The signature of rational languages

Victor Marsault[a], Jacques Sakarovitch[b]

[a] *IRIF, Université Denis Diderot, 8 place Aurélie Nemours, 75013 Paris, France.*
[b] *LTCI, CNRS / Telecom-ParisTech, 46 rue Barrault, 75013 Paris, France.*

Abstract

We present here the notion of *signature* of trees and of languages, and its relationships with the theory of numeration systems. The signature of an ordered infinite tree (of bounded degree) is an infinite (bounded) sequence of integers, the sequence of the degrees of the nodes taken in the visit order of *the canonical* breadth-first traversal of the tree. A prefix-closed language defines such a tree augmented with labels on arcs, hence is associated with a signature. This way of 'traversing' a language is related to the notion of abstract numeration system, due to Lecomte and Rigo.

After having set in detail the framework of signature, we study and characterise the signatures of rational languages. Using a known construction from numeration system theory, we show that these signatures form a special subclass of morphic words. We then use this framework to give an alternative definition to morphic numeration systems (also called Dumont-Thomas numeration systems). We finally highlight that the classes of morphic numeration systems and of (prefix-closed) rational abstract numeration systems are essentially the same.

1. Introduction

This work introduces the notion of *breadth-first signature* of a tree, or of a language. It consists of an infinite word describing the tree (or the language). Depending on the direction from the tree to the word, or conversely, it is either a *serialisation* of the tree into an infinite word or a *generation* of the tree by the word. Here, we study and characterise the serialisation of rational, or regular, languages.

The breadth-first signature or, for short, the *signature* of an *ordered* tree of finite degree is the sequence of the degrees of the nodes visited by a breadth-first traversal of the tree. Since the tree is ordered, there is a *canonical* breadth-first traversal; hence the signature is defined by the tree. Conversely, and under a validity condition, a signature characterises a tree.

*Corresponding Author
Email address: Jacques.Sakarovitch@telecom-paristech.fr (Jacques Sakarovitch)

Similarly, the *labelling* of a labelled tree is the infinite sequence of the labels of the arcs visited by the breadth-first traversal of this tree. The pair signature/labelling is once again characteristic of the labelled tree. It provides an effective serialisation of labelled trees, hence of prefix-closed languages.

This serialisation of a prefix-closed language over an ordered alphabet is very close, and in some sense, equivalent to the enumeration of the words of the language in the radix order. It makes then this notion particularly fit to describe the languages of integer representations in various numeration systems. It is of course the case for the representations in an integer base p which corresponds to the signature p^ω, the constant sequence. But it is also the case for non-standard numeration systems such as the Fibonacci numeration system, whose representation language has for signature the Fibonacci word. It is also the case for the *rational base numeration systems*, as defined in [1], and whose representation languages have periodic signatures, that is, signatures that are infinite periodic words. To tell the truth, it is the latter case that first motivated our study of signatures and a subsequent work is devoted to the characterisation of the trees and languages generated by periodic signatures [2].

In this work, we first show that the signatures of prefix-closed rational languages all belong to a special subclass of morphic infinite words that we call *s-morphic signatures*. An s-morphic signature is a morphic word where the projection (or 'coding') morphism is entirely determined by the prolongable morphism itself; more precisely it is a word of the form $f_\sigma(\sigma^\omega(a))$ where f_σ is the morphism that maps every letter b to the length of $\sigma(b)$ (considered as a digit). Conversely, we prove that every s-morphic signature, paired with an appropriate *morphic labelling*, generates a prefix-closed rational language. The proof of these results relies on a correspondence between morphic words and automata due to Maes and Rigo [3] or Dumont and Thomas [4, 5] and whose principle goes back to the work of Cobham [6].

The fact that the description of a language by its signature is, in some sense, equivalent to the enumeration of the words of the language in the radix order makes the notion of signature remarkably close to the one of Abstract Numeration System (ANS for short) as proposed by Lecomte and Rigo [7]. An ANS is defined by an arbitrary language L over an ordered alphabet; this language is ordered by the radix order and the representation of the integer n is by definition the $(n+1)$-th word of L, independently of any evaluation function as it is the case in 'concrete' numeration systems. An ANS L is said to be *rational* if L is a rational language. Note that Lecomte and Rigo usually consider rational ANSs only and call them simply ANSs; the systems that we call ANSs are referred to as *generalised numeration system* in the few cases when they consider them (such as in [8]).

We then use our framework to give an alternative definition of Morphic Numeration System (MNS) originally introduced by Dumont and Thomas in [9], hence often called *Dumont-Thomas numeration systems* in the literature.

By our definition, a MNS is a prefix-closed rational ANS of a special form and is canonically associated with an s-morphic (unlabelled) signature. We

then show that any given rational prefix-closed ANS L may be converted easily to the MNS K associated with the s-morphic signature of L: the conversion function $L \to K$, that maps the representation of every integer n in L to the representation of n in K, is realised by a letter-to-letter pure-sequential transducer. This conversion transducer is moreover graph-isomorphic to the automata accepting L and K; note that this second automaton is the so-called *prefix automaton*. These considerations result in the idea that prefix-closed rational ANSs and MNSs have the same expressive power.

Section 2 describes the correspondence between signatures, trees and prefix-closed languages. Section 3 defines s-morphic signature and establishes the characterisation theorem. Section 4 gives the definition of Dumont–Thomas numeration systems and shows their central position. A very preliminary version of this work, covering part of the content of sections 2 and 3 only, has appeared in [10]. Most of the results are also part of the PhD thesis of the first author [11].

Acknowledgements. The authors would like to thank the referees of the submitted version of [10] for their constructive comments. They are particularly grateful to Michel Rigo, who has pointed to several references among which the works of Dumont–Thomas, for his precious advices and friendly encouragements.

2. Signatures of trees and languages

We describe here a process of *serialisation* of (infinite) trees, (infinite) labelled trees, and (infinite) prefix-closed languages, that is, the representation of such objects by one, or two, (infinite) words, using the assumption of the existence of an underlying order. We also recall the related concept of abstract numeration system and introduce the one of padded language and, for the rational case, of padded finite automata.

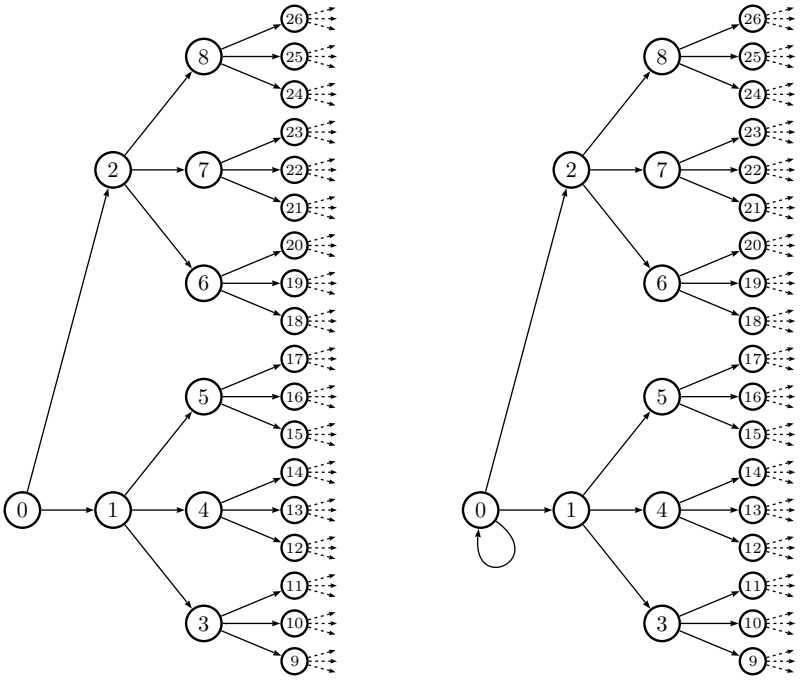
2.1. On trees

Classically, a tree is an undirected graph in which any two vertices are connected by exactly one path (*cf.* [12], for instance). Our point of view differs in two respects. First, a tree is a *directed* graph such that (i) there exists a *unique* vertex, called *root*, which has no incoming arc, and (ii) there is a *unique (oriented) path* from the root to every other vertex. Second, our trees are *ordered*, that is, the set of children of every node is totally ordered.

In the figures, we draw trees with the root on the left, arcs rightwards and the child order will be implicitly defined by the convention that children placed higher are greater (according to this order).

It will prove to be convenient to have a slightly different look at trees and to consider that the root of a tree is also a *child of itself*, that is, bears a loop

onto itself.[1] We call such a structure an *i-tree*. It is so close to a tree that we pass from one to the other with no further ado. Nevertheless, some definitions or results are easier or more straightforward when stated for i-trees, and others when stated for trees: it is then handy to have both available. A tree will usually be denoted by $\mathcal{T}_x$ for some index x and the associated i-tree by $\mathcal{I}_x$. Figure 1 shows such a pair of a tree and the associated i-tree.



(a) A tree that is almost ternary

(b) The ternary i-tree

Figure 1: The tree and i-tree associated with the base 3 numeration system

The degree of a node is the number of its children. In the sequel, we consider infinite (i-)trees of finite degree, that is, all nodes of which have finite degree. (We consider indeed infinite (i-)trees of *bounded* degree, but this restriction does not matter for the definitions to come.) The breadth-first traversal of such an ordered (i-)tree defines a total ordering of its nodes.

Convention. *The set of nodes of an (i-)tree is always the set* $\mathbb{N}$ *of the non-negative integers.*

With this convention, the root is 0 and n is the $(n+1)$-th node visited by the

[1] This convention is sometimes taken when implementing tree-like structures (for instance in the Unix/Linux file system).

traversal. For n, m in $\mathbb{N}$, we write

$$n \xrightarrow[\mathcal{T}]{} m$$

whenever m is a child of n in $\mathcal{T}$. We denote by $\mathsf{d}(n)$ the degree of the node n.

2.2. Signatures of trees

We call *signature* any infinite sequence $\boldsymbol{s} = s_0\, s_1\, s_2 \cdots$ of non-negative integers. Whenever the signature $\boldsymbol{s}$ is obvious from the context, we simply denote by S_j, for every integer j, the partial sum of the j first letters of $\boldsymbol{s}$:

$$\forall j \in \mathbb{N} \qquad S_j \;=\; \sum_{i=0}^{j-1} s_i \quad ,$$

that is, $S_0 = 0\,,\ S_1 = s_0$ and more generally $S_j = S_{j-1} + s_{j-1}$ for every $j > 0$.

Definition 1. *A signature $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ is* valid *if the following holds:*

$$\forall j \in \mathbb{N} \qquad S_{j+1} \;>\; j+1 \;\; . \tag{1}$$

In particular, the validity of $\boldsymbol{s}$ implies that s_0 is greater than, or equal to, 2.

Definition 2.

(i) *The* breadth-first signature *or, for short, the* signature*, of an i-tree $\mathcal{I}$ is the sequence $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ of the degrees of the nodes of the* i-tree *$\mathcal{I}$:*

$$\forall i \in \mathbb{N} \qquad s_i \;=\; \mathsf{d}(i) \;\; .$$

(ii) *The breadth-first signature of a tree $\mathcal{T}$ is the signature of* the corresponding i-tree.

Figure 1 shows both the tree and the i-tree the signature of which is 3^ω. Valid signatures are in bijection with infinite i-trees of finite degree, as expressed by the next proposition.

Proposition 3.

(i) *Let $\boldsymbol{s} = s_0\, s_1\, s_1 \cdots$ be a valid signature. There exists a unique i-tree $\mathcal{I}_{\boldsymbol{s}}$ whose signature is $\boldsymbol{s}$: the i-tree such that every node n has s_n children, the s_n nodes of the interval $\{S_n,\, S_n+1,\, \ldots,\, S_{n+1}-1\}$.*

(ii) *The signature of any infinite (i-)tree of finite degree is valid.*

Proof. The proof of (i) takes essentially the form of a procedure that generates an i-tree from a valid signature $\boldsymbol{s} = s_0 s_1 s_2 \cdots$. It maintains two integers: the node n to be processed and the number m of nodes created so far, both initially set to 0. At step $(n+1)$ of the procedure, s_n nodes are created, namely the nodes $m, m+1, \ldots, (m + s_n - 1)$, and s_n edges are created, all with starting point n, and one for each of these new nodes as end point. Then n is incremented by 1, and m by s_n.

