disjoint alphabets of the kind $L[n \mapsto n^\ell]$ and $L[n \mapsto n^{i_j} - t_j\, n^{i_j-1}]$ given by Lemma 3.3.12, we get a regular language L satisfying, for all $n \geq t$, $\mathcal{U}_L(n) = Q(n)$.

Since $t \geq 2$, we can assume that L contains exactly $P(t)$ words of length at most $t-1$. This can be achieved by adding or removing a finite number of words from the language L. Let $\mathcal{S}$ be an ANS based on the ordered regular language L. The first word of length t has a numerical value equal to $P(t)$ and, for all $n \geq t$, $\mathcal{U}_L(n) = P(n+1) - P(n)$. Then we get

$$\mathrm{rep}_{\mathcal{S}}(P(\mathbb{N})) = (\mathsf{minlg}\,(L) \cap A^{\geq t}) \cup \{\mathrm{rep}_{\mathcal{S}}(P(0)), \ldots, \mathrm{rep}_{\mathcal{S}}(P(t-1))\}$$

where $A^{\geq t}$ denotes the set of all words of length at least t. By Lemma 3.3.5, $\mathrm{rep}_{\mathcal{S}}(P(\mathbb{N}))$ is regular. □

As the third step we get a theorem of recognisability in the general case of polynomials with rational coefficients. Interestingly, the proof relies on the $\mathcal{S}$-recognisability of arithmetic progressions.

Proposition 3.3.14 *Let $P \in \mathbb{Q}[X]$ be such that $P(\mathbb{N}) \subseteq \mathbb{N}$. There exists an ANS $\mathcal{S} = (L, A, <)$ such that $P(\mathbb{N})$ is $\mathcal{S}$-recognisable.*

Proof Assume that $\deg(P) = d \geq 1$. Let $s_0, \ldots, s_d, c_d \in \mathbb{N} \setminus \{0\}$ and $c_0, \ldots, c_{d-1} \in \mathbb{Z}$ be such that

$$P(X) = \frac{c_d}{s_d}\, X^d + \frac{c_{d-1}}{s_{d-1}}\, X^{d-1} + \cdots + \frac{c_0}{s_0}.$$

Let s be the least common multiple of $s_0, \ldots, s_d$. One has $sP = Q$ with $Q \in \mathbb{Z}[X]$. Since $P(\mathbb{N}) \subseteq \mathbb{N}$, then $Q(\mathbb{N}) \subseteq s\,\mathbb{N}$. As in the proof of Proposition 3.3.13, there exist $t = \sup\{2, m\}$ where m is the least integer such that $P(n) < P(n+1)$, for all $n \geq m$, and a regular language M over a totally ordered alphabet $(A, <)$ such that, for all $n \geq t$,

$$\mathcal{U}_M(n) = Q(n+1) - Q(n) = s\,(P(n+1) - P(n)) \ .$$

We modify M by possibly adding or removing a finite number of words to get $\mathcal{V}_M(t-1) = s\,P(t) = Q(t)$. Otherwise stated, if we set $\mathcal{R} = (M, A, <)$ and w is the first word of length t in M, then $\mathrm{val}_{\mathcal{R}}(w) = Q(t)$. By Theorem 3.3.1, the arithmetic progression $s\,\mathbb{N}$ is $\mathcal{R}$-recognisable. Consequently, $L = \mathrm{rep}_{\mathcal{R}}(s\,\mathbb{N})$ is a regular language such that

$$\mathcal{V}_L(t-1) = P(t) \text{ and, } \forall n \geq t,\ \mathcal{U}_L(n) = P(n+1) - P(n)\ .$$

Indeed L is obtained by taking in the genealogically ordered language M the words at position $is+1$, $i \in \mathbb{N}$. Since the first word of length t in M is the first word of length t in L and its position in the genealogically ordered language L is $P(t)$, the conclusion follows from Lemma 3.3.5. □

Proposition 3.3.15 *Let $c \in \mathbb{N} \setminus \{0,1\}$ and P be a polynomial in $\mathbb{Q}[X]$ such that $P(\mathbb{N}) \subseteq \mathbb{N}$. There exists a numeration system $\mathcal{S}$ such that the set $\{P(n)\, c^n \mid n \in \mathbb{N}\}$ is $\mathcal{S}$-recognisable.*

Proof First assume that $P \in \mathbb{Z}[X]$ and that it is non-constant. We show how to construct a regular language L such that for all large enough n,

$$\mathcal{U}_L(n) = P(n+1)\, c^{n+1} - P(n)\, c^n = [c\, P(n+1) - P(n)]\, c^n \ .$$

The assumption $P(\mathbb{N}) \subseteq \mathbb{N}$ implies that the polynomial $c\, P(n+1) - P(n) \in \mathbb{Z}[X]$ has a positive leading coefficient. We can apply the same decomposition as in (3.9) and therefore proceed as in the proof of Proposition 3.3.13.

To get such a language L, it is enough to show how to construct, for all $k \geq 0$, a regular language $L[n \mapsto n^k c^n]$ having $n^k\, c^n$ words of length $n \geq 0$ and, for all $t \geq 1$, a regular language with $(n^k - t\, n^{k-1})\, c^n$ words of length $n \geq t$. As an intermediate step, also we construct, for all $k > i \geq 0$, regular languages $M_{k,i}$ having $n^i c^n$ words of length $n - k + i$ for all $n > k$.

If $\mathrm{Card}(A) = c$, note that $L[n \mapsto n^0 c^n] = A^*$. So we build $L[n \mapsto n^1 c^n]$ and $M_{1,0}$ first. Let $A_1, \ldots, A_c$ be c pairwise disjoint alphabets of cardinality c. The language $M_{1,0} = A_1^* \cup \cdots \cup A_c^*$ is such that $\mathcal{U}_{M_{1,0}}(n-1) = c^n$ for all $n > 1$. Let a_1 be a letter not in $\mathrm{alph}(M_{1,0})$. To obtain $L[n \mapsto n^1 c^n]$ we take the words of length at least 2 in $M_{1,0} \sqcup\!\sqcup \{a_1\}$ and add c distinct words of length 1.

Let $k \geq 2$. Assume that we have $M_{k-1,0}, \ldots, M_{k-1,k-2}$ at our disposal. We have to construct languages $M_{k,0}, \ldots, M_{k,k-1}$ and $L[n \mapsto n^k c^n]$. Let $A_1, \ldots, A_{c^k}$ be c^k pairwise disjoint alphabets of cardinality c. The language $M_{k,0} = A_1^* \cup \cdots \cup A_{c^k}^*$ is such that $\mathcal{U}_{M_{k,0}}(n-k) = c^n$ for all $n > k$. Now assume that we have $M_{k,i}$ for some $i < k-1$. Let a_{i+1} be a letter not in $\mathrm{alph}(M_{k,i})$. Then for $n > k$, $M_{k,i} \sqcup\!\sqcup \{a_{i+1}\}$ has $n^i(n-k+i+1)c^n$ words of length $n-k+i+1$ because for each word of length $n-k+i$ in $M_{k,i}$ we can put the extra letter a_{i+1} in $n-k+i+1$ positions. To get $M_{k,i+1}$ we make the union of $M_{k,i} \sqcup\!\sqcup \{a_{i+1}\}$ and $k-i-1$ copies over disjoint alphabets of languages of the kind $M_{k,i}\, a_{i+1}$. The extra letter concatenated at the end of each words in $M_{k,i}$ ensures that $M_{k,i}\, a_{i+1}$ has $n^i c^n$ words of length $n-k+i+1$. Now $M_{k,k-1}$ has, for all $n > k$, $n^{k-1} c^n$ words of length $n-1$. So if a_k does not belong to $\mathrm{alph}(M_{k,k-1})$, we consider the words of length at least $k+1$ in $M_{k,k-1} \sqcup\!\sqcup \{a_k\}$ and we add a suitable number of words of shorter length to get $L[n \mapsto n^k c^n]$. Since a shuffle operation is involved in this latter construction, using the same argument as in (3.8) we can build a regular language having a complexity function $(n^k - t\, n^{k-1})c^n$ for all $n \geq t$.

To conclude the proof, if $P \in \mathbb{Q}[X] \setminus \mathbb{Z}[X]$, then apply the same trick as in the proof of Proposition 3.3.14. □

Repeating the construction given in this latter proof to get several regular languages over distinct alphabets, the reader should be convinced that Theorem 3.3.7 stated at the beginning of this section is derived easily.

3.3.2 ANS based on slender languages

A language L is said to be *slender*, if there exists d such that $\mathcal{U}_L(n) \leq d$ for all $n \geq 0$. For ANS based on slender languages, $\mathcal{S}$-recognisable sets are completely characterised.

Theorem 3.3.16 *Let $L \subseteq A^*$ be a slender regular language and $\mathcal{S} = (L, A, <)$. A set $X \subseteq \mathbb{N}$ is $\mathcal{S}$-recognisable if, and only if, X is a finite union of arithmetic progressions.*

Proof By a well-known characterisation of slender languages†, there exist $k \geq 1$ and words x_i, y_i, z_i, $1 \leq i \leq k$, such that

$$L = \bigcup_{i=1}^{k} x_i\, y_i^* z_i \cup F,\ \ x_i, z_i \in A^*, y_i \in A^+$$

where the sets $x_i\, y_i^* z_i$ are pairwise disjoint and F is a finite set. The sequence $(\mathcal{U}_L(n))_{n\in\mathbb{N}}$ is eventually periodic of period $p = \mathrm{lcm}_i|y_i|$. Moreover, for n large enough, if $x_i\, y_i^n\, z_i$ is the m-th word of length $|x_i\, z_i| + n\,|y_i|$ then $x_i\, y_i^{n+p/|y_i|}\, z_i$ is the m-th word of length $|x_i\, z_i| + n\,|y_i| + p$. Roughly speaking, for sufficiently large n, the structures of the ordered sets of words of length n and $n + p$ are the same. The regular subsets of L are of the form

$$\bigcup_{j\in J} x_{i_j}\, (y_{i_j}^{t_j})^* z_{i_j} \cup F' \tag{3.10}$$

where J is a finite set, $i_j \in \{1, \ldots, k\}$, $t_j \in \mathbb{N}$ and F' is a finite subset of L. Now we can conclude. If X is $\mathcal{S}$-recognisable, then $\mathrm{rep}_{\mathcal{S}}(X)$ is a regular subset of L of the form (3.10). In view of the first part of the proof, it is clear that X is eventually periodic with period $\mathrm{lcm}(p, \mathrm{lcm}_j|y_{i_j}^{t_j}|)$. The converse follows from Theorem 3.3.1. □

Example 3.3.17 Consider the language $L = ab^*c \cup b(aa)^*c$. It contains exactly two words of each positive even length: $ab^{2i}c \prec ba^{2i}c$ and one word for each odd length larger than 2: $ab^{2i+1}c$. The sequence $\mathcal{U}_L(n)$ is eventually periodic of period two: $0, 0, 2, 1, 2, 1, \ldots$.

† See for instance (Păun and Salomaa 1995) or independently (Shallit 1994). Compare this result with the one given in Theorem 3.3.21.

Corollary 3.3.18 *Let $\mathcal{S}$ be a numeration system based on a slender language. If $X, Y \subseteq \mathbb{N}$ are $\mathcal{S}$-recognisable, then $X + Y$ and $t\,X$ are $\mathcal{S}$-recognisable for all $t \in \mathbb{N}$.*

Proof It is clear that if $X, Y \subseteq \mathbb{N}$ are eventually periodic, then $X + Y$ and $t\,X$ are also eventually periodic. □

3.3.3 Multiplication by a constant

From Corollary 3.3.18 given above, if $\mathcal{S} = (L, A, <)$ is an ANS based on a slender language and if $X \subseteq \mathbb{N}$ is $\mathcal{S}$-recognisable, then for any given non-negative integer t, the set tX is again $\mathcal{S}$-recognisable. Such a property is also well-known for the usual k-ary numeration systems and more generally for the Pisot numeration systems sketched in Example 3.1.14. One can therefore consider the following general problem of *characterising ANS $\mathcal{S}$ such that multiplication by a constant is $\mathcal{S}$-recognisability-preserving that is, for all $\mathcal{S}$-recognisable sets $X \subseteq \mathbb{N}$ and for all $t \in \mathbb{N}$, the set $t\,X = \{tx \mid x \in X\}$ is still $\mathcal{S}$-recognisable.* This is the most basic arithmetic operation to consider. A more ambitious task is to consider addition of two $\mathcal{S}$-recognisable sets $X + Y = \{x + y \mid x \in X, y \in Y\}$ and look for ANS such that the resulting set is again $\mathcal{S}$-recognisable. Note that even for the usual integer base systems, if X and Y are two k-recognisable sets of integers, then in general the set $X.Y = \{xy \mid x \in X, y \in Y\}$ is not k-recognisable.