This procedure indeed describes an i-tree. The first node created is 0 and the first arc created is the loop $0 \longrightarrow 0$ on the root. It is verified by induction that at every step, m is equal to S_n. The initial conditions ($n = m = 0$) indeed satisfy this equality since S_0 is an empty sum.

The validity of $\boldsymbol{s}$ ensures that at the end of every step of the procedure $n < m$ holds (but not at the beginning of the first step where $n = m = 0$). It follows that every node is strictly larger than its father, but for the root, whose father is itself.

(ii) Let $\mathcal{I}$ be an infinite i-tree and $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ its signature; S_n is the number of children of the first n nodes of $\mathcal{I}$. If $\boldsymbol{s}$ is not valid, the smallest integer j for which Equation (1) does not hold is such that $S_j = j$, in which case the set of the children of the j first nodes is of cardinal j, hence $\mathcal{I}$ has j nodes and is finite. □

Figure 2 shows the first eight steps of the generation process applied to the signature $\boldsymbol{s}_1 = (321)^\omega$. A slightly larger initial part of the resulting infinite i-tree $\mathcal{I}_{\boldsymbol{s}_1}$ together with a labelling is shown in Figure 3.

2.3. Labelled signatures of labelled trees

In our framework, alphabets are totally ordered. In the case of alphabets of digits, the natural order is of course implicitly used. A word $w = a_0 a_1 \cdots a_{k-1}$ is *increasing* if $a_0 < a_1 < \cdots < a_{k-1}$. As usual, the length of a finite word w is denoted by $|w|$.

A labelled tree $\mathcal{T}$, or i-tree $\mathcal{I}$, is an (i-)tree every arc of which holds a label taken in an alphabet A. Since both A and $\mathcal{T}$ (or $\mathcal{I}$) are ordered, the labels on the arcs have to be *consistent* with these two orders: two arcs originating from the same node n must be labelled by two letters whose order is the same as the endpoints of the arcs or, more intuitively, an arc to a greater child is labelled by a greater letter. For n, m in $\mathbb{N}$, and a in A,we write

$$n \xrightarrow[\mathcal{I}]{a} m \tag{2}$$

whenever m is a child of n in $\mathcal{I}$ and the arc from n to m holds the label a.

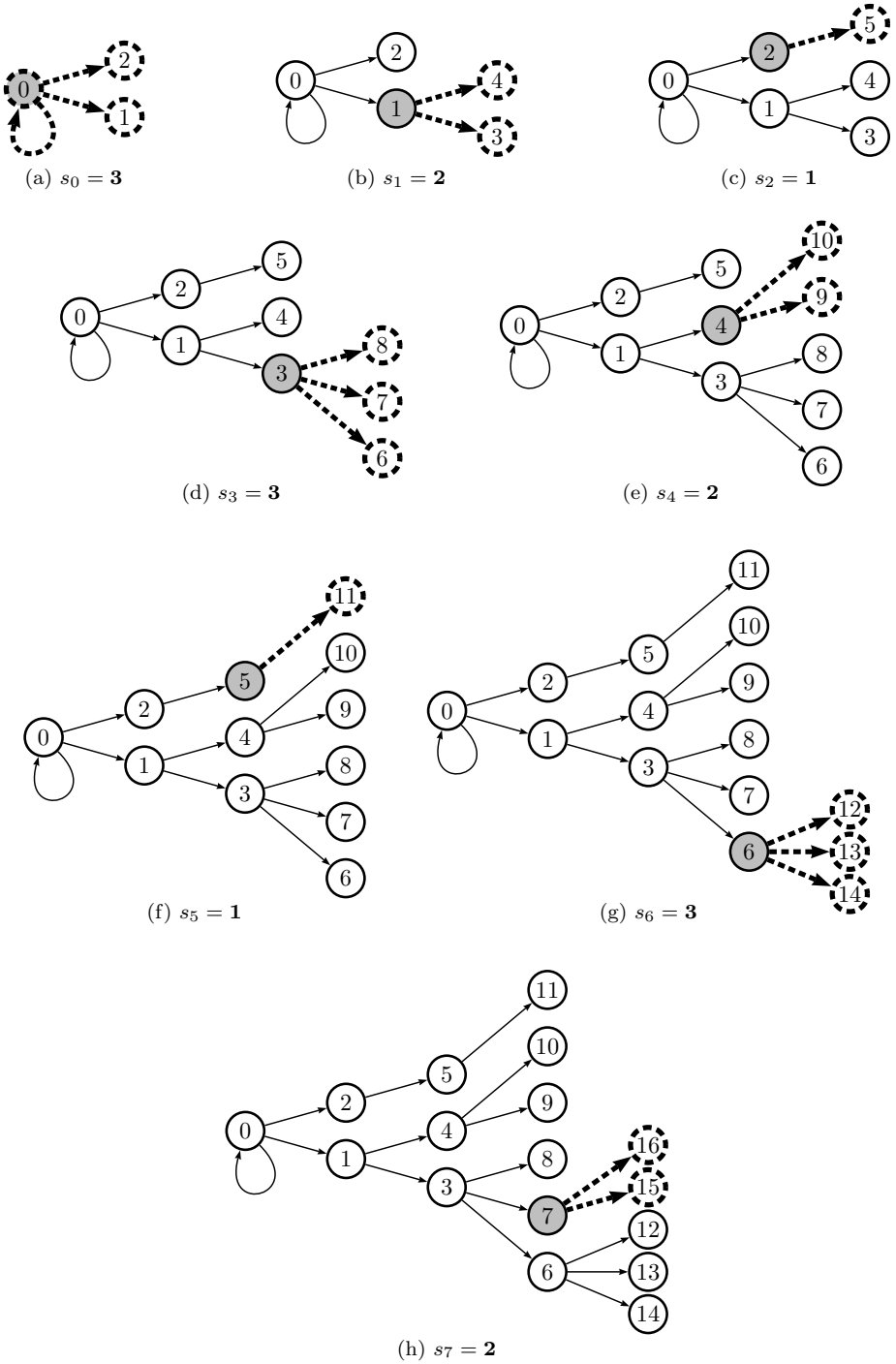
The labelling $\boldsymbol{\lambda} = \lambda_0 \lambda_1 \lambda_2 \cdots$ of a labelled i-tree $\mathcal{I}$ (labelled in A) is an infinite word of A^ω, the sequence of the labels of the arcs of $\mathcal{I}$ visited in a breadth-first traversal:

$$\forall m \in \mathbb{N} \qquad \lambda_m \text{ is the label of the unique arc incoming to the node } m \text{ in } \mathcal{I}\ .$$

It follows that λ_0 is the label of the loop on the root of $\mathcal{I}$.

As it is an infinite sequence of non-negative integers, a signature $\boldsymbol{s}$ naturally *determines a factorisation* of any other infinite word $\boldsymbol{\lambda}$: $\boldsymbol{\lambda} = w_0 w_1 w_2 \cdots$ by the condition that $|w_n| = s_n$ for every n in $\mathbb{N}$ (and thus $w_n = \varepsilon$ if $s_n = 0$).

Definition 4. *Let $\boldsymbol{s}$ be a signature. An infinite word $\boldsymbol{\lambda}$ in A^ω is* consistent *with $\boldsymbol{s}$ if the factorisation $\boldsymbol{\lambda} = w_0 w_1 w_2 \cdots$ determined by $\boldsymbol{s}$ has the property that every w_n is an* increasing word.



(a) $s_0 = \mathbf{3}$ (b) $s_1 = \mathbf{2}$ (c) $s_2 = \mathbf{1}$

(d) $s_3 = \mathbf{3}$ (e) $s_4 = \mathbf{2}$

(f) $s_5 = \mathbf{1}$ (g) $s_6 = \mathbf{3}$

(h) $s_7 = \mathbf{2}$

Figure 2: The first eight steps of the generation of $\mathcal{I}_{(321)^\omega}$

A pair $(\boldsymbol{s}, \boldsymbol{\lambda})$ *of infinite words is a* valid labelled signature *if* $\boldsymbol{s}$ *is a valid signature and if* $\boldsymbol{\lambda}$ *is consistent with* $\boldsymbol{s}$*.*

A simple and formal verification yields the following.

Proposition 5. *A labelled i-tree* $\mathcal{I}$ *uniquely determines a valid labelled signature and conversely any valid labelled signature* $(\boldsymbol{s}, \boldsymbol{\lambda})$ *uniquely determines a labelled i-tree* $\mathcal{I}_{(\boldsymbol{s},\boldsymbol{\lambda})}$ *whose labelled signature is precisely* $(\boldsymbol{s}, \boldsymbol{\lambda})$*.*

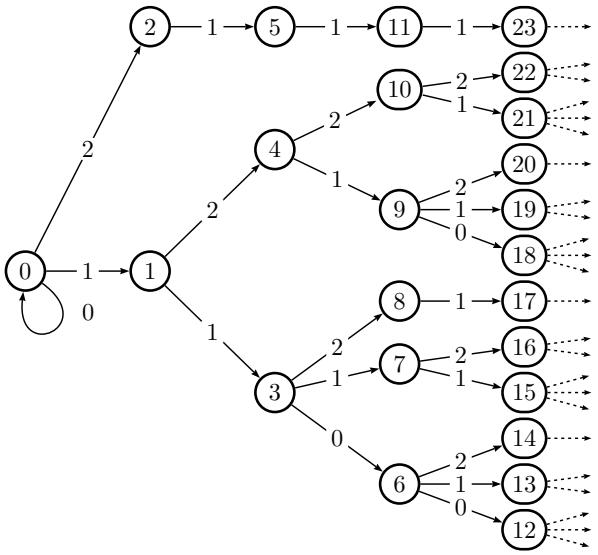


Figure 3: The labelled i-tree $\mathcal{I}_{(\boldsymbol{s}_1,\boldsymbol{\lambda}_1)}$ where $\boldsymbol{s}_1 = (321)^\omega$ and $\boldsymbol{\lambda}_1 = (012.12.1)^\omega$

Figure 3 shows the labelling of the i-tree whose signature is $\boldsymbol{s}_1 = (321)^\omega$ by the infinite periodic[2] word $\boldsymbol{\lambda}_1 = (012.12.1)^\omega$. (This is of course a very special labelling: labellings consistent with $\boldsymbol{s}$ do not need to be periodic, but periodic words are the easiest cases of finitely defined infinite words.)

2.4. Labelled signatures of languages

The *branch language* of a labelled tree is the set of words that label all paths from the root to every node of the tree. It is a *prefix-closed* language. Conversely, every prefix-closed language over a totally ordered alphabet uniquely defines a labelled ordered tree.

The branch language of a labelled *i-tree* is a language of a special form that we call *padded*. The most common example of a padded language is given by the writings of the integers in (an integer) base p. The representation of an integer

[2]The dots in the period are written to make obvious the factorisation of the labelling $\boldsymbol{\lambda}_1$ determined by the signature $\boldsymbol{s}_1$.

is a word over the alphabet $[\![p]\!] = \{0, 1, \ldots, p-1\}$ that does not begin with a 0 (and the set of representations is not a padded language). But there are cases where one wants to have the possibility to write a number differently. For the addition of two numbers for instance, it is convenient to have representations of the same length, and the shorter one is prefixed with the adequate numbers of 0's to match with the longer one. It is currently said that the shorter representation is *padded* with 0's.

The branch language K of an i-tree has clearly the property that any word of K can be prefixed by an arbitrary number of the label of the loop (on the root) and still be in K. The label of the loop of an i-tree is called *padding letter*. The notion of padded language can be given a purely language-theoretic definition as follows.

Definition 6. *Let A be a (totally ordered) alphabet and let a be a letter in A. A language K over A is said to be* a-padded *if the following conditions hold:*
(i) $u \in K \Leftrightarrow au \in K$;
(ii) *If bu is in K, with b in A, then b is not smaller than a.*
A language is padded *if it is a-padded for some letter a.*

If a language is padded, it is a-padded for a unique a: the second condition of Definition 6 implies that if K is both a-padded and a'-padded, then $a = a'$.

Notation. *A padded language is written either as a^*L, or as $\widehat{L}$ if the padding letter does not need to be specified; in both cases L is then implicitly defined as the set of the words of the padded language which do not start with the padding letter.*

It is easy to verify that if $\mathcal{I}$ is a labelled i-tree and $\mathcal{T}$ the corresponding tree, then the branch language of $\mathcal{I}$ is a padded language $\widehat{L}$ where L is the branch language of $\mathcal{T}$. Our notation transfers at the level of branch languages the correspondence between trees and i-trees.

To some extent, there is no difference, between a labelled (i-)tree and the prefix-closed language of its branches. We may thus speak of the labelled signature, and of the signature, of a prefix-closed language and take the corresponding notation: the branch language of a tree $\mathcal{T}_x$ (resp. an i-tree $\mathcal{I}_x$), for some index x, is denoted by L_x (resp. $\widehat{L}_x$). Proposition 5 may then be rephrased in the following way.

Proposition 7. *A prefix-closed padded language $\widehat{L}$ uniquely determines a labelled i-tree and hence a valid labelled signature,* the *labelled signature of $\widehat{L}$ and conversely any valid labelled signature $(\boldsymbol{s}, \boldsymbol{\lambda})$ uniquely determines a labelled i-tree $\mathcal{I}_{(\boldsymbol{s},\boldsymbol{\lambda})}$ and hence a prefix-closed padded language $\widehat{L}_{(\boldsymbol{s},\boldsymbol{\lambda})}$, whose signature is precisely $(\boldsymbol{s}, \boldsymbol{\lambda})$.*

Remark 8. *Any language L over a totally ordered alphabet A* can be made padded *by adding a new letter # to A and by setting # smaller than all letters in A. We then consider $K = \#^*L$ instead of L and L is rational if and only if so is K.*

Remark 9. *A very 'simple' tree may produce an artificially 'complex' language when paired with a 'complex' labelling. For instance, the infinite unary tree may be labelled by an infinite word whose prefixes form a non-recursive language. Therefore, any result relative to* languages *defined by signatures will always require some hypothesis to constrain the* labelling. *The notion of* periodic labelling *as in the example shown in Figure 3 or* s-morphic labelled signature *defined in the next Section 3 are examples of such hypotheses.*

2.5. Trees, languages and abstract numeration systems

The identification between a prefix-closed language L over a totally ordered alphabet and the ordered labelled tree $\mathcal{T}_L$ whose branch language is L (and whose set of nodes is $\mathbb{N}$) is very close to the notion of *Abstract Numeration Systems* (ANS) introduced by Lecomte and Rigo (*cf.* [7, 13]). In this setting, the language L over the totally ordered alphabet A is ordered by the trace of the *radix order* over A^* and — since it is meant to define a *numeration system* — the representation of an integer n in this system, also called the *L-representation* of n and denoted by $\langle n \rangle_L$, is the $(n+1)$-th word of L in the radix order.

This notion generalises the situation in classical numeration systems. Let us take for instance the numeration in base 3. The usual way for defining the representation of integers in that system is to define an *evaluation function* $\pi_3 \colon [\![3]\!]^* \to \mathbb{N}$ by the following: if $w = d_k d_{k-1} \cdots d_1 d_0$ is a word of length $k+1$, then

$$\pi_3(w) \;=\; \pi_3(d_k d_{k-1} \cdots d_1 d_0) \;=\; \sum_{i=0}^{k} d_i 3^i \quad . \tag{3}$$

Note that in this case, it is convenient to have the digits indexed *from right to left*.

As said above, every integer n is uniquely represented by a word $\langle n \rangle_3$ of $[\![3]\!]^* = \{0,1,2\}^*$ which does not begin with a 0, that is, the set L_3 of integer representations in base 3 is defined by

$$L_3 \;=\; \{\langle n \rangle_3 \mid n \in \mathbb{N}\} \;=\; \{1,2\}\{0,1,2\}^* \cup \{\varepsilon\}$$

(with the convention that the integer 0 is represented by ε rather than by the digit 0, which suits us better). It then turns out that $\langle n \rangle_3$ is the *$(n+1)$-th word* of L_3 in the radix order, that is, the representation of n in base 3 coincides with the representation of n in the ANS defined by L_3 over the ordered alphabet $\{0,1,2\}$:

$$\forall n \in \mathbb{N} \qquad \langle n \rangle_3 \;=\; \langle n \rangle_{L_3} \quad .$$

On the other hand, since $\mathcal{T}_L$ is visited by a *breadth-first search*, the $(n+1)$-th node of $\mathcal{T}_L$ — labelled with n — is reached from the root by the $(n+1)$-th word — in the radix order — of the branch language of $\mathcal{T}_L$, that is, L itself (under the hypothesis that L is prefix-closed, which is necessary for the identification between L and $\mathcal{T}_L$).

These two descriptions show that considering a prefix-closed language over an ordered alphabet as an ANS or as the branch language of a labelled ordered

tree are two ways of expressing the concept, namely the radix order over the language. The similarity between the two notions is further shown in the following equation

$$\forall n \in \mathbb{N} \qquad 0 \xrightarrow[\mathcal{T}_L]{\langle n \rangle_L} n \ , \tag{4}$$

which implies

$$\forall n, m \in \mathbb{N}\,,\ \forall a \in A \qquad \langle n \rangle_L a \ = \ \langle m \rangle_L \quad \Longleftrightarrow \quad n \xrightarrow[\mathcal{T}_L]{a} m \ . \tag{5}$$

3. S-morphic signatures

Now that the general framework of signature is set up, we may turn to the case of rational padded languages. We begin with the definition of the *folding automaton morphism* between a rational i-tree and the finite automaton that accepts its branch language. We then characterise the signature of these languages in terms of fixed point of iterated (word) morphisms. In the last two sections we show that the labelling does not really matter and we consider the special case of ultimately periodic signatures.

3.1. Finite automata and rational padded languages

For the terminology, notation and basic definitions on finite automata and *rational* (or *regular*) languages (and transducers in the forthcoming sections) we essentially follow [14] (*cf.* also [15]): a *deterministic automaton* $\mathcal{A}$ over A^* is written $\mathcal{A} = \langle\, Q, A, \delta, i, F \,\rangle$ where Q is the set of states, A the alphabet, δ the transition function, i the initial state and F the set of final states. For all p in Q and a in A, we write

$$p \xrightarrow[\mathcal{A}]{a} q \tag{6}$$

if $\delta(p, a) = q$. The transition function δ is extended to $Q \times A^*$ and it is convenient to write $\delta(i, w) = i \cdot w$ for w in A^*. A word w of A^* is *accepted* (or *recognised*) by $\mathcal{A}$ if $i \cdot w \in F$. The language $L(\mathcal{A})$ accepted (or recognised) by $\mathcal{A}$ is the set of words accepted (by $\mathcal{A}$).

All automata we deal with are finite and deterministic — and we thus call them simply automata — but the infinite trees and i-trees may also be seen as *infinite* (deterministic) automata, with $\mathbb{N}$ as set of states and where the root 0 is the unique initial state and all nodes are final. The writing (2) is then consistent with (6) and the language $\widehat{L}$ (resp. L) is accepted by the 'automaton' $\mathcal{I}_L$ (resp. $\mathcal{T}_L$).

An *automaton morphism* φ from an automaton $\mathcal{A} = \langle\, Q, A, \delta, i, F \,\rangle$ to an automaton $\mathcal{B} = \langle\, R, A, \eta, j, G \,\rangle$ is a map $\varphi\colon Q \to R$ such that

(i) $\varphi(i) = j$,

(ii) $\varphi(F) = G$, and

(iii) $\forall p, q \in Q\,,\ \forall a \in A \qquad p \xrightarrow[\mathcal{A}]{a} q \quad \Longrightarrow \quad \varphi(p) \xrightarrow[\mathcal{A}]{a} \varphi(q) \ .$

The morphism φ is a *covering* if moreover

(iv) $\varphi^{-1}(G) = F$, and

(v) $\forall p, q \in Q\,,\ \forall a \in A \qquad \varphi(p) \xrightarrow[\mathcal{A}]{a} \varphi(q) \quad \Longrightarrow \quad p \xrightarrow[\mathcal{A}]{a} q \ .$

The definition of covering is simpler here than in the general case (*cf.* [14]) since we consider deterministic automata only. Obviously, if $\varphi\colon \mathcal{A} \to \mathcal{B}$ is a covering then $L(\mathcal{A}) = L(\mathcal{B})$.

If a padded language $\widehat{L} = a^*L$ is a rational language, then its *minimal* automaton $\mathcal{A}_{\widehat{L}} = \langle\, Q, A, \delta, i, F \,\rangle$ has the property that the initial state i bears a loop whose label is a (since Definition 6 (i) implies $a^{-1}K = K$) and a is the smallest of all labels of transitions outgoing from i (as a consequence of Definition 6 (ii)). By metonymy and for conciseness, we call *padded* an automaton with such a property and the language accepted by a padded automaton is a padded language.

If a rational language L is prefix-closed, then any *trim* automaton $\mathcal{A}$ that accepts L has the property that every state is final and conversely an automaton every state of which is final (is trim if accessible and) accepts a prefix-closed language. For conciseness and by metonymy again, we call *prefix-closed* an automaton with such a property.

Let us note that any labelled i-tree $\mathcal{I}_L$ may bee seen as an *infinite prefix-closed padded automaton* which accepts $\widehat{L}$ and $\mathcal{T}_L$ as an *infinite prefix-closed automaton* which accepts L.

Proposition 10. *Let $\mathcal{A} = \langle\, Q, A, \delta, i, Q \,\rangle$ be a prefix-closed padded automaton, $\widehat{L} = L(\mathcal{A})$ the padded language it accepts and $\mathcal{I}_L$ the associated i-tree. Let $\varphi_{\mathcal{A}}\colon \mathbb{N} \to Q$ be the function that maps every node n of $\mathcal{I}_L$ to the state of $\mathcal{A}$ reached by the reading of $\langle n \rangle_L$:*

$$\forall n \in \mathbb{N} \qquad \varphi_{\mathcal{A}}(n) \ = \ i \cdot \langle n \rangle_L \ .$$

Then, $\varphi_{\mathcal{A}}$ is a covering from $\mathcal{I}_T$ onto $\mathcal{A}$ (which we call the folding morphism *on $\mathcal{A}$).*

Proof. Since $\varepsilon = \langle n \rangle_L$, $\varphi(0)$ is the initial state i, conditions (ii) and (iv) follow from the fact that both $\mathcal{I}_L$ and $\mathcal{A}$ are prefix-closed and conditions (ii) and (iv) both follow from Equation (5):

$$\forall n, m \in \mathbb{N}\,,\ \forall a \in A \qquad n \xrightarrow[\mathcal{I}_L]{a} m \quad \Longleftrightarrow \quad \varphi(n) \xrightarrow[\mathcal{A}]{a} \varphi(m) \ . \tag{7}$$

□

Indeed, $\mathcal{I}_L$ is the partial unfolding of the automaton $\mathcal{A}$, partial because the loop $i \xrightarrow[\mathcal{A}]{a} i$ is not unfolded.

Remark 11. *Let $\mathcal{A} = \langle\, Q, A, \delta, i, F \,\rangle$ be a padded automaton that accepts the padded language a^*L.* It is not true *that suppressing the loop $i \xrightarrow[\mathcal{A}]{a} i$ yields an automaton that accepts L. The latter property holds only if the loop*

suppression yields a standard *automaton, that is, an automaton in which there is no transition incoming to the initial state. This is not always the case, as witnessed for instance by the examples given in Figures 5(a) and 5(b) later on.*

3.2. Automatic and morphic words

The study of the relationship between morphic words and automata goes back to Cobham (in [6], *cf.* Theorem 14 below) and has been developed by Rigo and Maes (in [3], *cf.* Theorem 15 below). These results require some further definitions on automata and substitutions before being stated. We follow [16] for the terminology and basic definitions on *substitutions* which we rather call *morphisms* in order to have a better consistency with the whole field of automata theory.