Definition 3.3.19 If $w_1, \ldots, w_n$ are words over A of arbitrary, and not necessarily equal, lengths then the padding $(w_1, \ldots, w_n)^{\#}$ is defined as

$$(\#^{m-|w_1|}w_1, \ldots, \#^{m-|w_n|}w_n)$$

where $m = \max\{|w_1|, \ldots, |w_n|\}$ and $\#$ is a new padding symbol. Such an n-tuple can be considered as a single word over the alphabet $(A \cup \{\#\})^n$ obtained as the Cartesian product of n copies of $A \cup \{\#\}$. The concatenation of two words $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ over $(A \cup \{\#\})^n$ is $(u_1v_1, \ldots, u_nv_n)$ where the usual concatenation product is considered component-wise.

Remark 3.3.20 Assume that addition in an ANS $\mathcal{S} = (L, A, <)$ is computable by finite automaton, *i.e.*, its graph

$$G = \{(\mathrm{rep}_{\mathcal{S}}(x), \mathrm{rep}_{\mathcal{S}}(y), \mathrm{rep}_{\mathcal{S}}(x + y))^{\#} \mid x, y \geq 0\}$$

is regular where the shortest words are padded with an extra symbol $\#$ to get three components of same length, that is, we get words over the Cartesian product $(A \cup \{\#\}) \times (A \cup \{\#\}) \times (A \cup \{\#\})$ and the corresponding

DFA reads simultaneously one symbol from the three components. By considering $G \cap \{(v, v, w)^{\#} \mid v \in L, w \in A^*\}$ then multiplication by 2 is also computable by finite automaton and in particular, multiplication by 2 is therefore $\mathcal{S}$-recognisability-preserving. By iterating this kind of arguments, if addition is computable by finite automaton, then it is the same for multiplication by any constant $t \in \mathbb{N}$.

First let us recap some well-known facts about the complexity function of regular languages. A language is said to be *polynomial* (or *sparse*) if there exists some non-negative integer k such that $\mathcal{U}_L(n) \in \mathcal{O}(n^k)$. If the regular language is infinite, one can show that there exist a constant $C > 0$ and an infinite sequence of integers $n_1 < n_2 < \cdots$ such that $\mathcal{U}_L(n_j) \geq C n_j^k$ for all $j \geq 1$. Infinite slender languages are in particular polynomial. Deterministic finite automata accepting polynomial regular languages have some specific properties. A well-known description of the polynomial regular languages and the dichotomy existing with exponential languages are for instance given in (Szilard, Yu, Zhang, et al. 1994). We recall these two statements below. Obviously only states which are both accessible and co-accessible have an impact on the complexity function (see Chapter 1 for definition of accessibility).

Theorem 3.3.21 *Let $\mathcal{A} = (Q, A, E, \{q_0\}, T)$ be an accessible and co-accessible DFA. The language L accepted by $\mathcal{A}$ is polynomial if, and only if, all states $q \in Q$ belong to at most one cycle in $\mathcal{A}$. In particular, L is polynomial if, and only if, it is a finite union of languages of the form $u_1 v_1^* u_2 v_2^* \cdots u_t v_t^* u_{t+1}$.*

Note that it is algorithmically decidable whether or not the language accepted by a DFA given as an input is polynomial. This question is for instance considered in Theorem 11.1.27.

Theorem 3.3.22 *A regular language L is either polynomial or there exist $C > 0$ and an infinite sequence $n_1 < n_2 < \cdots$ such that the complexity function of L satisfies, for all $j \geq 1$, $\mathcal{U}_L(n_j) = 2^{f(n_j)}$ where $f(n_j) \geq C\, n_j$.*

If a regular language is not polynomial, then we shall say that it is *exponential*. Note that in general, we cannot give a suitable lower bound on $\mathcal{U}_L(n)$ for every large enough n. It is the reason why in the above theorem, we have not written $f \in \Omega(n)$ but instead have given a lower bound for infinitely many n. Consider for instance the regular language $L = \{aa, ab, ba, bb\}^*$. We have, for all $n \geq 0$, $\mathcal{U}_L(2n) = 2^{2n}$ and $\mathcal{U}_L(2n+1) = 0$. Therefore this language is exponential but even in this case, for infinitely many m, $\mathcal{U}_L(m) = 0$.

Next we sketch a picture of the main results about preservation of $\mathcal{S}$-recognisability after multiplication by a constant.

Proposition 3.3.23 *Let $\mathcal{S} = (a^*b^*, \{a, b\}, a < b)$. If $t \in \mathbb{N}$ is an odd square, then for all $\mathcal{S}$-recognisable set $X \subseteq \mathbb{N}$, $t\,X$ is again $\mathcal{S}$-recognisable. Otherwise, there exists an $\mathcal{S}$-recognisable set $Y \subseteq \mathbb{N}$ such that $t\,Y$ is not $\mathcal{S}$-recognisable.*

The proof is given in (Lecomte and Rigo 2001) and involves Pell equations. It is due to the expression (3.2) of $\mathrm{val}_{\mathcal{S}}(a^ib^j)$ which is a polynomial of degree 2 in i and j. Let us point out that the arguments developed in this proof are also useful to give a counter-example showing that $\mathcal{S}$-automaticity is not preserved after periodic deletion, see (Rigo and Maes 2002). The fact revealed in the previous proposition when t is not a square is a special case of some general phenomenom (Rigo 2002) about ANS on polynomial regular languages.

Theorem 3.3.24 *Let $\mathcal{S}$ be an ANS based on a polynomial regular language L. Suppose that there exist $C > 0$ and an integer $k \geq 1$ such that $\mathcal{U}_L(n) \in \mathcal{O}(n^k)$ and for infinitely many n, $\mathcal{U}_L(n) \geq Cn^k$. If t is not the $(k+1)$th power of an integer, then there exists an $\mathcal{S}$-recognisable set $Y \subseteq \mathbb{N}$ such that $t\,Y$ is not $\mathcal{S}$-recognisable.*

Having this theorem at hand, it seems natural to determine which suitable powers may preserve recognisability. For specific kind of polynomial languages of any degree, we have the following result. Notice that for $n = 1$, we have a slender language and the case $n = 2$ is exactly the one considered in Proposition 3.3.23.

Proposition 3.3.25 (Charlier, Rigo, and Steiner 2008) *Let $n \geq 3$. Let A be the ordered alphabet $\{a_1 < \cdots < a_n\}$ and $\mathcal{S} = (a_1^* \cdots a_n^*, A, <)$ be an ANS. For all $t \geq 2$, there exists an $\mathcal{S}$-recognisable set $Y \subseteq \mathbb{N}$ such that $t\,Y$ is not $\mathcal{S}$-recognisable.*

Also details are given in (Charlier 2009). In general, exponential languages with a polynomial complement do not preserve recognisability after multiplication by a constant.

Proposition 3.3.26 *Let A be an alphabet such that $\mathrm{Card}(A) \geq 2$. Let $L \subset A^*$ be an infinite polynomial regular language and $\mathcal{S}$ be an ANS based on its complement $A^* \setminus L$. There exists an $\mathcal{S}$-recognisable set $Y \subseteq \mathbb{N}$ and an integer t such that $t\,Y$ is not $\mathcal{S}$-recognisable.*

Arguments appearing in Example 3.1.7 and in the proof of this proposition as given in (Rigo 2002) are based on similar techniques involving the pumping lemma.

In view of these results and except for very special cases like the slender languages or A^*, we can conclude that the only regular languages that are possibly suited to define ANS for which recognisability is preserved after arithmetic operations, are necessarily exponential languages with exponential complement.

3.4 Automatic sequences

In (Cobham 1972), it is shown that an infinite word $x = x_0x_1x_2\cdots$ over an alphabet B is obtained as the image under a coding $\tau : A \to B$ of the fixed point of a morphism $\sigma : A \to A^*$ of *constant length* k if, and only if, there exists a DFAO $\mathcal{A} = (Q, A, \delta, \{q_0\}, B, \mu)$ such that $x_n = \mu(\delta(q_0, \mathrm{rep}_k(n)))$ for all $n \geq 0$. This result closely relates the constant length of the morphism to the base k numeration system and also explains the consecrated terminology of k-automatic sequences †, *i.e.*, the nth term of the sequence is generated by an automaton with output "fed" with the k-ary representation of n.

A natural generalisation of this iterative process used to define infinite words is to relax the hypothesis about the constant length of σ and to consider an arbitrary non-erasing prolongable morphism $\sigma : A \to A^*$ and an extra coding. Hence what kind of numeration system should replace the usual k-ary one? In this section, we prove the following analogue of Cobham's theorem from 1972 where ANS come into play and we discuss some of its applications to $\mathcal{S}$-recognisable sets.

Theorem 3.4.1 *Let $x = x_0x_1x_2\cdots$ be an infinite word over an alphabet B. This word is substitutive if, and only if, there exists an ANS $\mathcal{S} = (L, A, <)$ and a DFAO $(Q, A, \delta, \{q_0\}, B, \mu)$ such that for all $n \geq 0$, $x_n = \mu(\delta(q_0, \mathrm{rep}_{\mathcal{S}}(n)))$.*

This result splits into Propositions 3.4.12 and 3.4.16. The next definition naturally extends the classical generation process of k-automatic sequences.

† Originally A. Cobham was using the terminology of *tag sequences* referring to the generation process of the infinite word. Let us quote (Cobham 1972): "Adding a feedback feature which permits symbols produced at early stages of the generating process to be re-examined at later stages increases flexibility and the variety of sequences generable by devices so augmented is substantially richer.... Suppose we have generated symbols with index 0 through $2k-1$ and that our left hand points at the k-th of these, our right hand at the last. We observe the symbol at which our left hand is pointing and write with our right the $2k$-th and $(2k+1)$-st as prescribed. Moving our left hand one symbol to the right, we are in position to repeat the procedure.".

Definition 3.4.2 Let $\mathcal{S} = (L, A, <)$ be an ANS. We say that an infinite word $x = x_0x_1x_2\cdots \in B^{\mathbb{N}}$ is *$\mathcal{S}$-automatic*, if there exists a DFAO $(Q, A, \delta, \{q_0\}, B, \mu)$ such that $x_n = \mu(\delta(q_0, \mathrm{rep}_{\mathcal{S}}(n)))$ for all $n \geq 0$.

Example 3.4.3 We consider the alphabets $A = \{a, b\}$, $B = \{0, 1, 2, 3\}$, the ANS $\mathcal{S} = (a^*b^*, A, a < b)$ of Example 3.1.12 and the DFAO depicted in Figure 3.5 We obtain the first few terms of the corresponding $\mathcal{S}$-automatic

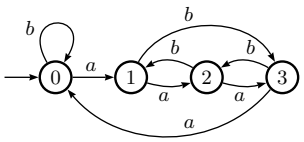

Fig. 3.5. A DFAO with output alphabet $\{0, 1, 2, 3\}$.

sequence $x = 0102303120023101012302303120312023100231012\cdots$.

Notice that taking the ANS $\mathcal{R} = (\{a, ba\}^*\{\varepsilon, b\}, \{a, b\}, a < b)$ of Example 3.1.15 we obtain another infinite word $y = 01023\underline{1}31\underline{023}\cdots$ which is $\mathcal{R}$-automatic (underlined letters indicate the differences between x and y). This stresses the fact that an $\mathcal{S}$-automatic sequence really depends on two ingredients: an ANS and a DFAO.