Automatic words are built via automata with final function, morphic words via prolongable morphisms. We first recall the classical instance of this equivalence in the case of p-automatic words and p-uniform morphic words. We then state the equivalence of these two generating devices. The classic reference for automatic words is the treatise of Allouche and Shallit [17]. We only recall what is necessary to set up the link with the notion of signature.

3.2.1. Automata with final function and automatic words

An *automaton with final function* $\mathcal{A}$ is an automaton endowed with a function from the set of final states to a set D called also *alphabet* here. An automaton with final function is then specified by a classical deterministic automaton $\mathcal{A} = \langle\, Q, A, \delta, i, F \,\rangle$ together with a total function $f\colon F \to D$. Such an automaton realises a map from A^* to D, denoted by $|\mathcal{A}|$, whose domain is $L(\mathcal{A})$ and defined by $|\mathcal{A}|\,(w) = f(i \cdot w)$ if $i \cdot w$ belongs to F and $|\mathcal{A}|\,(w)$ is undefined otherwise.

Let $p > 1$ be an integer that will be considered as a base. And let D be finite alphabet. An *infinite word* $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ of D^ω is said to be *p-automatic*[3] if there exists an automaton $\mathcal{A}$ over $[\![p]\!]^*$ with final function f in D such that for every n in $\mathbb{N}$, $|\mathcal{A}|\,(\langle n \rangle_p) = s_n$, that is, if the reading of the representation of the integer n in base p leads $\mathcal{A}$ to a state that is mapped to s_n by f.

3.2.2. Prolongable morphisms and morphic words

Let A and D be two alphabets. A *letter-to-letter morphism* from A^* to D^* is a morphism that maps every letter of A onto a letter of D (sometimes called a *strictly alphabetic* morphism). A *continuous morphism* from A^* to D^* is a morphism that maps no letter of A onto the empty word of D^* (sometimes called a *non-erasing* morphism). Let 'a' be a letter in A. A morphism (an endomorphism indeed) $\sigma\colon A^* \to A^*$ is said to be *prolongable on 'a'* if 'a' is the first letter of $\sigma(a)$ *and* if the length of the words of the sequence $(\sigma^n(a))_{n\in\mathbb{N}}$

[3] It is usually said that a *sequence* is p-automatic rather than an *infinite word*, *cf.* the eponymous work [17] already cited. We use the latter to be consistent with the definitions of signature and labelling.

tends to infinity. In the sequel, we say that a morphism is *prolongable* without mentioning *on 'a'* and this letter is thus kept by convention for this usage (and we enclose it between quotes in order to improve readability).

If σ is prolongable, there exists by definition a non-empty word u such that $\sigma(a) = au$ and the sequence $(\sigma^n(a))_{n\in\mathbb{N}}$ converges to the infinite word

$$\sigma^\omega(a) \;=\; au\sigma(u)\sigma^2(u)\cdots \quad . \tag{8}$$

Any infinite word of this form $\sigma^\omega(a)$ for a certain prolongable morphism σ is called a *pure morphic word*.

Definition 12. *An infinite word $\boldsymbol{s}$ is a* morphic word *if it is the image of a pure morphic word by a morphism, that is, if $\boldsymbol{s} = f(\sigma^\omega(a))$, where $\sigma\colon A^* \to A^*$ is a prolongable morphism and $f\colon A^* \to D^*$ a morphism.*

Without loss of generality, one can take more restrictive hypotheses to generate morphic words.

Lemma 13 ([18], *cf.* also [17])**.** *Let $\boldsymbol{s}$ be a morphic word of D^ω. Then, there exist an alphabet A, a* continuous *prolongable morphism $\sigma\colon A^* \to A^*$ and a* letter-to-letter *morphism $f\colon A^* \to D^*$ such that $\boldsymbol{s} = f(\sigma^\omega(a))$.*

3.2.3. The coincidence between automatic and morphic words

Let $p > 1$ be an integer that will be considered as a base. A morphism $\sigma\colon A^* \to A^*$ is said to be *p-uniform* if the image by σ of every letter of A is a *word of length p*. A morphic word $\boldsymbol{s}$ is *p-uniform* if there exist a letter-to-letter morphism f and a p-uniform prolongable morphism σ such that $\boldsymbol{s} = f(\sigma^\omega(a))$. Cobham's result reads then:

Theorem 14 ([6])**.** *Let $p > 1$ be an integer. An infinite word is p-automatic if and only if it is a p-uniform morphic word.*

The notion of rational abstract numeration system (rational ANS) has allowed Rigo and Maes to generalise this correspondence beyond the hypothesis of p-uniformity.

Let L be a *rational* language over a (totally ordered) alphabet A that will be considered as an ANS (see Section 2.5). Along the same line as the definition of p-automatic words, an *infinite word* $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ of D^ω is said to be *L-automatic* if there exists an automaton $\mathcal{A}$ over A with final function in D such that for every n in $\mathbb{N}$, $|\mathcal{A}|(\langle n\rangle_L) = s_n$, that is, if $\mathcal{A}$ accepts L — hence the hypothesis that L is rational — and the reading of the L-representation of the integer n leads $\mathcal{A}$ to a state which is mapped to s_n by f. The generalisation of Theorem 14 reads then:

Theorem 15 ([3])**.** *An infinite word $\boldsymbol{s}$ is L-automatic for a certain rational ANS L if and only if $\boldsymbol{s}$ is a morphic word.*

3.3. S-morphic signatures

If the alphabet D above is an alphabet of non-negative digits, the p-automatic words and the L-automatic words are infinite words of non-negative integers, hence signatures. We now define signatures that are morphic words of a special form.

Definition 16. *Let $\sigma\colon A^* \to A^*$ be a prolongable morphism.*

(i) *We denote by $f_\sigma\colon A^* \to D^*$ the letter-to-letter morphism defined by*

$$\forall b \in A \qquad f_\sigma(b) \;=\; |\sigma(b)|$$

(D is thus an alphabet of digits). The morphic word $f_\sigma(\sigma^\omega(a))$ is called an s-morphic signature.

(ii) *Let B be an ordered alphabet. A morphism $g\colon A^* \to B^*$ is* consistent *with σ if, for every b in A,*

$$|g(b)| \;=\; |\sigma(b)| \;=\; f_\sigma(b) \;, \quad \text{and } g(b) \text{ is increasing} \;. \tag{9}$$

If g is consistent with σ, the pair $\big(f_\sigma(\sigma^\omega(a)),\ g(\sigma^\omega(a))\big)$ is called an s-morphic labelled signature*, and also denoted by (σ, g) for convenience.*

If σ is a prolongable morphism, then for any prefix v of $\sigma^\omega(a)$, $|\sigma(v)| > |v|$ and then it holds:

Proposition 17. *An s-morphic signature is valid, and so is an s-morphic labelled signature.*

The morphism f_σ is entirely determined by σ and the set of s-morphic signatures is strictly contained in the one of morphic words; on the other hand, it is incomparable with the set of *pure* morphic words (*cf.* Remark 32).

Example 18. *The labelled signature $(\boldsymbol{s}_1, \boldsymbol{\lambda}_1)$ of Figure 3 with $\boldsymbol{s}_1 = (321)^\omega$ and $\boldsymbol{\lambda}_1 = (012121)^\omega$ is an s-morphic signature. Indeed, $\boldsymbol{s}_1 = f_{\sigma_1}({\sigma_1}^\omega(a))$ where $\sigma_1 : \{a,b,c\}^* \to \{a,b,c\}^*$ is defined by:*

$$\sigma_1(a) \;=\; abc \;, \qquad \sigma_1(b) \;=\; ab \qquad \text{and} \qquad \sigma_1(c) \;=\; c$$

and $\boldsymbol{\lambda}_1 = g_1({\sigma_1}^\omega(a))$ where $g_1 : \{a,b,c\}^* \to \{0,1,2\}^*$ is the morphism consistent with σ defined by:

$$g_1(a) \;=\; 012 \;, \qquad g_1(b) \;=\; 12 \qquad \text{and} \qquad g_1(c) \;=\; 1 \;.$$

An interpretation of the padded language $\widehat{L}_{(\sigma_1, g_1)}$ is given in Example 29.

Example 19 (The Fibonacci signature)**.** *The Fibonacci word is the pure morphic word ${\sigma_2}^\omega(a)$ defined by $\sigma_2(a) = ab$ and $\sigma_2(b) = a$:*

$${\sigma_2}^\omega(a) \;=\; abaababaabaab\cdots \;.$$

The s-morphic signature defined by σ_2 is

$$f_{\sigma_2}\big({\sigma_2}^{\omega}(a)\big) \;=\; 2122121221221\cdots$$

Let g_2 be the morphism consistent with σ_2 defined by $g_2(a) = 01$ and $g_2(b) = 1$:

$$g_2\big({\sigma_2}^{\omega}(a)\big) \;=\; 01.0.01.01.0.01.0.01.01.0.01.01.0\cdots$$

It is remarkable that the branch language $\widehat{L}_{(\sigma_2,g_2)}$ of the i-tree $\mathcal{I}_{(\sigma_2,g_2)}$ shown in Figure 4 is the language of the representations of the integers in the Fibonacci numeration system.

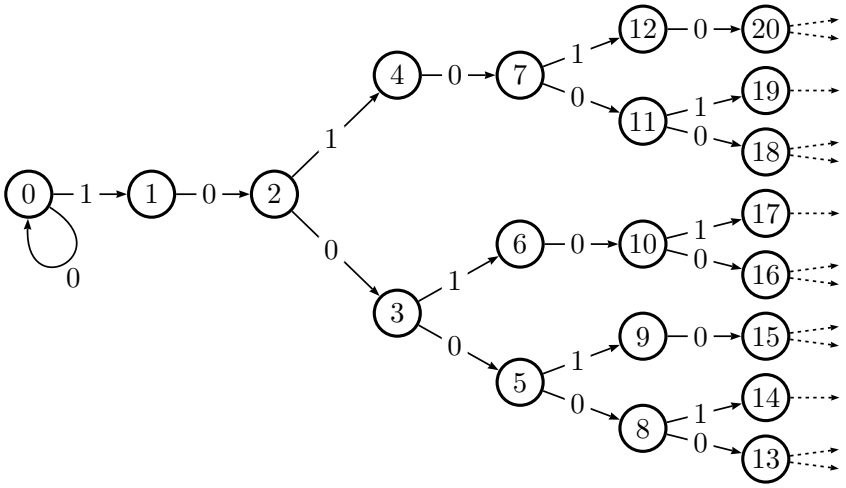


Figure 4: The labelled i-tree $\mathcal{I}_{(\sigma_2,g_2)}$

We may now state our characterisation, or serialisation, result.

Theorem 20. *A prefix-closed padded language is rational if and only if its labelled signature is s-morphic.*

This statement is close to be a consequence of Theorem 15, but not quite: it is in some sense more precise. Let L be a rational prefix-closed language and let $\mathcal{A}$ be a trim automaton accepting L, every state of which is thus final. If $\mathcal{A}$ is endowed with the final function that maps every state to its outgoing degree, then the L-automatic word it realises is precisely the signature $\boldsymbol{s}_L$ of L and from Theorem 15 follows then that $\boldsymbol{s}_L$ is a morphic word. Similarly, the labelling $\boldsymbol{\lambda}_L$ of L may be shown to be L-automatic, hence morphic by Theorem 15 again. However, we have to go from morphic to s-morphic and moreover, it is *the pair* $(\boldsymbol{s}_L, \boldsymbol{\lambda}_L)$ that is to be shown s-morphic, that is, generated by *the same* prolongable morphism (*cf.* Definition 16). The proof of Theorem 20 requires a more accurate construction.

3.4. The correspondence between automata and s-morphic signatures

The core of the proof of Theorem 20 lies indeed in a statement found also in the Rigo and Maes paper [3] and restated here as Lemma 25. It is based on two opposite constructions described in Definitions 21 and 22 that build an automaton from an s-morphic labelled signature and conversely.