Example 3.4.4 Let $\mathcal{S} = (B_k, \{0, \ldots, k-1\}, <)$ be the ANS corresponding to the usual k-ary numeration system where the language $B_k = \{0, \ldots, k-1\}^* \setminus 0\{0, \ldots, k-1\}^*$ is as in Example 3.1.11. By considering a DFAO $(Q, \{0, \ldots, k-1\}, \delta, \{q_0\}, B, \mu)$, the sequence defined, for all $n \geq 0$, by $x_n = \mu(\delta(q_0, \mathrm{rep}_{\mathcal{S}}(n)))$ is $\mathcal{S}$-automatic. So k-automatic sequences are special cases of $\mathcal{S}$-automatic sequences.

The next lemma is a powerful result that allows to get rid of the erasing behaviour that can appear in the two morphisms used for generating a substitutive word and restricts the second one to a coding. A proof† of this result is given in (Allouche and Shallit 2003). This result is also expressed by Theorem 4.6.1.

Lemma 3.4.5 (Cobham 1968) *Let A, B, C be three alphabets. Consider two arbitrary morphisms $\sigma : A \to A^*$ and $\tau : A \to B^*$ such that $\tau(\sigma^\omega(a))$ is an infinite word. There exist a non-erasing morphism $\alpha : C \to C^+$ prolongable on a letter $c \in C$ and a coding $\beta : C \to B$ such that*

$$\tau(\sigma^\omega(a)) = \beta(\alpha^\omega(c)) \ .$$

† Have a look, one needs to define *dead* and *moribund* letters.

The idea of the following result is to consider the end point, being or not a final state does not matter, of all paths that can be achieved in a DFA. These paths are naturally genealogically ordered with respect to their label. Recall that S is the shift operator introduced in Chapter 1.

Lemma 3.4.6 *Let $A = \{a_1 < \cdots < a_n\}$ be a totally ordered alphabet, $\mathcal{A} = (Q, A, E, \{q_0\}, T)$ be a DFA where E defines a partial function $\delta_{\mathcal{A}} : Q \times A \to Q$ and let $z \notin Q$. Define the morphism $\psi_{\mathcal{A}} : Q \cup \{z\} \to (Q \cup \{z\})^*$ by $\psi_{\mathcal{A}}(z) = z\, q_0$ and, for all $q \in Q$,*

$$\psi_{\mathcal{A}}(q) = \delta_{\mathcal{A}}(q, a_1) \cdots \delta_{\mathcal{A}}(q, a_n) \ .$$

In this latter expression, if $\delta_{\mathcal{A}}(q, a_i)$ is not defined for some i, then it is replaced by ε. Let L be the regular language accepted by $(Q, A, E, \{q_0\}, Q)$ where all states of $\mathcal{A}$ are final. Then the shifted sequence $S(\psi^{\omega}_{\mathcal{A}}(z))$ is the sequence $(x_n)_{n\in\mathbb{N}}$ of the states reached in $\mathcal{A}$ by the words of L in genealogical order, i.e.*, for all $n \in \mathbb{N}$,*

$$x_n = \delta_{\mathcal{A}}(q_0, w_n)$$

where w_n is the $(n+1)$st word of the genealogically ordered language L.

Prior to the proof, let us give a short example to set properly the framework.

Example 3.4.7 Consider the DFA given in Figure 3.6. Note that the automaton is not complete, the transition function δ is partial: from q_1 one cannot read b. Assume $a < b$. The sequence of the ordered words in the

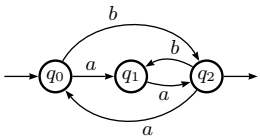

Fig. 3.6. A DFA.

language accepted by the automaton where all states are considered as final states are

$$(w_n)_{n\geq 0} = \varepsilon,\ a,\ b,\ aa,\ ba,\ bb,\ aaa,\ aab,\ baa,\ bab,\ bba,\ aaaa, \dots \ .$$

The corresponding sequence of states is

$$(\delta(q_0, w_n))_{n\geq 0} = q_0, q_1, q_2, q_2, q_0, q_1, q_0, q_1, q_1, q_2, q_2, q_1, \dots \ .$$

For instance, the second q_2 in the sequence, *i.e.*, its fourth element, is the

state reached by the DFA when reading aa, *i.e.*, the fourth word w_3, from q_0. Now consider the morphism

$$\psi : \begin{cases} z & \mapsto & z\,q_0 \\ q_0 & \mapsto & q_1\,q_2 \\ q_1 & \mapsto & q_2 \\ q_2 & \mapsto & q_0\,q_1\ . \end{cases}$$

One can observe that the introduction of the extra letter z gives a prolongable morphism: in this example only $\psi(z)$ begins with z, for $x \in \{q_0, q_1, q_2\}$, $\psi(x)$ does not start with x. Now, one can compute the prefix of $\psi^\omega(z)$,

$$\psi^\omega(z) = z\ q_0\ q_1\ q_2\ q_2\ q_0\ q_1\ q_0\ q_1\ q_1\ q_2\ q_2\ q_1\ q_2\ q_2\ q_2\ \cdots\ .$$

Now let us give the proof of Lemma 3.4.6.

Proof First observe that we have the following factorisation:

$$\psi^\omega_{\mathcal{A}}(z) = z x_0 x_1 x_2 \cdots = z\,q_0\,\psi_{\mathcal{A}}(q_0)\,\psi^2_{\mathcal{A}}(q_0)\cdots$$

and $x_0 = q_0 = \delta_{\mathcal{A}}(q_0, \varepsilon)$. Then by definition of $\psi_{\mathcal{A}}$, if $x_n = \delta_{\mathcal{A}}(q_0, w_n)$, $n \geq 0$, then the factor

$$u_n = \psi(x_n) = \delta_{\mathcal{A}}(q_0, w_n a_1) \cdots \delta_{\mathcal{A}}(q_0, w_n a_n) \tag{3.11}$$

appears in $\psi^\omega_{\mathcal{A}}(z)$ with the usual convention of replacing with ε the undefined transitions. Indeed, $z x_0 x_1 x_2 \cdots$ is a fixed point of $\psi_{\mathcal{A}}$ and each x_n produces a factor $\psi_{\mathcal{A}}(x_n) = u_n$ appearing later on in the infinite word. Moreover this factor is preceded by $\delta_{\mathcal{A}}(q_0, w_{n-1}a_1) \cdots \delta_{\mathcal{A}}(q_0, w_{n-1}a_n)$ and followed by $\delta_{\mathcal{A}}(q_0, w_{n+1}a_1) \cdots \delta_{\mathcal{A}}(q_0, w_{n+1}a_n)$. It is therefore clear that we get all states reached from the initial state when considering in increasing genealogical order the labels of all the paths in $\mathcal{A}$. □

Remark 3.4.8 We use the notation of Lemma 3.4.6. Note that the morphism $\psi_{\mathcal{A}}$ given in the previous statement depends only on Q, A and E but not on the set of final states T. In the literature, one can find the terminology *transition structure* when final states are unspecified or unimportant. In particular, there is no relation between the language recognised by $\mathcal{A}$ and the infinite word $S(\psi^\omega_{\mathcal{A}}(z))$.

If $\mathcal{A}$ contains no cycle, *i.e.*, if the language recognised by $\mathcal{A}$ is finite, then the fixed point of $\psi_{\mathcal{A}}$ starting with z is also finite.

Example 3.4.9 Also the reader could have the feeling that the introduction of the extra letter z such that $\psi_{\mathcal{A}}(z) = z q_0$ is artificial. In particular, if

the initial state has a loop. Let us consider the following example given by the DFA depicted in Figure 3.7. Let us compare the infinite words generated

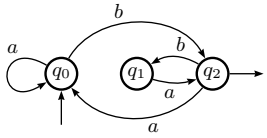

Fig. 3.7. Another DFA.

by the morphism

$$\mu : \begin{cases} q_0 & \mapsto & q_0\, q_2 \\ q_1 & \mapsto & q_2 \\ q_2 & \mapsto & q_0 q_{\,1} \end{cases}$$

and by the morphism $\psi_{\mathcal{A}}$ given by Lemma 3.4.6 and defined by $\psi_{\mathcal{A}}(z) = zq_0$ and $\psi_{\mathcal{A}}(x) = \mu(x)$ for $x \in \{q_0, q_1, q_2\}$. We get

$$\psi_{\mathcal{A}}^{\omega}(z) = zq_0q_0q_2q_0q_2q_0q_1q_0q_2q_0q_1q_0q_2q_2q_0q_2\cdots$$

but

$$\mu^{\omega}(q_0) = q_0q_2q_0q_1q_0q_2q_2q_0q_2q_0q_1q_0q_1q_0q_2\cdots\ .$$

One can show that the sequence $\mu^{\omega}(q_0)$ is the sequence of states reached from q_0 by considering only words in the DFA starting with b instead of taking into account all the possible paths. Of course, one cannot simply remove the loops of label a from q_0 because it may be used not only by paths starting with a.

In Lemma 3.4.6, we have in a canonical way associated with any DFA, even with any labelled directed graph, a morphism. Now we present some kind of converse construction. Note that this construction is very close to the prefix-suffix graph introduced in Definition 5.2.4.

Definition 3.4.10 Let us adapt a classical construction encountered in the case of k-automatic sequences. Any pair given by a morphism $\sigma : A \to A^*$ and a letter $a \in A$ can be canonically associated with a DFA denoted $\mathcal{A}_{\sigma,a}$ and defined as follows. Let $\|\sigma\| = \max_{b\in A} |\sigma(b)|$. The alphabet of $\mathcal{A}_{\sigma,a}$ is $\{1, \ldots, \|\sigma\|\}$, its set of states is A. The initial state is a. For all $b \in A$ and $i \in \{1, \ldots, |\sigma(b)|\}$, we set $\delta(b, i) = \sigma(b)[i, i]$ to define the partial transition function of $\mathcal{A}_{\sigma,a}$. There is usually no need to specify the final states. One can for instance set $T = A$ as the set of final states.

Moreover, if an extra morphism $\tau : A \to B^*$ is given, then we extend

$\mathcal{A}_{\sigma,a}$ to define a DFAO $\mathcal{A}_{\sigma,a,\tau}$ where the output function is given precisely by τ.

Example 3.4.11 With the notation of the previous definition, consider the alphabets $A = \{a, b, c\}$, $B = \{d, e\}$ and the morphisms

$$\sigma : A \to A^+, \begin{cases} a & \mapsto & abc \\ b & \mapsto & bc \\ c & \mapsto & aac \end{cases} \quad \text{and} \quad \tau : A \to B, \begin{cases} a & \mapsto & d \\ b, c & \mapsto & e \end{cases}.$$

The corresponding automaton $\mathcal{A}_{\sigma,a,\tau}$ is given in Figure 3.8 and the output function is represented on the outgoing arrows.

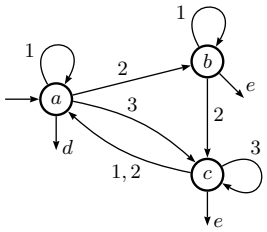

Fig. 3.8. The automaton $\mathcal{A}_{\sigma,a,\tau}$.

Proposition 3.4.12 *Let $\sigma : A \to A^*$ be a morphism prolongable on the letter $a \in A$ and $\tau : A \to B^*$ be a morphism such that $x = \tau(\sigma^\omega(a))$ is infinite. There exists an abstract numeration system $\mathcal{S}$ such that $x \in B^{\mathbb{N}}$ is an $\mathcal{S}$-automatic sequence.*

Proof Thanks to Lemma 3.4.5, we can assume that σ is non-erasing and that τ is a coding. Let $C = \{1, \ldots, \|\sigma\|\} \subset \mathbb{N}$ and consider the automaton $\mathcal{A}_{\sigma,a} = (A, C, E, \{a\}, T)$ as given in Definition 3.4.10 where $T = A$, *i.e.*, all states are final.

Let $L \subseteq C^*$ be the language recognised by $\mathcal{A}_{\sigma,a}$. This language will be used to build an abstract numeration system $\mathcal{S}$ to show that x is $\mathcal{S}$-automatic. The alphabet C being a subset of $\mathbb{N}$, we will consider the natural ordering of C. Since $\sigma(a) \in aA^+$, it is clear that if $w \in L$ then $1w \in L$. Indeed by definition of $\mathcal{A}_{\sigma,a}$, its initial state a has a loop labelled by 1, the first letter in C. If we apply Lemma 3.4.6 to this automaton $\mathcal{A}_{\sigma,a}$, we obtain a morphism $\psi_{\mathcal{A}_{\sigma,a}}$ generating the sequence of the states reached by the words of L. This morphism is defined as follows. Let $z \notin A$. We have $\psi_{\mathcal{A}_{\sigma,a}}(z) = za$ and, for all $b \in A$, $\psi_{\mathcal{A}_{\sigma,a}}(b) = \sigma(b)$.