Definition 21. *Let (σ, g) be an s-morphic labelled signature where $\sigma\colon A^* \to A^*$ is prolongable on 'a' and $g\colon A^* \to B^*$ a morphism consistent with σ. This signature defines the automaton $\mathcal{A}_{(\sigma,g)}$ over B^*:*

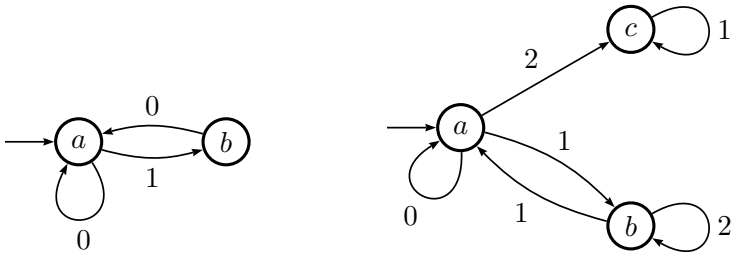
$$\mathcal{A}_{(\sigma,g)} \;=\; \langle\, A, B, \delta, a, A \,\rangle \;\;,$$

with A as set of states and 'a' as initial state and whose transitions are defined in the following way: from every state b, there is a transition to every letter *of $\sigma(b)$ and the transition to the k-th letter c of $\sigma(b)$ is labelled by the k-th letter y of $g(b)$:*

$$b \xrightarrow[\mathcal{A}_{(\sigma,g)}]{y} c \;\;.$$

Since $g(b)$ is an increasing word for every b, $\mathcal{A}_{(\sigma,g)}$ is deterministic. Since σ is prolongable on 'a' and $g(a)$ an increasing word, $\mathcal{A}_{(\sigma,g)}$ is a padded automaton.

Figure 5(a) shows the automaton associated with the Fibonacci signature, Figure 5(b) the one associated with $(\boldsymbol{s}_1, \boldsymbol{\lambda}_1)$ (*cf.* Example 18).



(a) The automaton $\mathcal{A}_{(\sigma_2,g_2)}$ (b) The automaton $\mathcal{A}_{(\sigma_1,g_1)}$

Figure 5: Two automata built from an s-morphic signature

Notation. *For every state p of an automaton $\mathcal{A} = \langle\, Q, B, \delta, i, Q \,\rangle$, we write k_p for the number of transitions going out of p* minus 1. *For instance, in the automaton $\mathcal{A}_{(\sigma,g)}$ of Definition 21, $k_b = |\sigma(b)| - 1 = |g(b)| - 1$ for every b in B. We take this convention for the easiness of writing, as a consequence of the fact that the* first *letter of a word is* indexed by 0*, the k-*th *by $k-1$, and we rather not have $k_p - 1$ written as an index (*cf. *Equation (10)).*

Definition 22. *Let B be an ordered alphabet and $\mathcal{A} = \langle\, Q, B, \delta, i, Q \,\rangle$ a padded, deterministic and prefix-closed automaton over B. Two morphisms $\sigma_{\mathcal{A}}\colon Q^* \to Q^*$ and $g_{\mathcal{A}}\colon Q^* \to B^*$ are associated with $\mathcal{A}$ in the following way. For every state p*

in Q, let $p \xrightarrow{b_0} q_0$, $p \xrightarrow{b_1} q_1$, …, $p \xrightarrow{b_{k_p}} q_{k_p}$ be the (k_p+1) transitions going out from p, listed in the increasing order of the labels: $b_0 < b_1 < \cdots < b_{k_p}$. The values of $\sigma_{\mathcal{A}}$ and $g_{\mathcal{A}}$ on p are then defined by:

$$\sigma_{\mathcal{A}}(p) \;=\; q_0\, q_i \cdots q_{k_p} \qquad \textit{and} \qquad g_{\mathcal{A}}(p) \;=\; b_0\, b_1 \cdots b_{k_p} \;\; . \tag{10}$$

Since $\mathcal{A}$ is padded, the loop $i \xrightarrow{b_0} i$ is the first in the list of transitions outgoing from 'i' and $\sigma_{\mathcal{A}}$ is prolongable on 'i'.

An easy and formal verification shows that the constructions described in Definitions 21 and 22 are opposite of each other:

Proposition 23.
(i) *If $\mathcal{B}$ is a trim prefix-closed padded automaton, then $\mathcal{A}_{(\sigma_{\mathcal{B}}, g_{\mathcal{B}})} = \mathcal{B}$.*
(ii) *If (τ, h) is an s-morphic labelled signature, then $\sigma_{\mathcal{A}_{(\tau,h)}} = \tau$ and $g_{\mathcal{A}_{(\tau,h)}} = h$.*

Theorem 20 is the direct consequence of the following proposition which lifts the correspondence between automata and the signature stated in Proposition 23 to a correspondence between automata and the languages and which also show the consistency of Definitions 21 and 22.

Proposition 24.
(i) *The signature of the language accepted by a padded automaton $\mathcal{A}$ is $(\sigma_{\mathcal{A}}, g_{\mathcal{A}})$.*
(ii) *The automaton $\mathcal{A}_{(\sigma,g)}$ associated with the s-morphic (valid) signature (σ, g) accepts the padded language $\widehat{L}_{(\sigma,g)}$.*

The core of the proof of the lifting consists in the description of the relationship between the computations in an automaton $\mathcal{A}$ and the infinite word generated by $\sigma_{\mathcal{A}}$. It could be found in [3] under a different formulation.

Lemma 25 ([3])**.** *Let $\mathcal{A} = \langle\, Q, A, \delta, i, Q \,\rangle$ be a padded automaton, $\widehat{L}$ the padded language it accepts and $\sigma_{\mathcal{A}}$ the associated prolongable morphism on 'i'. Let $\sigma_{\mathcal{A}}^{\omega}(i) = q_0\, q_1\, q_2 \cdots$ be the pure morphic word of Q^{ω} generated by $\sigma_{\mathcal{A}}$. Then, for every integer n, q_n is the state of $\mathcal{A}$ reached after the reading of the representation of n in the ANS L:*

$$\forall n \in \mathbb{N} \qquad q_n \;=\; i \cdot \langle n \rangle_L \;\; .$$

Proof. Let $\varphi_{\mathcal{A}} \colon \mathcal{I}_L \to \mathcal{A}$ be folding morphism on $\mathcal{A}$ which we rather write φ here to lighten the notation. Proving the lemma amounts to establish

$$\sigma_{\mathcal{A}}{}^{\omega}\big(\varphi(0)\big) \;=\; \varphi(0)\,\varphi(1)\,\varphi(2) \cdots \tag{11}$$

Since φ is an automaton morphism, if $m, (m+1), \ldots, (m+k)$ are all the successors of n in $\mathcal{I}_L$ then

$$\sigma_{\mathcal{A}}\big(\varphi(n)\big) \;=\; \varphi(m)\,\varphi(m+1) \cdots \varphi(m+k) \;\; . \tag{12}$$

Let us establish by induction on d the following claim.

Claim 25.1. *For every integer d*

$$\sigma_{\mathcal{A}}{}^{d}\big(\varphi(0)\big) \;=\; \varphi(0)\varphi(1)\varphi(2)\cdots\varphi(m)$$

where m is the greatest node of $\mathcal{I}_L$ at depth d.

Applying the previous Equation (12) to the root 0 yields

$$\sigma_{\mathcal{A}}\big(\varphi(0)\big) \;=\; \sigma_{\mathcal{A}}(i) \;=\; iu \qquad \text{with} \qquad u \;=\; \varphi(1)\varphi(2)\cdots\varphi(k_0)$$

where $1, 2, \ldots, k_0$ are all the nodes *at depth* 1 in $\mathcal{I}_L$. The same equation applied to every n in $\{1, 2, \ldots, k_0\}$ implies then that

$$\sigma_{\mathcal{A}}(u) \;=\; \sigma_{\mathcal{A}}\big(\varphi(1)\ldots\varphi(k_0)\big) \;=\; \varphi(r)\varphi(r+1)\cdots\varphi(r+s) \;\;,$$

where $r, (r+1), \ldots, (r+s)$ are all the nodes *at depth* 2 in $\mathcal{I}_L$. The same argument used inductively shows that the following equation holds for every integer d:

$$\sigma_{\mathcal{A}}{}^{d}(u) \;=\; \varphi(r)\varphi(r+1)\cdots\varphi(r+s) \;\;,$$

where $r, (r+1), \ldots, (r+s)$ are the nodes *at depth* $(d+1)$ in $\mathcal{I}_L$. The whole claim follows from the previous equation, since for every integer d, $\sigma^d(i)$ is equal to $iu\sigma(u)\sigma^2(u)\cdots\sigma^{d-1}(u)$. And Equation (11) is in turn a direct consequence of the claim, hence the lemma holds. □

Proof of Proposition 24. (i) Since the folding morphism is a covering, the degree of the node n in $\mathcal{I}_L$ is equal to the out-degree of $\varphi(n)$ in $\mathcal{A}$. Besides, it follows from Definition 22 that $f_{\sigma_{\mathcal{A}}}$ maps every state of $\mathcal{A}$ to its outgoing degree, hence from Lemma 25 that the signature of $\mathcal{I}_L$, hence of $\widehat{L}$, is $f_{\sigma_{\mathcal{A}}}(\sigma_{\mathcal{A}}{}^{\omega}(i))$.

Similarly, it follows from Definition 22 that $g_{\mathcal{A}}$ maps every state of $\mathcal{A}$ to the concatenation of its outgoing labels taken in increasing order, hence that the labelling of $\mathcal{I}_L$ is $g_{\mathcal{A}}(\sigma_{\mathcal{A}}{}^{\omega}(i))$.

(ii) We write $\mathcal{A} = \mathcal{A}_{(\sigma,g)}$ and $\widehat{L}$ the language it accepts. It follows from (i) that the labelled signature of $\widehat{L}$ is $(\sigma_{\mathcal{A}}, g_{\mathcal{A}})$, which is equal to (σ, g) from Proposition 23 (ii), hence that $\widehat{L} = \widehat{L}_{(\sigma,g)}$ since labelled signatures and languages are in bijection (Proposition 5). □

For further use (in proof of Theorem 39, Section 4) we give a more precise version of Lemma 25.

Proposition 26. *Let (σ, g) be an s-morphic (valid) signature and $\mathcal{I}_L$ the i-tree it generates. Let n be an integer, w the prefix of length $(n+1)$ of $\sigma^{\omega}(a)$ and m the greatest successor of n in $\mathcal{I}_L$. Then, $\sigma(w)$ is the prefix of length $(m+1)$ of $\sigma^{\omega}(a)$.*

Proof. Let d be the depth of the node n, let n' and m' be the greatest nodes of $\mathcal{I}_L$ at depth $(d-1)$ and d respectively. It then holds $n' < n \leqslant m' < m$ and m is at depth $(d+1)$. The same argument as the one used for proving Claim 25.1 yields

$$\sigma\big(\varphi(0)\varphi(1)\cdots\varphi(n')\big) \;=\; \varphi(0)\varphi(1)\cdots\varphi(m')$$

and it remains to show that

$$\sigma\big(\varphi(n'+1)\,\varphi(n'+2)\cdots\varphi(n)\big) \;=\; \sigma\big(\varphi(m'+1)\,\varphi(m'+2)\cdots\varphi(m)\big)\;. \quad (13)$$

The nodes $(n'+1),(n'+2),\ldots,n$ are the smallest $(n-n')$ nodes at depth d, hence their successors are the smallest j nodes at depth $(d+1)$, for some integer j. Since $(m'+1)$ is the smallest node at depth $(d+1)$ and m is the maximal successor of n, the successors of the nodes $(n'+1),(n'+2),\ldots,n$ are $(m'+1),(m'+2),\ldots,m$.

Applying Equation (12) to every node $(n'+1),(n'+2),\ldots,n$ successively then yields Equation (13) and concludes the proof. □

3.5. The unimportance of labelling

By definition, the language generated by the s-morphic signature (σ,g) depends upon the two parameters σ and g. The intuition is that it depends 'heavily' on σ and 'lightly' on g or, to state it in another way, the languages generated by two (valid) signatures (σ,g) and (σ,h) are 'very similar'. Let us formalise this notion before we give the statement that applies to our case.