The main point leading to the conclusion is to compare $\psi^\omega_{\mathcal{A}_{\sigma,a}}(z)$ and

$\sigma^\omega(a)$. There exists $u \in A^+$ such that $\sigma(a) = au$. We have the following factorisations

$$\sigma^\omega(a) = au\,\sigma(u)\,\sigma^2(u)\,\sigma^3(u)\cdots$$

and

$$\psi^\omega_{\mathcal{A}_{\sigma,a}}(z) = za\,a\,u\,\sigma(a)\,\sigma(u)\,\sigma^2(a)\,\sigma^2(u)\,\sigma^3(a)\,\sigma^3(u)\cdots\ .$$

If we erase the factors z, a, $\sigma(a)$, $\sigma^2(a), \ldots$ occurring in that order in the above factorisation of $\psi^\omega_{\mathcal{A}_{\sigma,a}}(z)$, we recover $\sigma^\omega(a)$. Recall that $\psi^\omega_{\mathcal{A}_{\sigma,a}}(z)$ is, except for z, the sequence of states reached in $\mathcal{A}_{\sigma,a}$ by considering all the possible paths in genealogical order. The second occurrence of a in $\psi^\omega_{\mathcal{A}_{\sigma,a}}(z)$ is the state reached in $\mathcal{A}_{\sigma,a}$ when reading $1 \in L$. By the property (3.11) of $\psi_{\mathcal{A}_{\sigma,a}}$, the factor $\sigma^n(a)$ in the above factorisation corresponds to the states reached in $\mathcal{A}_{\sigma,a}$ when reading the words in L of length $n+1$ starting with 1. Consequently, when giving to $\mathcal{A}_{\sigma,a}$ the words of $L \setminus 1C^*$ in increasing genealogical order, we build exactly the sequence $\sigma^\omega(a) = (y_n)_{n\geq 0}$, *i.e.*, if $w_0 \prec w_1 \prec w_2 \prec \cdots$ are the words of $L \setminus 1C^*$ in genealogical order, then $y_n = \delta(a, w_n)$ where δ is the transition function of $\mathcal{A}_{\sigma,a}$. To conclude, one has to consider the automaton $\mathcal{A}_{\sigma,a,\tau}$ as a DFAO with the ANS $\mathcal{S}$ built over $L \setminus 1C^*$ to see that the sequence $\tau(\sigma^\omega(a))$ is $\mathcal{S}$-automatic. □

Example 3.4.13 We illustrate the previous proof by considering the morphisms of Example 3.4.11 and the automaton $\mathcal{A}_{\sigma,a,\tau}$ given in Figure 3.8. We thus have a morphism $\psi_{\mathcal{A}_{\sigma,a}}$

$$\psi_{\mathcal{A}_{\sigma,a}} : A \cup \{z\} \to (A \cup \{z\})^+ : \begin{cases} z & \mapsto & za \\ a & \mapsto & \sigma(a) = abc \\ b & \mapsto & \sigma(b) = bc \\ c & \mapsto & \sigma(c) = aac\ . \end{cases}$$

Let $u = bc$ and $\sigma(a) = au$. If we underline the factors z, a, $\sigma(a)$, $\sigma^2(a), \ldots$ we have

$$\psi^\omega_{\mathcal{A}_{\sigma,a}}(z) = \underline{z}\underline{a}\underline{abc}\underline{abc}bcaac\underline{abcbcaac}bcaacabcabcaac\cdots\ .$$

Erasing the underlined factors, we get $abcbcaacbcaacabcabcaac\cdots$ which is exactly $\sigma^\omega(a)$.

The statement of the next result explicitly introduces the language that was built in the proof of Proposition 3.4.12. We can say that the language $L \setminus 1C^*$ is the *directive language* of σ: if the letters in $\sigma^\omega(a)$ are indexed by the words in $L \setminus 1C^*$, then we know precisely which letter is producing which factor through the morphism.

Corollary 3.4.14 *Let $\sigma : A \to A^*$ be a non-erasing morphism prolongable on the letter $a \in A$ such that $x = (x_n)_{n\geq 0} = \sigma^\omega(a)$ is infinite. Consider the ANS $\mathcal{S}$ built over $L \setminus 1C^*$ where $C = \{1, \ldots, \max_{b\in A} |\sigma(b)|\}$ and L is the language accepted by $\mathcal{A}_{\sigma,a}$. Let $w \in L$ be such that $|\sigma(x_{\mathrm{val}_\mathcal{S}(w)})| = \ell$. Then*

$$\sigma(x_{\mathrm{val}_\mathcal{S}(w)}) = x_{\mathrm{val}_\mathcal{S}(w1)} \cdots x_{\mathrm{val}_\mathcal{S}(w\ell)} \ .$$

In the above formula, for $i \in \{1, \ldots, \ell\}$, wi has to be understood as the concatenation of $w \in L \subseteq C^*$ and $i \in C$.

Proof It is a direct consequence of the proofs of Lemma 3.4.6 and Proposition 3.4.12.

An independent proof is the following one. We even get another proof that $x_n = \delta_\sigma(a, \mathrm{rep}_S(n))$ where δ_σ is the partial transition function of $\mathcal{A}_{\sigma,a}$.

Consider the adjacency matrix $\mathbf{M} \in \mathbb{N}^{A\times A}$ of $\mathcal{A}_{\sigma,a}$, see Section 1.4 for the definition. For all $s > 0$ and $b, c \in A$, $[\mathbf{M}^s]_{b,c}$ is the number of paths of length s from b to c in $\mathcal{A}_{\sigma,a}$. Since all states of this latter automaton are final, the number N_s of words of length s accepted by $\mathcal{A}_{\sigma,a}$ is obtained by summing up all the entries of $\mathbf{M}^s$ in the row corresponding to the initial state a. Because $\mathcal{A}_{\sigma,a}$ has a loop of label 1 in a, the number of words of length s accepted by $\mathcal{A}_{\sigma,a}$ and starting with 1 is equal to the number N_{s-1} of words of length $s-1$ accepted by $\mathcal{A}_{\sigma,a}$. Consequently, the number of words of length s in the language $L \setminus 1C^*$ is exactly $N_s - N_{s-1}$. From the definition of $\mathcal{A}_{\sigma,a}$, the matrix $\mathbf{M}$ can also be related to the morphism σ and $\mathbf{M}_{b,c}$ is the number of occurrences of c in $\sigma(b)$. Summing up all entries in the row of $\mathbf{M}^s$ corresponding to a gives $|\sigma^s(a)|$. Therefore, the number of words of length s in $L \setminus 1C^*$ is $|\sigma^s(a)| - |\sigma^{s-1}(a)|$ and we get that

$$|\mathrm{rep}_S(n)| = s \Leftrightarrow n \in \{|\sigma^{s-1}(a)|, \ldots, |\sigma^s(a)| - 1\}. \tag{3.12}$$

In particular, if $0 < n < |\sigma(a)|$, we have $|\mathrm{rep}_S(n)| = 1$ and in this case† $\mathrm{rep}_S(n) = n + 1$. Since we have $\mathrm{rep}_S(0) = \varepsilon$ and $\sigma(a) = au$, for some $u \in \Sigma^*$, we get $x_0 = a = \delta_\sigma(a, \mathrm{rep}_S(0))$. Hence, by the definition of $\mathcal{A}_{\sigma,a}$, we have that $x_n = \delta_\sigma(a, \mathrm{rep}_S(n))$ for $n < |\sigma(a)|$. Now let $s > 0$ and assume that $x_n = \delta_\sigma(a, \mathrm{rep}_S(n))$ for all $n < |\sigma^s(a)|$. Let $|\sigma^s(a)| \leq n < |\sigma^{s+1}(a)|$. There exists a unique $|\sigma^{s-1}(a)| \leq m < |\sigma^s(a)|$ such that

$$\sigma^{s+1}(a) = \underbrace{\sigma^{s-1}(a)\, u\, x_m\, v}_{\sigma^s(a)}\ \sigma(u)\ \underbrace{y\, x_n\, z}_{\sigma(x_m)}\sigma(v),$$

for some words u, v, y, z. Therefore $x_n = (\sigma(x_m))_{i-1}$ for some $i \in$

† In order to have $\mathrm{rep}_S(n) = n$, one could work instead with the alphabet $C' = \{0, \ldots, \max_{b\in A} |\sigma(b)| - 1\}$.

$\{1, \ldots, |\sigma(x_m)|\}$. Then by the definition of $\mathcal{A}_{\sigma,a}$, we have

$$x_n = \delta_\sigma(x_m, i) = \delta_\sigma(\delta_\sigma(a, \mathrm{rep}_S(m)), i) = \delta_\sigma(a, \mathrm{rep}_S(m)i)$$

and in view of condition (3.12) and again by the definition of $\mathcal{A}_{\sigma,a}$, we get

$$\mathrm{val}_S(\mathrm{rep}_S(m)i) = |\sigma^s(a)| + |\sigma(x_{|\sigma^{s-1}(a)|})| + \cdots + |\sigma(x_{m-1})| + i = n.$$

Hence, $\mathrm{rep}_S(n) = \mathrm{rep}_S(m)i$ and the result follows. □

Example 3.4.15 The infinite word generated by the morphism σ given in Example 3.4.11 is $(x_n)_{n\geq 0} = abc\underline{b}caac\underline{bc}aacabcabcaac\cdots$. The first few words without leading 1 accepted by the automaton given in Figure 3.8 where all states are final are ε, 2, 3, 21, 22, 31, 32, 33, 211, ... This provides us with an ANS $\mathcal{S}$.

For instance, we consider the element $x_3 = b$. This is why it has been underlined. We know that $\sigma(b) = bc$. So this latter factor should appear later on in the infinite word and the previous corollary permits us to find where it occurs. The $\mathcal{S}$-representation of 3 is 21. So we have to consider the words 211 and 212 — only these two words because $|\sigma(b)| = 2$ — and $\mathrm{val}_\mathcal{S}(211) = 8$, $\mathrm{val}_\mathcal{S}(212) = 9$. Therefore, one can check that $x_8x_9 = \sigma(b) = bc$.

Now we turn to the converse of Proposition 3.4.12.

Proposition 3.4.16 *Every $\mathcal{S}$-automatic sequence is substitutive.*

Proof Let $\mathcal{S} = (L, A, <)$ be an ANS. Let $\mathcal{A} = (Q, A, E, \{q_0\}, T)$ be a complete DFA with transition function $\delta_\mathcal{A} : Q \times A^* \to Q$ recognising L and $\mathcal{B} = (R, A, \delta_\mathcal{B}, \{r_0\}, B, \mu)$ be a DFAO generating an $\mathcal{S}$-automatic sequence $x = (x_n)_{n\geq 0}$ over B, *i.e.*, for all $n \geq 0$, $x_n = \mu(\delta_\mathcal{B}(r_0, \mathrm{rep}_\mathcal{S}(n)))$.

Consider the Cartesian product automaton $\mathcal{P} = \mathcal{A} \times \mathcal{B}$ defined as follows. The set of states of $\mathcal{P}$ is $Q \times R$. The initial state is (q_0, r_0) and the alphabet is A. For any word $w \in A^*$, the transition function $\Delta : (Q\times R)\times A^* \to Q\times R$ is given by

$$\Delta((q, r), w) = (\delta_\mathcal{A}(q, w), \delta_\mathcal{B}(r, w)) \ .$$

This means that the product automaton mimics in a single automaton, the behaviours of both $\mathcal{A}$ and $\mathcal{B}$. In particular, after reading w in $\mathcal{P}$, $\Delta((q_0, r_0), w)$ belongs to $F \times R$ if, and only if, w belongs to L. Moreover, if $\mathrm{rep}_\mathcal{S}(n) = w$ and $\Delta((q_0, r_0), w) = (q, r)$, then $x_n = \mu(r)$.