Let $\widehat{L}$ and $\widehat{K}$ be two padded languages over A^* and B^* respectively. We call *conversion function from L to K* the function $\chi\colon A^*\to B^*$ whose domain is L and such that for every integer n, $\chi\left(\langle n\rangle_L\right)=\langle n\rangle_K$ (and hence its image is K). The 'complexity' of the function χ is a good measure for the similarity between L and K, both considered as ANSs: the simpler the function, the closer the ANSs. We will not engage into a theory for the complexity of word functions. It will however easily be accepted that functions realised by finite automata are among the simplest, that no function (but the identity) will be simpler than a *strictly alphabetic*, that is, *letter-to-letter*, *morphism*. It is known from a theorem of Cobham [19] that the conversion between the two (very 'simple') languages $\{1\}\{0,1\}^*$ and $\{1,2\}\{0,1,2\}^*$, that is, the representations of integers in base 2 and in base 3, *is not* a function realised by a finite automaton. On the other hand, the conversion between the representations of integers in base 4 and in base 2 is a morphism ($0\mapsto 00$, $1\mapsto 01$, etc.) followed by the removal of a possible leading zero.

Automata that realise (word) functions are called *transducers*: they are automata whose transitions are labelled with *pairs of words*. We follow [14] (and [15]) for definitions on transducers and give here as few definitions as possible.

A transducer is *letter-to-letter* if the labels are pairs of letters, that is, taken in a product alphabet $A\times B$. A letter-to-letter transducer over $(A\times B)^*$ is *sequential* if the projection on A yields a *deterministic* automaton over A^*; it is *pure sequential* if moreover every state is final.[4] Two automata or transducers are said to be *graph-isomorphic* if their underlying graphs (obtained by erasing the transition labels) are isomorphic.

[4] Letter-to-letter pure-sequential transducers are also sometimes called *Mealy machines*.

Pure-sequential letter-to-letter transducers over $(A\times B)^*$ realise the partial functions from A^* to B^* that can be considered as the simplest next to strictly alphabetic morphisms. The following statement shows that the actual labelling of a rational ANS is not really important and was established in [20] in a more general framework.

Proposition 27 ([20])**.** *Let (σ, g) and (σ, h) be two s-morphic (valid) signatures. The conversion function from $L_{(\sigma,g)}$ to $L_{(\sigma,h)}$ is realised by a letter-to-letter pure-sequential transducer $\mathcal{T}$ that is graph-isomorphic to $\mathcal{A}_{(\sigma,g)}$ and $\mathcal{A}_{(\sigma,h)}$.*

Proof. Let $\sigma\colon A^* \to A^*$ be a prolongable morphism, and let $g\colon A^* \to B^*$ and $h\colon A^* \to C^*$ be two morphisms consistent with σ, that is, according to Definition 16, such that

$$\forall a \in A \qquad |\sigma(a)| \;=\; |g(a)| \;=\; |h(a)| \;.$$

We then define the morphism $t\colon A^* \to (B \times C)^*$ by

$$\forall a \in A \qquad t(a) \;=\; (b_0, c_0)(b_1, c_1)\cdots(b_k, c_k) \qquad \begin{array}{rcl} \text{where} \quad g(a) & = & b_0\, b_1 \cdots b_k \;; \\ \text{and} \quad h(a) & = & c_0\, c_1 \cdots c_k \;. \end{array}$$

The *automaton* $\mathcal{A}_{(\sigma,t)}$ is then a prefix-closed automaton whose alphabet is $(B \times C)$, that is, a letter-to-letter *transducer* $\mathcal{T}$ every state of which is final. This identification corresponds to the one that maps the free monoid $(A\times B)^*$ to the submonoid of $A^*\times B^*$ generated by $A\times B$: sequences of pairs of letters are mapped onto pairs of words of equal lengths.

The definition of $\mathcal{T}$ as the automaton $\mathcal{A}_{(\sigma,t)}$ associated with (σ, t) by the construction described in Definition 21 has several outcomes.

First, $\mathcal{T}$ is graph-isomorphic to any other automata associated with σ, in particular with $\mathcal{A}_{(\sigma,g)}$ and $\mathcal{A}_{(\sigma,h)}$.

Second, if $a \xrightarrow[\mathcal{T}]{(b,c)} a'$, then a' is the i-th letter of $\sigma(a)$ and (b, c) the i-th letter of $t(a)$ for a certain i, from which follows that b is the i-th letter of $g(a)$ and c the i-th letter of $h(a)$. Hence $\mathcal{A}_{(\sigma,g)}$ is the underlying input automaton and $\mathcal{A}_{(\sigma,h)}$ the underlying output automaton of $\mathcal{T}$. And then, first, $\mathcal{T}$ is pure sequential, second, $\mathcal{T}$ maps $L\big(\mathcal{A}_{(\sigma,g)}\big) = \widehat{L}_{(\sigma,g)}$ onto $L\big(\mathcal{A}_{(\sigma,h)}\big) = \widehat{L}_{(\sigma,h)}$.

Finally, since for every a in A both $g(a)$ and $h(a)$ are increasing words, it follows that $\mathcal{T}$ is *locally increasing*: for every pair of transitions $a \xrightarrow[\mathcal{T}]{(b,c)} a'$ and $a \xrightarrow[\mathcal{T}]{(b',c')} a''$ of $\mathcal{T}$ originating from the same state a, the following equivalence holds: $b < b' \iff c < c'$. It follows by an easy induction that $\mathcal{T}$ preserves the strict radix order: if $\mathcal{T}(u) = v$ and $\mathcal{T}(u') = v'$ then $u <_{\mathrm{rad}} u' \iff v <_{\mathrm{rad}} v'$. This, combined with the previous property, implies that $\mathcal{T}$ realises the conversion function from $L_{(\sigma,g)}$ to $L_{(\sigma,h)}$. □

3.6. The case of ultimately periodic signatures

Let $\boldsymbol{s} = u\,v^\omega = s_0 s_1 \cdots s_{m-1}(s_m s_{m+1} \cdots s_{m+q-1})^\omega$ be an ultimately periodic signature (remember that every letter of $\boldsymbol{s}$ is a digit). We call *growth ratio*

of $\boldsymbol{s}$ (or, alternatively, the growth ratio of v), and denote by $\mathsf{gr}(\boldsymbol{s})$, the average of the letters of v (which is also the limit of the average of the first n letters of $\boldsymbol{s}$, when n tends to infinity):

$$\mathsf{gr}(u\,v^\omega) \;=\; \mathsf{gr}(s_m s_{m+1}\cdots s_{m+q-1}) \;=\; \frac{1}{q}\sum_{i=0}^{q-1} s_{m+i} \;=\; \frac{S_{m+q}-S_m}{q}\quad .$$

We treat here the case where $\mathsf{gr}(v)$ is an integer, that is, when the sum of the letters of v is a multiple of its length q.

Proposition 28. *Let $\boldsymbol{s}$ be an ultimately periodic (valid) signature. If the growth ratio of $\boldsymbol{s}$ is an integer, then $\boldsymbol{s}$ is an s-morphic signature.*

Proof. We first detail the proof for purely periodic signatures, as it exhibits the core of the property.

Let $\boldsymbol{s} = s_0\,s_1\,s_2\cdots = (s_0\,s_1\cdots s_{q-1})^\omega$ be a purely periodic signature of period q. We denote by k the growth ratio of $\boldsymbol{s}$, that is, satisfying $S_q = k\,q$.

We consider the two alphabets $[\![q]\!]$ and $[\![k\,q]\!]$. Let $\varphi\colon [\![q]\!]^* \to [\![k\,q]\!]^*$ be the morphism defined by

$$\forall i \in [\![q]\!] \qquad \varphi(i) \;=\; S_i\,(S_i+1)\cdots(S_i+s_i-1)\quad .$$

Since $S_{i+1} = (S_i + s_i)$ for every integer i, it follows immediately that

$$\varphi\big(01\cdots(q-1)\big) \;=\; 012\cdots(kq-1)\quad .$$

Let ψ be the letter-to-letter morphism $[\![kq]\!] \to [\![q]\!]$ projecting the bigger alphabet to the smaller: $\forall i \in [\![k\,q]\!] \quad \psi(i) = (i \bmod q)$. We write $\sigma = \psi \circ \varphi$, an endomorphism on $[\![q]\!]^*$ which satisfies, from the previous equation, that

$$\sigma\big(01\cdots(q-1)\big) \;=\; \big(012\cdots(q-1)\big)^k\quad .$$

It follows that $(012\cdots(q-1))^\omega$ is a fixed point of σ.

The validity of $\boldsymbol{s}$ insures that for every integer i, $0 < i \leqslant q$, it holds $S_i > i$, hence the prefix w of length i of $(012\cdots(q-1))$ satisfies

$$|\sigma(w)| \;=\; |\varphi(w)| \;=\; |01\cdots(S_i-1)| \;=\; S_i \;>\; i \;=\; |w|\quad .$$

It follows that σ is prolongable on 0. The s-morphic signature induced by σ is by definition $f_\sigma\big[(012\cdots(q-1))^\omega\big]$; since $f_\sigma(i) = |\sigma(i)| = s_i$ for every integer $i < q$, the s-morphic signature induced by σ is $(s_0\,s_1\cdots s_{q-1})^\omega = \boldsymbol{s}$, concluding the proof in the purely periodic case.

The generalisation to an ultimately periodic signature $u\,v^\omega$ is easy but less elegant. We write $j = |u|$ and we introduce j new letters $\overline{0}, \overline{1}, \ldots, \overline{j-1}$ gathered within an alphabet denoted by A.

We consider S_j, which is the sum of the letters of u, and more precisely the euclidean division of $(S_j - j)$ by $(k\,q)$, of which we denote the quotient and remainder respectively by Q and R:

$$(S_j - j) \;=\; Q(kq) + R \qquad \text{and} \qquad 0 \;\leqslant\; R \;<\; kq\quad .$$

The word

$$w \;=\; \overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(012\cdots(kq-1)\big)^{Q}\,012\cdots(R-1)$$

is then of length S_j. The function φ is now a morphism $A\cup[\![q]\!]\to A\cup[\![kq]\!]$, defined implicitly by:

$$\begin{array}{ll}\forall i,\ 0\ \leqslant\ i\ <\ j & \varphi\big(\overline{0}\,\overline{1}\cdots\overline{i}\big)\ \text{ is the prefix of length } S_{i+1} \text{ of } w\ \ ,\\ \forall i,\ 0\ \leqslant\ i\ <\ q & \varphi(i)\ =\ (R+V_i)(R+V_i+1)\cdots(R+V_i+v_i-1)\ \ ,\end{array}$$

where for all integers $i,\ 0\leqslant i\leqslant q$, $V_i=(S_{i+j}-S_i)$ and $v_i=s_{i+j}$ (which implies that $V_{i+1}=V_i+v_i$ and $V_q=kq$); the integer V_i is thus the sum of the first i letters of v.

It follows that for every integer i,

$$\varphi\Big(\overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(q-1)\big)^{i}\Big)\ =\ \overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(kq-1)\big)^{Q+i}\,01\cdots(R-1)\ .$$

We extend the morphism ψ to $A\cup[\![kq]\!]$ by the identity over A, thus $\sigma=\psi\circ\varphi$ is now an endomorphism of $A\cup[\![q]\!]$; it then follows from the previous equation that, for every integer i:

$$\sigma\Big(\overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(q-1)\big)^{i}\Big)\ =\ \overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(q-1)\big)^{k(Q+i)}\,(0\bmod q)(1\bmod q)\cdots(R-1\bmod q)\ ,$$

hence, when i tends to infinity,

$$\sigma\Big(\overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(q-1)\big)^{\omega}\Big)\ =\ \overline{0}\,\overline{1}\cdots\overline{j-1}\,\big(01\cdots(q-1)\big)^{\omega}\ .$$

Similarly to the purely periodic case, the validity of the signature then ensures that σ is prolongable (on $\overline{0}$). □

Example 29. *The signature $\boldsymbol{s}_1=(321)^\omega$ considered in Example 18 is purely periodic. It is generated by the endomorphism σ_1 (defined there) which corresponds to the construction described in the previous proof. The padded language $\widehat{L}_{(\sigma_1,g_1)}$ is shown in Figure 3; it is accepted by the automaton $\mathcal{A}_{(\sigma_1,g_1)}$, shown in Figure 5(b).*

Applying a result from [2] yields that this language is a language of non-canonical representations of the integers in base 2 (that is, the growth ratio of $\boldsymbol{s}_1$): the $(n+1)$-th word of $L_{(\sigma_1,g_1)}$ in the radix order is a word $d_kd_{k-1}\cdots d_0$ over the (non-canonical) alphabet $\{0,1,2\}$ and its binary value $\sum_{i=0}^{k}d_i2^i$ is equal to n.