Now we can apply Lemma 3.4.6 to $\mathcal{P}$ and define a morphism $\psi_\mathcal{P}$ prolongable on a letter z which does not belong to $Q \times R$. In view of the

previous paragraph, we define $\nu : (Q \times R) \cup \{z\} \to B^*$ by

$$\nu(q,r) = \begin{cases} \mu(r), & \text{if } q \in F, \\ \varepsilon, & \text{otherwise} \end{cases}$$

and $\nu(z) = \varepsilon$. As Lemma 3.4.6 can be used to describe the sequence of reached states, $\nu(\psi_{\mathcal{P}}(z))$ is exactly the sequence $(x_n)_{n\geq 0}$. □

Note that the morphisms obtained at the end of this proof are erasing. Again, if needed, Lemma 3.4.5 can be used.

3.4.1 Some properties of $\mathcal{S}$-automatic sequences

Here we give a characterisation of $\mathcal{S}$-automatic sequences in terms of finiteness of its $\mathcal{S}$-kernel and then study the relationship between $\mathcal{S}$-automaticity and $\mathcal{S}$-recognisability. We conclude this subsection with some discussion about the theorem of Cobham.

Definition 3.4.17 Let $\mathcal{S} = (L, A, <)$ be an ANS. For each word w in A^*, we define a, possibly finite or empty, ordered set of integers:

$$\mathcal{I}_{\mathcal{S}}(w) = \{n \mid \mathrm{rep}_{\mathcal{S}}(n) \in A^*w\} = \mathrm{val}_{\mathcal{S}}(L \cap A^*w) = \{i_{w,0} < i_{w,1} < \cdots\} .$$

In particular if s is a suffix of w, then $\mathcal{I}_{\mathcal{S}}(w) \subseteq \mathcal{I}_{\mathcal{S}}(s)$. Also we define a partial function $\alpha_{\mathcal{S}}$ mapping $(w,n) \in A^* \times \mathbb{N}$ onto $i_{w,n}$. For all $w \in A^*$, defining the ANS $\mathcal{S}_w = (L \cap A^*w, A, <)$, we get

$$\alpha_{\mathcal{S}}(w,n) = \mathrm{val}_{\mathcal{S}}(\mathrm{rep}_{\mathcal{S}_w}(n)), \text{ whenever defined .}$$

Example 3.4.18 Consider the language L accepted by the DFA depicted in Figure 3.2 from Example 3.2.1. Since the first few words in L are

$\mathrm{rep}_{\mathcal{S}}(n)$	b	aa	abb	bab	bba	$aaab$	$aaba$	$abaa$	$baaa$	$bbbb$	$\ldots$
n	0	1	2	3	4	5	6	7	8	9	$\ldots$

we get $\mathcal{I}_{\mathcal{S}}(\varepsilon) = \mathbb{N}$, $\mathcal{I}_{\mathcal{S}}(a) = \{1,4,6,7,8,\ldots\}$, $\mathcal{I}_{\mathcal{S}}(b) = \{0,2,3,5,9,\ldots\}$, $\mathcal{I}_{\mathcal{S}}(aa) = \{1,7,8,\ldots\}$, *etc.* So $\alpha_{\mathcal{S}}(\varepsilon,n) = n$ for all $n \geq 0$, $\alpha_{\mathcal{S}}(a,0) = 1$, $\alpha_{\mathcal{S}}(a,1) = 4$, $\alpha_{\mathcal{S}}(a,2) = 6$, $\alpha_{\mathcal{S}}(b,0) = 0$, $\alpha_{\mathcal{S}}(b,1) = 2$, $\alpha_{\mathcal{S}}(b,2) = 3$, *etc.*

Recall that, for $k \geq 2$, the *k-kernel* (also see Definition 9.1.1 where it is used to define the concept of automatic sequence, as recalled below) of a sequence $(x_n)_{n\geq 0}$ is the set of subsequences defined as

$$\{(x_{k^j n+r})_{n\geq 0} \mid j \geq 0, 0 \leq r < j\} .$$

Otherwise stated, we consider all the subsequences obtained by taking all the indices that are congruent modulo a power of k. It is well-known that a

sequence is k-automatic if, and only if, its k-kernel is finite, see for instance (Allouche and Shallit 2003). With the definition of the map $\alpha_{\mathcal{S}}$ introduced above, but by writing α_k when dealing with the usual k-ary numeration system on $B_k = \{0,\ldots,k-1\}^* \setminus 0\{0,\ldots,k-1\}^*$, the usual k-kernel of a sequence $(x_n)_{n\geq 0}$ can be rewritten as

$$\{(x_{\alpha_k(w,n)})_{n\geq 0} \mid w \in \{0,\ldots,k-1\}^*\}$$

because a word u over $\{0,\ldots,k-1\}$ ends with a suffix w of length j if, and only if, $\mathrm{val}_k(u) \mod k^j = \mathrm{val}_k(w)$.

Definition 3.4.19 Let $\mathcal{S} = (L, A, <)$ be an ANS. The *$\mathcal{S}$-kernel* of the sequence $(x_n)_{n\geq 0}$ is the set of subsequences $\{(x_{\alpha_{\mathcal{S}}(w,n)})_{n\geq 0} \mid w \in A^*\}$.

Theorem 3.4.20 (Rigo and Maes 2002) *A sequence $x = (x_n)_{n\geq 0} \in A^{\mathbb{N}}$ is $\mathcal{S}$-automatic if, and only if, its $\mathcal{S}$-kernel is finite.*

The proof is similar to the classical one.

Remark 3.4.21 It is obvious that a set of integers is $\mathcal{S}$-recognisable if, and only if, its characteristic word is $\mathcal{S}$-automatic. Therefore a sequence $x = (x_n)_{n\geq 0} \in A^{\mathbb{N}}$ is $\mathcal{S}$-automatic if, and only if, for all $a \in A$, the *a-fiber*, *i.e.*, the set $\{n \mid x_n = a\}$, is $\mathcal{S}$-automatic.

Properties of the complexity function p_w counting the number of factors of a substitutive word w are well-known, see Chapter 4. These facts can be taken into account to show that some sets are $\mathcal{S}$-recognisable for *no* ANS $\mathcal{S}$. Take the *Champernowne word* over $\{0,1\}$ $c = 011011100101110111000\cdots$ obtained as the ordered juxtaposition of the binary representations of the integers. It is the characteristic word of a set of integers $\{1,2,4,5,6,\ldots\}$ which is never $\mathcal{S}$-recognisable because the complexity function of c is $p_c(n) = 2^n$ which is not an admissible behaviour for a substitutive word. Also it can be shown that the set of primes is never $\mathcal{S}$-recognisable (Mauduit 1988), (Mauduit 1992), (Rigo 2000).

Remark 3.4.22 It is not difficult to prove that the characteristic sequence of the set of squares can be generated using the morphism $\sigma : a \mapsto abcd$, $b \mapsto b$, $c \mapsto cdd$, $d \mapsto d$ iterated on a and a coding $\tau : a, b \mapsto 1$, $c, d \mapsto 0$ (also see the morphism and the coding given in Example 1.2.23). We can compare this result with Example 3.3.8 and observe that the construction developed in Section 3.3.1 can also be presented in the context of substitutive words. In particular, one can notice that the same kind of results have been obtained independently in (Carton and Thomas 2002) where some decidability of the logical structure $\langle \mathbb{N}, + \rangle$ extended with a substitutive predicate is sought.

It is time to come back to the Cobham theorem (Theorem 1.5.5) and its generalisation to ANS. Indeed, now we hope that thanks to Theorem 3.4.1 the reader is convinced that both formalisms of substitution or ANS are well suited to define and study a relevant notion of recognisable sets of integers. Let x, y be infinite fixed points of two morphisms μ, ν: $\mu(x) = x$, $\nu(y) = y$ and α, β be two codings. Roughly speaking, we would like to have a result of the kind: if μ and ν are "independent" in a sense to be defined and if $\alpha(x) = \beta(y)$, then the word $\alpha(x)$ is eventually periodic. Following G. Hansel's work about syndeticity (Hansel 1982, Hansel 1998), F. Durand has made a lot of progress in that direction. For instance, if μ and ν are primitive and if the corresponding dominating eigenvalues are multiplicatively independent then the theorem of Cobham still holds, see (Durand 1998a). Later on more cases can be taken into account, see (Durand 1998c), (Durand 2002) and also (Durand and Rigo 2009) where the situation of two ANS, one defined on a polynomial language and the other on an exponential one, is considered. To obtain full generality, only a few cases remain unsolved.

Remark 3.4.23 Up to now there is no proof of a Cobham-like theorem for a substitution having no main sub-substitution having the same dominating eigenvalue like $a \mapsto aa0$, $0 \mapsto 01$ and $1 \mapsto 0$. In this latter example, the dominating eigenvalue is 2 but the substitution restricted to $\{0, 1\}$ has $(1 + \sqrt{5})/2$ as dominating eigenvalue.

3.4.2 The HD0L ω-equivalence and periodicity problems

We recall some definitions about the so-called *Lindenmayer systems*. For an account of these systems we refer to (Kari, Rozenberg, and Salomaa 1997), also see Section 10.1. A *D0L system* is a triple $G = (A, \sigma, u)$ where A is a finite alphabet, u is a word over A and $\sigma : A^* \to A^*$ is a morphism, the acronym D stands for "deterministic" and 0 stands for "zero-sided". An *HD0L system*, where H stands for "homomorphism", is a 5-tuple $G = (A, B, \sigma, \tau, u)$ where (A, σ, u) is a D0L system, B is a finite alphabet and $\tau : A^* \to B^*$ is a morphism. If u is a prefix of $\sigma(u)$ and the set $\{\sigma^n(u) \mid n \geq 0\}$ is infinite, we denote $\sigma^\omega(u) = \lim_{n\to\infty} \sigma^n(u)$. Similarly, if $G = (A, B, \sigma, \tau, u)$ is an HD0L system, u is prefix of $\sigma(u)$ and the set $\{\tau(\sigma^n(u)) \mid n \geq 0\}$ is infinite, we denote $\omega(G) = \lim_{n\to\infty} \tau(\sigma^n(u))$. The *HD0L ω-equivalence problem* is stated as follows. Let $G_i = (A_i, B_i, \sigma_i, \tau_i, u_i)$, $i = 1, 2$, be two HD0L systems such that $\omega(G_1)$ and $\omega(G_2)$ exist. If $\omega(G_1) = \omega(G_2)$, then the two HD0L systems G_1 and G_2 are said to be *ω-equivalent*. Is it possible to decide whether or not G_1 and G_2 are ω-equivalent? In fact, HD0L systems are closely related to substitutive words.