Example 30. *The ultimately periodic signature $\boldsymbol{s}_3=3\,1\,1\,8\,1\,0\,(3\,2\,1)^\omega$ is s-morphic, generated by the endomorphism σ_3 of $\big\{\overline{0},\overline{1},\overline{2},\overline{3},\overline{4},\overline{5},0,1,2\big\}$ defined as*

follows.

$$
\begin{array}{lll}
 & & \textit{prefix of } \boldsymbol{s}_3 \\
 & & \downarrow \\
\sigma_3(\overline{0}) = \overline{0}\overline{1}\overline{2} & (\ f_{\sigma_3}(\overline{0}) = 3\ ) \\
\sigma_3(\overline{1}) = \overline{3} & (\ f_{\sigma_3}(\overline{1}) = 1\ ) \\
\sigma_3(\overline{2}) = \overline{4} & (\ f_{\sigma_3}(\overline{2}) = 1\ ) \\
\sigma_3(\overline{3}) = \overline{5}0120120 & (\ f_{\sigma_3}(\overline{3}) = 8\ ) \\
\sigma_3(\overline{4}) = 1 & (\ f_{\sigma_3}(\overline{4}) = 1\ ) \\
\sigma_3(\overline{5}) = \varepsilon & (\ f_{\sigma_3}(\overline{5}) = 0\ ) \\
\sigma_3(0) = 201 & (\ f_{\sigma_3}(0) = 3\ ) \\
\sigma_3(1) = 20 & (\ f_{\sigma_3}(1) = 2\ ) \\
\sigma_3(2) = 1 & (\ f_{\sigma_3}(2) = 1\ )
\end{array}
$$

The language generated by σ_3 and the appropriate morphism g_3 is also a language of non-canonical representations of the integers in base 2.

Remark 31. *It can be shown that a language with an ultimately periodic signature $\boldsymbol{s}$, the growth ratio of which is not an integer,* cannot be a rational language. *Hence $\boldsymbol{s}$ is not an s-morphic signature. The proof of this statement is however more convoluted; it is the subject of another work of the authors [2].*

Remark 32. *It is obvious that any* purely periodic *word is a pure morphic word (every letter is sent to the period). It is hardly more difficult to see that an* ultimately periodic *word is also a pure morphic word. Hence an ultimately periodic signature is a pure morphic word independently of its growth ratio.*

It thus follows from the previous remark that a pure morphic signature is not necessarily an s-morphic signature.

4. Morphic numeration systems

In this section, we use our framework to describe a family of numeration systems originally defined in a series of papers authored by Dumont and Thomas [9, 4, 5], hence often called *Dumont-Thomas numeration systems* in the literature (*e.g.* [21]). The authors themselves spoke of numeration systems *associated with (the fixed point of) a substitution*; Berthé and Rigo call them *substitution numeration systems* in [22]. Since we systematically use the term of *morphism* rather than *substitution*, we suggest to call these systems *morphic numeration systems* (MNS). These systems have been considered in many developments in the fields of numeration, symbolic dynamics, and word combinatorics (*cf.* [23, 24, 25]).

In the following, we consider an alphabet B whose *letters* are *words* over another alphabet A and the free monoid it generates. For the sake of clarity, if u, or $a_0\,a_1\cdots a_k$, denotes a word of A^*, the corresponding letter of B is denoted by $[u]$, or $[a_0\,a_1\cdots a_k]$. Let us emphasise that a word of B^* is *not* a word of A^*, for instance the words $[\varepsilon]$ and $[\varepsilon][\varepsilon]$ are two different words of B^*; neither of them is equal to the empty word (still denoted by ε) of B^*. Moreover, let us

recall that by *strict* prefix of a word u, it is understood a prefix of u different from u but that may be equal to the empty word.

Definition 33. *A morphism $\sigma : A^* \to A^*$ prolongable on 'a' determines an* alphabet B_σ *and a* morphism $g_\sigma \colon A^* \to B_\sigma^*$ *in the following way.*

(i) *We denote by B_σ the set of the* strict *prefixes of the images of the letters of A by σ:*

$$B_\sigma \;=\; \big\{\, [u] \;\big|\; u \text{ is a strict prefix of } \sigma(b) \text{ for some } b \in A \,\big\} \;. \qquad (14)$$

(ii) *We denote by g_σ the morphism $g_\sigma \colon A^* \to B_\sigma^*$ which maps every letter b of A^* to the concatenation of all the strict prefixes of $\sigma(b)$ (each one being considered as a letter of B_σ):*

$$\forall b \in A \quad g_\sigma(b) \;=\; [u_0][u_1] \cdots [u_{k-1}] \;,$$
$$\text{where } \; k \;=\; |\sigma(b)| \;\text{ and }\; u_i \text{ is the prefix of } \sigma(b) \text{ of length } i. \qquad (15)$$

Intuitively, g_σ is a kind of 'stuttering expression' of σ; for instance,

$$\text{if } \; \sigma(a) \;=\; abccba \;, \;\; \text{then } \; g_\sigma(a) \;=\; [\varepsilon][a][ab][abc][abcc][abccb] \;.$$

Note that $g_\sigma(b)$ does **not** contain the letter $[\sigma(b)]$ (which is, a priori, not even a letter of B_σ) and, since it contains the letter $[\varepsilon]$, it has *the same length* as $\sigma(b)$. In particular, if $\sigma(b)$ is the empty word, then $g_\sigma(b)$ is the empty word of B_σ^* (and not equal to $[\varepsilon]$).

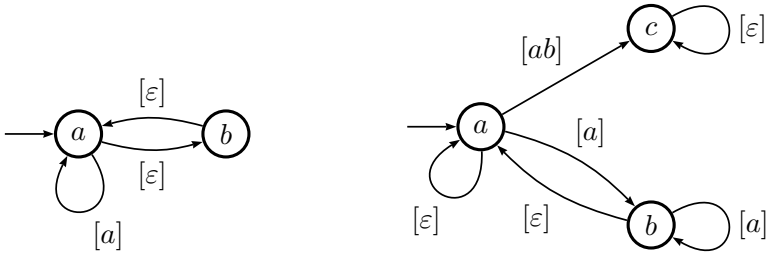
If the alphabet B_σ is ordered by the trace of the radix order of A^* on its subset B_σ, it immediately follows from the definition of g_σ itself that it holds:

Lemma 34. *The morphism g_σ is consistent with σ, hence the labelled signature (σ, g_σ) is valid.*

To lighten the writing, we denote by L_σ the language $L_{(\sigma,g_\sigma)}$ and by $\langle n \rangle_\sigma$ the representation of an integer n in the ANS L_σ, which is then entirely determined by σ. The valid signature (σ, g_σ) allows to build the automaton $\mathcal{A}_{(\sigma,g_\sigma)}$ as described by Definition 21. This automaton, written $\mathcal{A}_\sigma$ for short, is called the *prefix automaton associated with σ*. Figure 6 shows the prefix automata associated with the substitutions σ_1 and σ_2 from Section 3 (*cf.* Examples 18 and 19).

This automaton is implicitly present as soon as the original paper [9] through the notion of *suite admissible* which simulates the run of an automaton; it is then explicitly defined as an automaton in a subsequent paper [5]. The next statement gives a characterisation of the transitions of $\mathcal{A}_\sigma$ along the line of the original definition.

Lemma 35. *The automaton $\mathcal{A}_\sigma$ contains the transition $b \xrightarrow{[u]} c$ if and only if uc is a prefix of $\sigma(b)$.*



(a) The automaton $\mathcal{A}_{\sigma_2}$ (b) The automaton $\mathcal{A}_{\sigma_1}$

Figure 6: The prefix automata associated with two substitutions

Proof. From Definition 21, $\mathcal{A}_\sigma$ contains the transition $b \xrightarrow{[u]} c$ if and only if there is an integer i such that $[u]$ is the i-th letter of $g_\sigma(b)$ and c is the i-th letter of $\sigma(b)$. From Equation (15), such an i exists if and only if $u = u_{i-1}$ is the prefix of $\sigma(b)$ of length $(i-1)$ and c is the i-th letter of $\sigma(b)$, hence if and only if uc is the prefix of $\sigma(b)$ of length i. □

A direct consequence of this statement, used in the proof of Theorem 39, is a relationship between labels and states in prefix automata.

Lemma 36. *Let $b \xrightarrow{[u]} c$ be a transition of the automaton $\mathcal{A}_\sigma$. Then, the states reached from b by reading letters strictly smaller than $[u]$, taken in increasing order, form the state sequence spelled by u.*

The characteristic property of *morphic numeration systems* requires the definition of a new function.

Definition 37. *Let $\sigma : A^* \to A^*$ be a prolongable morphism and B_σ the associated (word) alphabet. The function ρ_σ from B_σ^* into A^* is defined in the following way: $\rho_\sigma(\varepsilon) = \varepsilon$ and for every word w of B_σ^* of length $(k+1)$, $w = [v_k][v_{k-1}] \cdots [v_0]$, it holds:*

$$\rho_\sigma\Big([v_k][v_{k-1}] \cdots [v_0]\Big) \;=\; \sigma^k(v_k)\sigma^{k-1}(v_{k-1}) \cdots \sigma^0(v_0) \;\; . \tag{16}$$

Alternatively, ρ_σ may be defined recursively by $\rho_\sigma(\varepsilon) = \varepsilon$ and

$$\forall w \in B_\sigma^* \,,\; \forall v \in A^* \qquad \rho_\sigma\Big(w\,[v]\Big) \;=\; \sigma\big(\rho_\sigma(w)\big)\,v \;\; . \tag{17}$$

The function ρ_σ is the combination of a convoluted use of the substitution σ and of the flattening map from B_σ^* to A^* and will play the role of the evaluation function in a numeration system. We shall see with Theorem 40 that the evaluation of a word w is indeed equal to $|\rho_\sigma(w)|$. We first illustrate this intuition with the translation of a classical numeration system in base p into a morphic numeration system.

Example 38 (Base 3)**.** *Let $\sigma_3 : \{a\}^* \to \{a\}^*$ be the morphism (prolongable on 'a') defined by $\sigma_3(a) = aaa$. We have*

$$B_{\sigma_3} \;=\; \big\{[\varepsilon], [a], [aa]\big\} \qquad \text{and} \qquad g_{\sigma_3}(a) \;=\; [\varepsilon][a][aa] \;\; .$$

The automaton $\mathcal{A}_{\sigma_3}$ is shown in Figure 7. If we think of the letters of B_{σ_3} as transcription of the digits 0, 1 and 2 respectively, then any word of ${\mathcal{A}_{\sigma_3}}^*$ which does not begin with the letter $[\varepsilon]$ is the transcription of the representation in base 3 of an integer n and (16) immediately yields that $|\rho_{\sigma_3}(w)| = n$.

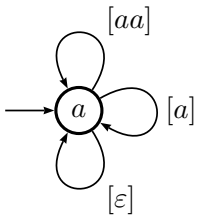


Figure 7: The automaton $\mathcal{A}_{\sigma_3}$

The essence of the result and construction of Dumont and Thomas may now be expressed in the following statement.