Lemma 3.4.24 *Let $G_1 = (A, B, \mu, \nu, w)$ be an HD0L system such that $\omega(G_1)$ exists and $|w| > 1$. Then there exists an HD0L system $G_2 = (C, B, \sigma, \tau, c)$ ω-equivalent to G_1 where the letter $c \in C$ is prefix of $\sigma(c)$.*

Proof Assume that $\mu(w) = wu$ for some $u \in A^+$ and $w = w_1 \cdots w_\ell$, $\ell \geq 2$, with $w_i \in A$ for $1 \leq i \leq \ell$. We have $\mu^n(w) = w\, u\, \mu(u) \cdots \mu^{n-1}(u)$ for all $n \geq 1$. Let us introduce $\ell + 1$ new letters $c, \overline{w}_1, \ldots, \overline{w}_\ell$ which do not belong to A. The alphabet C is defined by $C = A \cup \{c, \overline{w}_1, \ldots, \overline{w}_\ell\}$. The morphism $\sigma : C^* \to C^*$ is defined as follows, $\sigma : c \mapsto c\overline{w}_1$, $\overline{w}_1 \mapsto \overline{w}_2, \ldots$, $\overline{w}_{\ell-1} \mapsto \overline{w}_\ell$, $\overline{w}_\ell \mapsto u$ and for $a \in A$, $\sigma(a) = \mu(a)$. We get

$$\lim_{n\to\infty} \sigma^n(c) = c\overline{w}_1 \cdots \overline{w}_\ell\, u\, \mu(u)\, \mu^2(u)\, \mu^3(u) \ldots \ .$$

To conclude the proof, we define τ by $\tau(c) = \varepsilon$, $\tau(\overline{w}_i) = \nu(w_i)$ for $1 \leq i \leq \ell$ and $\tau(a) = \nu(a)$ for $a \in A$. It is obvious that $\tau(\sigma^\omega(c)) = \nu(\mu^\omega(w))$. □

Remark 3.4.25 With the above lemma and Theorem 3.4.1, many classical open decision problems about HD0L systems can be restated in the framework of ANS. See in particular Chapter 10. The HD0L ω-equivalence problem is equivalent to the following problem expressed in terms of ANS. Let $\mathcal{S}_i = (L_i, A_i, <_i)$, $i = 1, 2$, be two abstract numeration systems. Is it decidable, given regular languages $K_i \subseteq L_i$, $i = 1, 2$, whether or not $\mathrm{val}_{\mathcal{S}_1}(K_1) = \mathrm{val}_{\mathcal{S}_2}(K_2)$? In the same way, the problem of deciding whether or not a given infinite HD0L word $\omega(G)$ is eventually periodic is equivalent to the following problem. Let $\mathcal{S} = (L, A, <)$ be an ANS. *Is it decidable, given a regular language $K \subseteq L$, whether or not* $\mathrm{val}_{\mathcal{S}}(K)$ *is eventually periodic?*

3.4.3 Multidimensional setting

If one goes to the multidimensional case, it is not difficult to mimic as follows the construction of (Salon 1987), where images of letters are finite multidimensional words with square or cube shapes of same dimension. Let $d \geq 2$, $\mathcal{S} = (L, A, <)$ and $\#$ be a symbol not in A. The idea to define an $\mathcal{S}$-automatic d-dimensional sequence $\mathbf{x} = (x_{i_1,\ldots,i_d})_{i_1,\ldots,i_d \geq 0}$ over an alphabet B (*i.e.*, a map from $\mathbb{N}^d$ onto B) is to consider a DFAO $\mathcal{B} = (Q, (A \cup \{\#\})^d, \delta_{\mathcal{B}}, \{q_0\}, B, \mu)$ over the alphabet $(A \cup \{\#\})^d$ and to define

$$x_{i_1,\ldots,i_d} = \mu(\delta_{\mathcal{B}}(q_0, (\mathrm{rep}_{\mathcal{S}}(i_1), \ldots, \mathrm{rep}_{\mathcal{S}}(i_d))^{\#})) \ .$$

The padding operator $\#$ has been given in Definition 3.3.19.

One can therefore ask if Theorem 3.4.1 can be extended to this setting.

We mention the following result without much details about it, merely some informal description is given. Also see (Charlier 2009).

Theorem 3.4.26 (Charlier, Kärki, and Rigo 2010) *Let $d \geq 1$. The d-dimensional infinite word $\mathbf{x} = (x_{i_1,\ldots,i_d})_{i_1,\ldots,i_d \geq 0}$ is $\mathcal{S}$-automatic for some abstract numeration system $\mathcal{S} = (L, A, <)$ where $\varepsilon \in L$ if, and only if, $\mathbf{x}$ is the image under a coding of a morphic shape-symmetric infinite d-dimensional word.*

Observe that the proof of Proposition 3.4.12 makes use at the very beginning of Lemma 3.4.5. So one of the main difficulties occurring in the proof of the above theorem is that Lemma 3.4.5 has to be generalised to a multi-dimensional setting. This is some technical business that we do not want to present here. Nevertheless, we briefly describe using an example what is the idea of the shape-symmetry introduced in (Maes 1999). Indeed, this notion can be defined with plenty details and glory indices but a glimpse should be enough to have a good idea of the result above, also see (Maes 1998), (Maes 2000).

Example 3.4.27 Consider a map μ defined on the alphabet $\{a, \ldots, h\}$ and whose images are finite rectangular arrays. We can use the same formalism as in the definition of words and also define the concatenation of words in any of the two directions provided that they have compatible shapes. For instance, $\mu(a)$ and $\mu(b)$ can be concatenated horizontally but not vertically.

$$\mu(a) = \mu(f) = \begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array},\ \mu(b) = \begin{array}{|c|}\hline e \\ \hline c \\ \hline \end{array},\ \mu(c) = \begin{array}{|c|c|}\hline e & b \\ \hline \end{array}, \mu(d) = \begin{array}{|c|}\hline f \\ \hline \end{array},$$

$$\mu(e) = \begin{array}{|c|c|}\hline e & b \\ \hline g & d \\ \hline \end{array},\ \mu(g) = \begin{array}{|c|c|}\hline h & b \\ \hline \end{array},\ \mu(h) = \begin{array}{|c|c|}\hline h & b \\ \hline c & d \\ \hline \end{array}.$$

In Figure 3.9 we have represented the first iterations of μ on the letter a. As for prolongable morphisms, one can expect that this process will lead to some bidimensional fixed point $(x_{i,j})_{i,j \geq 0}$. The shape-symmetry mainly refers to the fact that, for all i, j, if the image by μ of $x_{i,j}$ is a rectangle of size $\ell \times m$, then the image by μ of $x_{j,i}$ is a rectangle of size $m \times \ell$. Also one must ensure that the images of all the letters in a given column (respectively row) have images which are rectangles with same length (respectively height). An equivalent formulation is that the image by μ of any diagonal element $x_{i,i}$ is a square. Some details are omitted, see (Maes 1999) for a complete description.

$\mu(a) =$ $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array}$, $\mu^2(a) =$ $\begin{array}{|c|c|c|}\hline a & b & e \\ \hline c & d & c \\ \hline e & b & f \\ \hline \end{array}$, $\mu^3(a) =$ $\begin{array}{|c|c|c|c|c|}\hline a & b & e & e & b \\ \hline c & d & c & g & d \\ \hline e & b & f & e & b \\ \hline e & b & e & a & b \\ \hline g & d & c & c & d \\ \hline \end{array}$

Fig. 3.9. The first few iterations of μ.

3.5 Representing real numbers

The basic aim of this section is to introduce and summarise the material found in (Lecomte and Rigo 2002), (Lecomte and Rigo 2004). The concern is to extend the use of an ANS $\mathcal{S} = (L, A, <)$ to represent real numbers and to study the properties of the proposed extension.

Roughly, the problem is reduced to the representation of numbers belonging to some subinterval of $[0, 1]$. The idea is to associate with a real number x an infinite word $w \in A^\omega$ that plays a role similar to its decimal expansion: w will be the limit of a sequence of words $w^{(n)}$ in L used to produce more and more accurate rational approximations $x^{(n)}$ of x that eventually converge to it. Let us explain how they mimic the decimal system or more generally, β-numeration systems, see Chapter 2. The rational approximations provided by the decimal expansion $.d_1d_2\cdots d_\ell\cdots$ of a real number in $(1/10, 1)$ are $\frac{d_1}{10}, \frac{d_1d_2}{100}, \ldots, \frac{d_1\cdots d_\ell}{10^\ell}, \ldots$. They all take the form of a fraction whose numerator is a prefix of some length ℓ of the expansion and the denominator is the number of integers whose decimal representation has *length at most* ℓ. With that scheme in mind, for a sequence $(w^{(n)})_{n\geq 0}$ of words in L converging to an infinite word w, we set

$$x^{(n)} = \frac{\mathrm{val}_{\mathcal{S}}(w^{(n)})}{\mathcal{V}_L(|w^{(n)}|)} \ . \tag{3.13}$$

Under some suitable assumptions, $(x^{(n)})_{n\geq 0}$ is a converging numerical sequence and its limit x belongs to some interval canonically associated with the language L. The prefixes $w^{(n)}$ of the representation w can be used to approximate x. In the sequel, we assume that $\mathcal{A}$ is the minimal automaton of the regular language L.

3.5.1 Extending $\mathrm{val}_{\mathcal{S}}$

The set of the representations of the real numbers that we will describe is

$$\mathrm{Adh}(L) := \{w \in A^\omega \mid \exists (w^{(n)})_{n\geq 0} \in L^{\mathbb{N}},\ \lim_{n\to\infty} w^{(n)} = w\} \ .$$

This notion of adherence appears in (Nivat 1978) and is studied in (Boasson and Nivat 1980).

Proposition 3.5.1 *The set Adh(L) is uncountable if, and only if, there exist two cycles $\mathcal{C}$ and $\mathcal{C}'$ in any DFA accepting L such that $\mathcal{C} \cap \mathcal{C}' \neq \emptyset$ and $\mathcal{C} \cup \mathcal{C}'$ contains an accessible state and a co-accessible state.*

In the sequel, $\mathrm{Adh}(L)$ is obviously supposed to be uncountable. Also we make some additional assumptions in order that the sequences (3.13) converge when w belongs to $\mathrm{Adh}(L)$.

Hypothesis 3.5.2 For each $q \in Q$, either

(i) there exists $N_q \in \mathbb{N}$ such that $\mathcal{U}_q(n) = 0$, for all $n > N_q$, or
(ii) there exist $\theta_q \geq 1$, $P_q(x) \in \mathbb{R}[x]$ and $c_q > 0$ such that

$$\lim_{n\to\infty} \frac{\mathcal{U}_q(n)}{P_q(n)\theta_q^n} = c_q \ .$$

Since $\mathrm{Adh}(L)$ is uncountable, it follows from the above proposition that the language L has an exponential growth and therefore that $\theta := \theta_{q_0} > 1$. Replacing P_{q_0} by P_{q_0}/c_{q_0}, we may assume in what follows that

$$\lim_{n\to\infty} \frac{\mathcal{U}_{q_0}(n)}{P_{q_0}(n)\theta^n} = 1 =: a_{q_0} \quad \text{and, for all } q \in Q, \quad a_q := \lim_{n\to\infty} \frac{\mathcal{U}_q(n)}{P_{q_0}(n)\theta^n} \geq 0 \ .$$

Proposition 3.5.3 *Let $\mathcal{S}$ be an ANS based on a language satisfying Hypothesis 3.5.2. If $w = w_0 w_1 \cdots \in Adh(L)$ is the limit of a sequence $(w^{(n)})_{n\geq 0}$ of words in L, then*

$$x := \lim_{n\to\infty} \frac{\mathrm{val}_{\mathcal{S}}(w^{(n)})}{\mathcal{V}_L(|w^{(n)}|)} = \frac{\theta - 1}{\theta^2} \sum_{q\in Q} a_q \sum_{j=0}^{\infty} b_{q,j}\theta^{-j}$$

with the coefficients $b_{q,j}$ defined in (3.5)*. In particular, x is independent of the sequence in $L^{\mathbb{N}}$ converging to w. Moreover, it belongs to $[1/\theta, 1]$. Conversely every element in $[1/\theta, 1]$ is the limit of a sequence of the form* (3.13) *for some $w \in Adh(L)$.*

In this latter proposition, x is said to be the *numerical value* $\mathrm{val}_{\mathcal{S}}(w)$ of w. In the same way, the infinite word w is said to be an *$\mathcal{S}$-representation* of the real number x.

Proposition 3.5.4 *Let $\mathcal{S}$ be an ANS based on a language satisfying Hypothesis 3.5.2. The map $\mathrm{val}_{\mathcal{S}} : Adh(L) \to [1/\theta, 1]$ is increasing and uniformly continuous.*

Some elements in $[1/\theta, 1]$ may have more than one $\mathcal{S}$-representation in $\mathrm{Adh}(L)$ and possibly infinitely many. This problem will be discussed in the next subsection.

Recall that each Pisot number β defines a unique positional and linear Bertrand numeration system $U_\beta = (U_n)_{n\in\mathbb{N}}$. See (Bruyère and Hansel 1997) and recall Example 3.1.14. One can show (Frougny and Solomyak 1996) that the language L_β of all the normalised representations computed by the greedy algorithm satisfies Hypothesis 3.5.2, with $\theta = \beta$. In (Lecomte and Rigo 2004), also it is shown that the $\mathcal{S}$-representations of the elements of $[1/\beta, 1]$ in the ANS based upon L_β and the classical β-developments of these numbers coincide. In particular, $\mathrm{Adh}(L_\beta)$ is the set of these developments.