Theorem 39. *Let $\sigma\colon A^* \to A^*$ be a prolongable morphism on 'a', $\mathcal{A}_\sigma$ the prefix automaton associated with σ, and, for every integer n, $\langle n\rangle_\sigma$ the representation of n in the ANS L_σ. Then, $\rho_\sigma\left(\langle n\rangle_\sigma\right)$ is the prefix of length n of $\sigma^\omega(a)$.*

This result immediately implies the original statements of Dumont and Thomas:

Corollary 40 ([9])**.** *Let $\sigma\colon A^* \to A^*$ be a prolongable morphism. For every integer n, there exists a unique w in B_σ^* such that:*
(i) *w does not start with $[\varepsilon]$,*
(ii) *w is accepted by $\mathcal{A}_\sigma$,*
(iii) *and $\rho_\sigma(w)$ is of length n.*

Corollary 41 ([9])**.** *Let $\sigma\colon A^* \to A^*$ be a prolongable morphism and w a word of B_σ^* accepted by $\mathcal{A}_\sigma$. If the word $\rho_\sigma(w)$ has length n, then it is the prefix of length n of $\sigma^\omega(a)$.*

Similarly, the result shown later on by Berthé and Rigo in [22] is contained in Theorem 39.

Corollary 42 ([22])**.** *Every morphic numeration system is a prefix-closed rational abstract numeration system.*

Proof of Theorem 39. By induction on n. The theorem is verified for $n = 0$: indeed $\langle 0\rangle_\sigma = \varepsilon$ hence $\rho_\sigma(\varepsilon) = \varepsilon$, which is the prefix of length 0 of $\sigma^\omega(a)$.

Let m be a positive integer and n its predecessor, hence $n < m$. We show that the statement holds for m.

We write $\sigma^\omega(a) = a_0\, a_1\, a_2 \cdots$. From Lemma 25 follows that for every integer ℓ, the word $\langle \ell \rangle_\sigma$ leads $\mathcal{A}_\sigma$ to the state a_ℓ. Let k be the smallest positive integer such that $(n-k)$ has at least one outgoing arc and let $(m-j)$ be the maximal successor of $(n-k)$; necessarily, $j > 0$ and $(m-(j-1))$, $(m-(j-2)), \ldots, m$ are all successors of n.

It follows from Lemma 36 that the letter $b \in B_\sigma$ which labels the arc $n \xrightarrow{b} m$ (or, in other words, the rightmost letter of $\langle m \rangle_\sigma$) is $[u]$, where u is the sequence of the states reachable from a_n by reading letters strictly smaller than $[u]$, that is:

$$b \;=\; [u] \;=\; [a_{(m-j+1)}\, a_{(m-j+2)} \cdots a_{m-1}] \;. \tag{18}$$

Applying Proposition 26 to $(n-k)$ yields

$$\sigma(a_0\, a_1 \cdots a_{n-k}) \;=\; a_0\, a_1 \cdots a_{m-j} \tag{19}$$

Since (from the definition of k) for every integer i, $0 < i < k$, the node $(n-i)$ has no outgoing arc, the word $\langle n-i \rangle_\sigma$ reaches in $\mathcal{A}_\sigma$ the state a_n which then must have no outgoing transition, hence

$$\forall i\,,\; 0 < i < k \qquad \sigma(a_{n-i}) \;=\; \varepsilon \;. \tag{20}$$

Combining Equations (18), (19) and (20) finally yields that

$$\sigma(a_0\, a_1 \cdots a_{n-1})\, u \;=\; a_0\, a_1 \cdots a_{m-1} \;. \tag{21}$$

Since n is the predecessor of m, $\langle m \rangle_\sigma = \langle n \rangle_\sigma\, b$ holds and the induction hypothesis (IH) applied to n implies that $\rho_\sigma(\langle n \rangle_\sigma) = a_0\, a_1 \cdots a_{n-1}$. The following computation:

$$\begin{aligned}
\rho_\sigma\big(\langle m \rangle_\sigma\big) \;&=\; \rho_\sigma\big(\langle n \rangle_\sigma\, b\big) && \\
&=\; \sigma\big(\rho_\sigma(\langle n \rangle_\sigma)\big)\, u && \text{from Equation (17)} \\
&=\; \sigma(a_0\, a_1 \cdots a_{n-1})\, u && \text{from IH applied to } n \\
&=\; a_0\, a_1 \cdots a_{m-1} && \text{from Equation (21)}
\end{aligned}$$

concludes the proof. □

By Theorem 20, any rational ANS has an s-morphic signature and by definition every s-morphic signature defines a MNS; Proposition 27 implies thus that even if not every rational ANS is a MNS, it is very close to one.

Proposition 43 ([20])**.** *Let $\mathcal{A}$ be a prefix-closed padded automaton that accepts the prefix-closed padded rational ANS $\widehat{L}$. The conversion function from $\widehat{L}$ to $\widehat{L}_{\sigma_\mathcal{A}}$ is realised by a letter-to-letter pure-sequential transducer which is graph-isomorphic with $\mathcal{A}$.*

This statement results in the idea that prefix-closed rational ANSs and MNSs are essentially the same and have the same expressive power. To put it in another way, every prefix-closed rational ANS is 'equivalent' to a MNS which could

be considered as a representative of its class. Figure 8 shows the transducer that realises the conversion function from $\widehat{L}_{(\sigma_1,g_1)}$ to $\widehat{L}_{\sigma_1}$.

On the other hand, MNSs have a feature that ANSs are missing in general: the existence of an *evaluation function*. Given an ANS $\widehat{L} \subseteq A^*$, there is no natural way to attribute values to the words of A^* that do not belong to $\widehat{L}$ whereas given a prolongable morphism σ any word of B_σ^* may be evaluated as $|\rho_\sigma(w)|$. In a MNS σ, the evaluation function attributes to each letter $[u_i]$ of a word $[u_k][u_k-1]\cdots[u_0]$ the value $|\sigma^i(u_i)|$ that depends both on the letter $[u_i]$ itself and on the position i in the word w. We conclude this work with the example of the evaluation functions in the case of the two morphisms σ_1 and σ_2 that served as running examples. They show that understanding the true nature of the evaluation function is another problem that remains to be investigated.

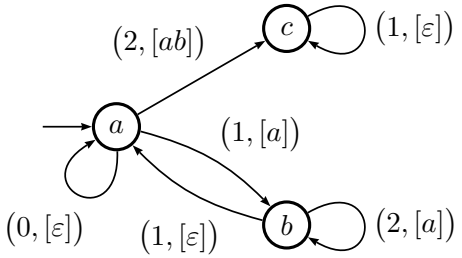


Figure 8: The transducer realising the conversion function from $\widehat{L}_{(\sigma_1,g_1)}$ to $\widehat{L}_{\sigma_1}$

Example 44 (Fibonacci)**.** *The* Fibonacci *morphism σ_2 is defined by $\sigma_2(a) = ab$ and $\sigma_2(b) = a$.*

The value given to the letter $[\varepsilon]$ is always 0 $(=\sigma_2^i(\varepsilon))$ independently of the position and the value given to the letter $[a]$ at position i is $|{\sigma_2}^i(a)|$, which is known to be equal to the the i-th number of the Fibonacci sequence.

In this case, the evaluation function ρ_σ corresponds to the evaluation function in the Fibonacci numeration system by applying the transcription $[\varepsilon] \mapsto 0$, $[a] \mapsto 1$ and the MNS L_{σ_2} is a positional numeration system*.*

Example 45 (Pseudo-base 2)**.** *Let $\sigma_1 : \{a,b,c\}^* \to \{a,b,c\}^*$ be the morphism previously defined in Example 18 by:*

$$\sigma_1(a) \;=\; abc \;\;, \qquad \sigma_1(b) \;=\; ab \qquad \text{and} \qquad \sigma_1(c) \;=\; c \;\;.$$

- *The value given to the letter $[\varepsilon]$ is always 0, independently of the position.*
- *The value given to the letter $[a]$ is 1 if it is the rightmost letter, or 32^{i-1} if it is at position $i > 0$.*
- *The value given to the letter $[ab]$ at position i is $(32^i - 1)$.*

It appears clearly that the evaluation function computed by ρ_{σ_1} does not correspond to a positional numeration system since the ratio of the values given to the letters $[a]$ and $[ab]$ is not independent of the position.

On the other hand, we know (from [2] for instance) that the ANS $L_{(\sigma_1,g_1)}$ (note that $g_1 \neq g_{\sigma_1}$) is a numeration system in base 2 with a non-canonical alphabet of digits (hence positional).

References

[1] S. Akiyama, C. Frougny, J. Sakarovitch, Powers of rationals modulo 1 and rational base number systems, Israel J. Math. 168 (2008) 53–91.

[2] V. Marsault, J. Sakarovitch, Trees and languages with periodic signature, in: E. Kranakis, G. Navarro, E. Chávez (Eds.), LATIN 2016, no. 9644 in Lect. Notes in Comput. Sci., 2016, pp. 605–618.

[3] M. Rigo, A. Maes, More on generalized automatic sequences, J. of Automata, Languages and Combinatorics 7 (3) (2002) 351–376.

[4] J.-M. Dumont, A. Thomas, Digital sum problems and substitutions on a finite alphabet, J. Number Theor. 39 (3) (1991) 351–366.

[5] J.-M. Dumont, A. Thomas, Digital sum moments and substitutions, Acta Arith. 64 (3) (1993) 205–225.

[6] A. Cobham, Uniform tag sequences, Math. Systems Theory 6 (1972) 164–192.

[7] P. Lecomte, M. Rigo, Numeration systems on a regular language, Theory Comput. Syst. 34 (2001) 27–44.

[8] E. Charlier, M. L. Gonidec, M. Rigo, Representing real numbers in a generalized numeration system, J. Comput. Syst. Sci. 77 (4) (2011) 743–759.

[9] J.-M. Dumont, A. Thomas, Systèmes de numération et fonctions fractales relatifs aux substitutions, Theoret. Computer Sci. 65 (2) (1989) 153–169.

[10] V. Marsault, J. Sakarovitch, Breadth-first generation of infinite trees and rational languages, in: A. M. Shur, M. V. Volkov (Eds.), DLT 2014, no. 8633 in Lect. Notes in Comput. Sci., 2014, pp. 252–259.

[11] V. Marsault, Énumeration et numération, Ph.D. thesis, Télécom–ParisTech (2016).

[12] R. Diestel, Graph Theory, Springer, 1997.

[13] P. Lecomte, M. Rigo, Abstract numeration systems, Ch. 3, in: Berthé and Rigo [26], pp. 108–162.

[14] J. Sakarovitch, Elements of Automata Theory, Cambridge University Press, 2009, corrected English translation of *Éléments de théorie des automates*, Vuibert, 2003.

[15] C. Frougny, J. Sakarovitch, Number representation and finite automata, Ch. 2, in: Berthé and Rigo [26], pp. 34–107.

[16] V. Berthé, M. Rigo, Preliminaries, Ch. 1, in: Encyclopedia Math. Appl. [26], pp. 1–33.

[17] J.-P. Allouche, J. Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, 2003.

[18] A. Cobham, On the Hartmanis-Stearns problem for a class of tag machines, in: 9th Symp. Switching and Automata Theory, 1968, pp. 51–60.

[19] A. Cobham, On the base-dependance of the sets of numbers recognizable by finite automata, Math. Systems Theory 3 (1969) 186–192.

[20] V. Marsault, Surminimisation of automata, in: I. Potapov (Ed.), DLT 2015, no. 9168 in Lect. Notes in Comput. Sci., Springer, 2015, pp. 352–363.

[21] M. Rigo, Formal Languages, Automata and Numeration Systems, ISTE-Wiley, 2014.

[22] V. Berthé, M. Rigo, Odometers on regular languages, Theory Comput. Syst. 40 (1) (2007) 1–31.

[23] V. Berthé, A. Siegel, J. Thuswaldner, Substitutions, Rauzy fractals and tilings, Ch. 5, in: Berthé and Rigo [26], pp. 248–323.

[24] M. Drmota, P. J. Grabner, Analysis of digital functions and applications, Ch. 9, in: Berthé and Rigo [26], pp. 452–504.

[25] J. Honkala, The equality problem for purely substitutive words, Ch. 10, in: Berthé and Rigo [26], pp. 505–529.

[26] V. Berthé, M. Rigo (Eds.), Combinatorics, Automata and Number Theory, no. 135 in Encyclopedia Math. Appl., Cambridge University Press, 2010.