Example 3.5.5 Consider the classical Fibonacci system (for integers) or the β-numeration system related to the Golden Ratio φ. The language of all the representations of integers not starting with 0 is accepted by the DFA depicted in Figure 3.10. Let us consider the ANS based on this language and show that we get back the usual φ-development. If we set $\lambda = \frac{5+\sqrt{5}}{10}$,

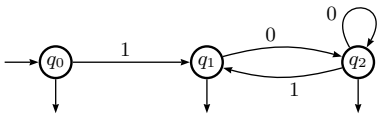

Fig. 3.10. A Fibonacci ANS.

then an easy computation shows that $\mathcal{U}_{q_0}(n) \sim \lambda\,\varphi^{n-1}$, $\mathcal{U}_{q_1}(n) \sim \lambda\,\varphi^n$ and $\mathcal{U}_{q_2}(n) \sim \lambda\,\varphi^{n+1}$. Setting P_{q_0} to the constant λ/φ and dividing $\mathcal{U}_q(n)$ by $P_{q_0}\,\varphi^n$, we get $a_{q_0} = 1$, $a_{q_1} = \varphi$ and $a_{q_2} = \varphi^2$. For any infinite word $w_0 w_1 \cdots \in \mathrm{Adh}(L)$, the formula of Proposition 3.5.3 becomes

$$\frac{\varphi-1}{\varphi^2}\sum_{j=0}^{\infty}\varphi^{-j} + (\varphi-1)\sum_{j=0}^{\infty} b_{q_2,j}\varphi^{-j} = \varphi^{-1} + \sum_{j\geq 2} w_j\,\varphi^{-j-1}$$

because for all $j \geq 0$ we have $b_{q_0,j} = 1$, $b_{q_1,j} = 0$, $b_{q_2,0} = 0$ and for $j > 0$ $b_{q_2,j} = w_j$. To obtain the last equality we used the fact that $\varphi - 1 = \varphi^{-1}$.

3.5.2 The intervals I_u

Let us have a closer look at the approximations of the elements in $[1/\theta, 1]$ by finite words. Let u be a word of length ℓ and denotes by I_u the set of real numbers $x \in [1/\theta, 1]$ having an $\mathcal{S}$-representation starting with u. If I_u is non-empty, *i.e.*, if u is a prefix of some element in $\mathrm{Adh}(L)$, then it is a

closed interval. In particular, $I_\varepsilon = [1/\theta, 1]$. Moreover, if u is a prefix of v, then $I_u \supset I_v$ and if $w = w_0 w_1 \cdots \in \mathrm{Adh}(L)$ is an $\mathcal{S}$-representation of x, then x belongs to $I_{w_0 \cdots w_{\ell-1}}$ for all ℓ.

The set $\mathcal{I}_\ell$ of non-empty intervals I_u such that $|u| = \ell$ defines a covering of $[1/\theta, 1]$ made of closed subintervals with disjoint interiors. Otherwise stated, there exist $k(\ell)$ and real numbers $\kappa_1^\ell = 1/\theta \leq \cdots \leq \kappa_{k(\ell)+1}^\ell = 1$ such that

$$\mathcal{I}_\ell = \left\{ [\kappa_j^\ell, \kappa_{j+1}^\ell] \mid j = 1, ..., k(\ell) \right\} .$$

Each $\kappa_j^\ell, 1 < j \leq k(\ell)$, has at least two representations, as it is the upper bound of some I_u and the lower bound of some I_v. It may well occur that $\mathrm{Adh}(L)$ contains infinitely many words having prefix u although I_u contains exactly one element, *i.e.*, $\kappa_j^\ell = \kappa_{j+1}^\ell$ for some j. This one has then infinitely many representations. Obviously, vanishing constants a_p are of no use to compute $\mathrm{val}_\mathcal{S}(w)$ and this causes that phenomenon, see Proposition 3.5.3. It follows easily from Hypothesis 3.5.2 that if $a_q = 0$ and $p = q.u$ for some word u, then $a_p = 0$. Thanks to Proposition 3.5.3, we may delete from $\mathcal{A}$ the states q such that $a_q = 0$ and the corresponding edges without changing the representations of real numbers, up to the fact that we replace $\mathcal{A}$ and L by the simplified automaton and its language. Now, a real number in $(1/\theta, 1)$ has exactly one representation if it is not the endpoint of some I_u and exactly two representations otherwise.

Proposition 3.5.6 *Let $\mathcal{S}$ be an ANS based on a language satisfying Hypothesis 3.5.2 and let $\mathbf{M} \in \mathbb{N}^{Q \times Q}$ be the adjacency matrix of $\mathcal{A}$. For all states p, we have*

$$a_p = \frac{1}{\theta} \sum_{q \in Q} \mathbf{M}_{pq} a_q .$$

If u is a prefix of length ℓ of some element of $Adh(L)$, then

$$I_u = \left[\frac{1}{\theta} + \frac{\theta - 1}{\theta^{\ell+1}} \sum_{\substack{|t|=\ell \\ t<u}} a_{q_0.t}, \frac{1}{\theta} + \frac{\theta - 1}{\theta^{\ell+1}} \sum_{\substack{|t|=\ell \\ t \leq u}} a_{q_0.t} \right] .$$

In particular, θ and the numbers κ_j^ℓ are algebraic.

3.5.3 A dynamical point of view

In order to understand the structure of the set of intervals I_u, it is useful, say, to normalise them in some way. This will allow us to design an algorithm to compute representations of real numbers in $[1/\theta, 1]$ and to study the set

of these which have an eventually periodic expansion. To that end, for any interval $I = [s, t]$, $s < t$, we use the increasing bijection

$$f_I : I \to [0, 1],\ x \mapsto \frac{x - s}{t - s}$$

and we say that $f_I(x)$ is the *relative position* of $x \in I$ (inside I). More generally, the relative position of a subset E of I inside I will be $f_I(E)$.

Proposition 3.5.7 *Let u and v be two words such that $q_0.u = q_0.v$. For each $a \in A$, $I_{ua} \neq \emptyset$ if, and only if, $I_{va} \neq \emptyset$. Moreover, if $I_{ua} \neq \emptyset$, then the relative positions of I_{ua} inside I_u is equal to that of I_{va} inside I_v.*

The above proposition is a key point in our study as it tells that the interval I_u only depends upon the states $q_0.u$, and not specifically upon u. We use it to construct a dynamical system $(Q \times [0, 1], T)$ which encodes the relationship between the sets $\mathcal{I}_\ell$.

Let q be any given state of $\mathcal{A}$. As the latter is accessible, $q = q_0.u$ for some u of length say ℓ. Then

$$I_u = [\kappa_i^{\ell+1}, \kappa_j^{\ell+1}]$$

for some $i < j$ and we get a partition

$$[0, 1] = [\kappa'_i, \kappa'_{i+1}) \cup \cdots \cup [\kappa'_r, \kappa'_{r+1}) \cup \cdots \cup [\kappa'_{j-1}, \kappa'_j]$$

where, for simplicity, κ'_r denotes the relative position of $\kappa_r^{\ell+1}$ inside I_u. Of course, the elements of that partition are nothing but the non-empty intervals among the I_{ua}, $a \in A$. We let $R_{q,a}$ denote the relative position of such an I_{ua} inside I_u and we define the function T by

$$T : Q \times [0, 1] \to Q \times [0, 1],\ (q, x) \mapsto (q.a, f_{R_{q,a}}(x))$$

where a is the unique letter such that $x \in R_{q,a}$.

The following algorithm 3.3 computes prefixes of the representation of a real number $x \in [1/\theta, 1]$ by applying iteratively T to the initial data (q_0, x). The length ℓ of the prefixes is determined by some halting condition.

Example 3.5.8 Let us continue Example 3.5.5. Clearly, the state q_0 occurs only once and a representation always starts with 1. From q_1, one can only reach q_2 reading 0. So a discussion has to be made only for state q_2. The interval $I_{10} =$ splits into I_{100} and I_{101}. The relative position of $\varphi^{-1} + \varphi^{-3}$ inside $[\varphi^{-1}, 1]$ is φ^{-1}. So we get the partition $[0, \varphi^{-1}) \cup [\varphi^{-1}, 1]$. A scheme of application of Algorithm 3.3 is given in Figure 3.11.

```
INPUT : x ∈ [1/θ, 1]
q ← q_0
u ← ε
I ← I_ε
y ← f_I(x)
REPEAT
   DETERMINE a ∈ A such that y ∈ R_{q,a}
   q ← q.a
   u ← CONCATENATE(u, a)
   I ← R_{q,a}
   y ← f_I(x)
UNTIL |u| = ℓ
```

Table 3.3. *An algorithm computing a prefix of length ℓ of an $\mathcal{S}$-representation of the real x.*

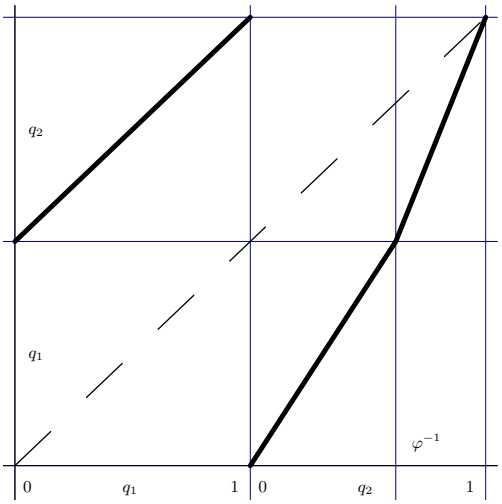

Fig. 3.11. The map T restricted to $\{q_1, q_2\} \times [0, 1]$.

3.5.4 Real numbers with eventually periodic representations

In the k-ary numeration system, the set of real numbers having an eventually periodic representations is $\mathbb{Q}$. In particular, it is dense in $\mathbb{R}$ and has strong algebraic properties (it is a field). It is of course tempting to investigate the properties of the set of eventually periodic words belonging to $\mathrm{Adh}(L)$. This is a hard problem and only a few results are known up to now in that direction.

We begin with some interesting facts. For any set of (infinite) words X, we let $\mathrm{eper}(X)$ denote the set of eventually periodic elements in X, by $\mathrm{per}(X)$ the set of their periods and by $\mathrm{preper}(X)$ the set of their preperiods,

i.e., if uv^ω belongs to $\mathrm{eper}(X)$, then v belongs to $\mathrm{per}(X)$ and u belongs to $\mathrm{preper}(X)$.

Proposition 3.5.9 *The sets per(Adh(L)) and preper(Adh(L)) are regular. Moreover, eper(Adh(L)) is dense in Adh(L).*

The real numbers having an eventually periodic representation can be characterised in terms of T.

Theorem 3.5.10 *A number $x \in [1/\theta, 1]$ has an eventually periodic representation if, and only if, there exists $r < s$ such that*

$$T^r(q_0, x) = T^s(q_0, x).$$

In particular, each number κ_i^ℓ has an eventually periodic representation.

Let us explain why κ_i^ℓ has an eventually periodic representation. There is a word m such that $I_m = [\kappa_i^\ell, \kappa_{i+1}^\ell]$. The representation of κ_i^ℓ given by Algorithm 3.3 starts with m. In other words, $T^\ell(q_0, \kappa_i^\ell) = (q_0.m, 0)$. But then, clearly, for $r > \ell$, $T^r(q_0, \kappa_i^\ell)$ is of the form $(q, 0)$ for some q. As there are finitely many states, it follows that $T^r(q_0, \kappa_i^\ell) = T^s(q_0, \kappa_i^\ell)$ for some $r < s$.

As for the algebraic properties of the set of numbers having an eventually periodic representation, we have the following result similar to the one given independently in (Bertrand 1977), (Schmidt 1980a) (also see Theorem 2.3.20 and Section 2.3.2.1).

Theorem 3.5.11 (Rigo and Steiner 2005) *Let $\mathcal{S}$ be an ANS based on a language satisfying Hypothesis 3.5.2. If the corresponding real number θ is a Pisot number, then*

$$\mathrm{val}_{\mathcal{S}}(\mathit{eper}(\mathit{Adh}(L))) = \mathbb{Q}(\theta) \cap [1/\theta, 1]$$

but if θ is neither a Pisot number nor a Salem number, then

$$\mathbb{Q}(\theta) \cap [1/\theta, 1] \not\subseteq \mathrm{val}_{\mathcal{S}}(\mathit{eper}(\mathit{Adh}(L))) \ .$$

3.6 Exercises and open problems

Exercise 3.1 (Charlier 2009) Consider the sequence $U = (U_n)_{n \geq 0}$ given by $U_i = i + 1$ for $i = 0, 1, 2, 3$ and $U_n = 2U_{n-1}$ for all $n \geq 4$. Show that $\mathbb{N}$ is U-recognisable. Show that for all $k \geq 1$, there exist no $a_{k-1}, \ldots, a_0 \in \mathbb{C}$ with $a_0 \neq 0$ such that, for all $n \geq 0$,

$$U_{n+k} = a_{k-1}U_{n+k-1} + \cdots + a_0 U_n.$$

In the terminology of (Berstel and Reutenauer 1988), U does not satisfy any *strict* linear recurrence relation. Hint: $a_0 \neq 0$ is invertible.

Exercise 3.2 Give a proof of (3.4) given in page 134 using Lemma 3.2.2.

Exercise 3.3 Consider the two abstract numeration systems based on a^*b^* obtained by changing the ordering on the alphabet, $\mathcal{S} = (a^*b^*, \{a, b\}, a < b)$ and $\mathcal{R} = (a^*b^*, \{a, b\}, b < a)$. Study the function $f_{\mathcal{S},\mathcal{R}} : \mathbb{N}^2 \to \mathbb{N}^2, (i, j) \mapsto (x, y)$ such that $\mathrm{rep}_{\mathcal{R}}(\mathrm{val}_{\mathcal{S}}(a^i b^j)) = a^x b^y$.

Exercise 3.4 Show that, in general, changing the ordering of the alphabet is not a recognisability-preserving operation. A counter-example is given in (Lecomte and Rigo 2001).

Exercise 3.5 Find a closed formula for the expression of $\mathrm{val}_{\mathcal{S}}(a^i b^j c^k)$ for the ANS $\mathcal{S} = (a^*b^*c^*, \{a, b, c\}, a < b < c\}$. In this system which set of integers is represented respectively by a^*, b^* and c^*?

Exercise 3.6 Generalise ANS on a^*b^* or $a^*b^*c^*$ by considering the ANS $\mathcal{S} = (a_1^* \cdots a_t^*, \{a_1, \ldots, a_t\}, a_1 < \cdots < a_t)$ where $a_1, \ldots, a_t$ are t distinct letters. Show that this system is equivalent to the so-called *binomial numeration system* defined as follows, see (Fraenkel 1985). Any integer $n \geq 0$ can be uniquely written as

$$n = \binom{z_t}{t} + \binom{z_{t-1}}{t-1} + \cdots + \binom{z_1}{1}$$

with $z_t > z_{t-1} > \cdots > z_1 \geq 0$. Indeed, show that we have

$$\mathrm{val}_{\mathcal{S}}(a_1^{n_1} \cdots a_t^{n_t}) = \sum_{i=1}^{t} \binom{n_i + \cdots + n_t + t - i}{t - i + 1} .$$

For details see (Charlier, Rigo, and Steiner 2008). Also see connection with (Lew, Morales, and Sánchez-Flores 1996).

Exercise 3.7 Consider the ANS given in Example 3.1.15. This system seems to be related to the Fibonacci numeration system. Is it possible to assign weights $v(a)$ and $v(b)$ to a and b to recover the usual Fibonacci system, *i.e.*, such that, for all $w_\ell \cdots w_0 \in L$, $\mathrm{val}_{\mathcal{S}}(w) = \sum_{k=0}^{\ell} v(w_i)\, F_i$?

Exercise 3.8 Let $\mathcal{S} = (a^*b^*, \{a, b\}, a < b\}$. Show that the formal series $\sum_{w \in L} \mathrm{val}_{\mathcal{S}}(w)\, w$ is rational in the sense of (Berstel and Reutenauer 1988) (Also see the definition given in Section 2.6.1). In particular, we get the

linear representation (λ, μ, γ) where $\mu : \{a, b\}^* \to \mathbb{N}^{3\times 3}$ is a morphism of monoids defined by

$$\mu(a) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \ \mu(b) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

$\lambda = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$ and $\gamma = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$ such that $\mathrm{val}_{\mathcal{S}}(w) = \lambda\, \mu(w)\, {}^t\gamma$. This result holds for any ANS, see (Rigo 2002) and independently (Choffrut and Goldwurm 1995) where $\mathrm{val}_{\mathcal{S}}$ is called *ranking*.

Exercise 3.9 Consider the ANS $\mathcal{S} = (\{a, b\}^* \setminus a^*, \{a, b\}, a < b)$ and the set Y such that $\mathrm{rep}_{\mathcal{S}}(Y) = a^*b$. Is the set $\mathrm{val}_{\mathcal{S}}(2\,\mathrm{rep}_{\mathcal{S}}(Y)) = Z$ still recognisable? We suggest to write a small computer program to list the first 100 elements in $\mathrm{rep}_{\mathcal{S}}(Z)$.

Exercise 3.10 Let $L \subset A^*$ be a cofinite language not equal to A^*. Study the preservation of $\mathcal{S}$-recognisability after multiplication by a constant for ANS based on L. Notice that if $L = A^*$, then ANS defined on L is equivalent to the usual integer base $(\mathrm{Card}\, A)$-ary system.

Exercise 3.11 (Open problem) For the usual k-ary numeration system, a logical characterisation of the k-recognisable sets by first order logical formula from $\langle \mathbb{N}, +, V_k \rangle$ is well known. Could one imagine a logical characterisation of the $\mathcal{S}$-recognisable sets in a suitable logical structure?

Exercise 3.12 (Open problem) Assume that $P \in \mathbb{Q}[X]$ is such that $P(\mathbb{N}) \subseteq \mathbb{N}$. If $P(\mathbb{N})$ is $\mathcal{S}$-recognisable for $\mathcal{S} = (L, A, <)$, what information on L can be obtained? For instance, is L polynomial?

Exercise 3.13 (Open problem) Let $\mathcal{S}$ be an ANS. Find necessary and/or sufficient conditions for the existence of an increasing sequence $(U_n)_{n\geq 0}$ of integers such that $U_0 = 1$ and a map $v : A \to \mathbb{N}$ such that $\mathrm{val}_{\mathcal{S}}(w) = \sum_{i=0}^{\ell} v(w_i)\, U_i$ for all $w = w_\ell \cdots w_0 \in L$.

Exercise 3.14 (Open problem) This problem is also discussed in the bibliographic notes and in Remark 3.4.25. Does there exist an algorithm, given an ANS $\mathcal{S} = (L, A, <)$ and any $\mathcal{S}$-recognisable set X of integers given by a DFA, which can be used to decide whether or not X is eventually periodic?

Exercise 3.15 (Open problem) Obtain a general Cobham-like theorem for ANS. See in particular, Remark 3.4.23.

3.7 Notes

Properties of k-recognisable sets are well-known. For a survey, see for instance (Bruyère, Hansel, Michaux, et al. 1994). This paper explains in particular the logical characterisation of the k-recognisable sets in terms of first order logical formulas in an extension of the Presburger arithmetic $\langle \mathbb{N}, + \rangle$ with a valuation V_k. Most of the characterisations encountered for k-ary systems can be extended to the Pisot systems of Example 3.1.14. See (Bruyère and Hansel 1997) which partially relies on (Frougny 1992) about the *normalisation* function computable by a finite automaton. Also it is probably worth to have a look at (Shallit 1994) which has been cited many times in this chapter.

Slender languages have been considered in several contexts and particularly in some decision problems. See (Honkala 1997), (Honkala 1998), (Honkala 2001b). For a study of morphisms and/or languages with polynomial growth, also see (Mauduit 1986).

It is interesting to note that Lemma 3.3.5 about the regularity of the set of minimal words of each length in a regular language has been extended to context-free languages. If L is context-free, then $\mathsf{minlg}\,(L)$ is again context-free (Berstel and Boasson 1997). It is also shown that if $\mathrm{Pref}(\mathrm{Adh}(L)) = L$, then $\mathsf{minlg}\,(L)$ is regular.

State complexity issues about *decimations* treated in Theorem 3.3.2 are considered in (Krieger, Miller, Rampersad, et al. 2009). In this paper, the authors moreover provide an example of an ANS $\mathcal{G}$ based on a non-regular context-free language such that $\mathrm{rep}_{\mathcal{G}}(2\mathbb{N})$ is not context-free. In (Berstel, Boasson, Carton, et al. 2006), some operations preserving regular languages are discussed.

The idea of Definition 3.4.10 associating with a morphism some canonical automaton already appears in the seminal paper (Cobham 1972). Of course, when considering a uniform morphism, the resulting automaton is complete.

More on extension of $\mathcal{S}$-automaticity to the multidimensional case can be found in (Rigo and Maes 2002) and (Nicolay and Rigo 2007). Following the work of A. Fraenkel, applications of ANS and in particular the use of Corollary 3.4.14, to combinatorial game theory appear in (Duchêne and Rigo 2008a) and (Duchêne and Rigo 2008b). See (Duchêne, Fraenkel, Nowakowski, et al. 2009) for an application of shape-symmetric bidimensional morphisms to Wythoff's game. In this paper, it is proved that the set of losing positions defines a shape-symmetric morphic array.

Consider the following decision problem. Let $\mathcal{S} = (L, A, <)$ be an ANS. For any $\mathcal{S}$-recognisable set of integers given by a DFA, decide

whether or not this set is eventually periodic. As explained in Remark 3.4.25 this problem can also be stated in terms of HD0L systems. The purely substitutive case is settled positively in (Harju and Linna 1986), (Pansiot 1986). For k-automatic sequences, the problem is solved in (Honkala 1986). See (Leroux 2005) where a polynomial time general procedure for d-dimensional subsets of $\langle \mathbb{Z}, + \rangle$ is given. An elegant and simple approach based on the construction of an NFA can be found in (Allouche, Rampersad, and Shallit 2009). In (Honkala and Rigo 2004), some special cases expressed in terms of ANS are treated. Recently, (Bell, Charlier, Fraenkel, et al. 2009) covers a large class of ANS for which the problem is decidable, also see (Charlier 2009). The general problem is still open.

Several other topics related to ANS and in particular to the representation of real numbers are the following ones. For most of the situations described below some extra assumptions on the language are often required like having a DFA with a dominating eigenvalue. The introduction in the sense of (Grabner, Liardet, and Tichy 1995) of the *odometer* for ANS is made in (Berthé and Rigo 2007b). The idea is to define in a proper way a map sending an infinite word onto its "successor", also see Section 6.5. As an example, for positional numeration systems like the Fibonacci numeration system, this map sends $010100(10)^\omega$ onto $0000(10)^\omega$ and one has to study carry propagation. The definition of the odometer for ANS acts on a pair made of an infinite word and the infinite sequence of states corresponding to the path followed in the automaton when reading this word. The idea is to replace a non-maximal prefix of length ℓ read from some state q by the next word of same length ℓ accepted from q. Continuing the Fibonacci example, the successor of 1010 is 10000 which explains what we get above by taking mirror images. Note that considering pairs of letters and states is equivalent to considering local automata. Some *tilings* given in the framework of ANS have been presented in (Berthé and Rigo 2007a), see connection with Chapters 2 and 5. An analogue to the classical *sum-of-digits* function (see Chapter 9) can be defined as follows. Consider a map $f : A \to \mathbb{R}$ to define a completely *additive function*, *i.e.*, for all $w = w_1 \cdots w_\ell \in A^*$, $f(w) = \sum_{i=1}^{\ell} f(w_i)$. The behaviour and distribution of the corresponding summatory function $\sum_{w \in L} f(w)$ is studied in (Grabner and Rigo 2003) and (Grabner and Rigo 2007). Extensions of β-expansions and of the map $T_\beta : x \mapsto \{\beta x\}$, see Section 2.3.2, is presented in (Rigo and Steiner 2005). As in Lemma 3.2.2, it involves for ANS as many maps as states in the minimal automaton of L.

The framework of Section 3.5 extends to a larger class a numeration systems in (Charlier, Le Gonidec, and Rigo) including numeration systems

based on a non-regular language such as the one coming from rational base numeration systems (see Section 2.5).