Breadth-first serialisation of trees and rational languages

Victor Marsault[*,†] and Jacques Sakarovitch[†]

2014 – 04 – 03

Abstract

We present here the notion of *breadth-first signature* and its relationship with numeration system theory. It is the serialisation into an *infinite word* of an ordered infinite tree of finite degree. We study which class of languages corresponds to which class of words and, more specifically, using a known construction from numeration system theory, we prove that the signature of rational languages are substitutive sequences.

1 Introduction

This work introduces a new notion: the 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 tree to word, or conversely), it is either a *serialisation* of the tree into an infinite word or a *generation* of the tree by the word. We study here the serialisation of rational, or regular, languages.

The (breath-first) signature of an ordered tree of finite degree is a sequence of integers, 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 uniquely defined and characteristic of the tree.

Similarly, we call *labelling* the infinite sequence of the labels of the edges visited by the breadth-first traversal of a labelled 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.

The serialisation of a (prefix-closed) language 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 describing 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

[*] Corresponding author, victor.marsault@telecom-paristech.fr

[†] LTCI, CNRS / Telecom ParisTech

signature the Fibonacci word (*cf.* Section 4); and 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. In another work still in preparation [10], we study trees and languages that have periodic signatures.

In the present work, we first study in detail the notion of signature of trees (Section 2) and of languages (Section 3). Then, in Section 4, we give with Theorem 14 a characterisation of the signatures of (prefix-closed) rational languages as those whose signature is a substitutive sequence. The proof of this result relies on a correspondence between substitutive sequences and automata due to Maes and Rigo [12] and whose principle goes back indeed to the work of Cobham [4].

2 On trees

Classically, trees are undirected graphs in which any two vertices are connected by exactly one path (*cf.* [6], for instance). Our view differs in two respects.

First, a tree is a *directed* graph $T = (V, \Gamma)$ such that there exist a *unique* vertex, called *root*, which has no incoming arc, and there is a *unique (oriented) path* from the root to every other vertex. Elements of the tree T gets particular names: vertices are called *nodes*; if (x, y) is an arc, y is called *a child* of x and x *the father* of y; a node without children is a *leaf*. We draw trees with the root on the left and arcs rightward.

Second, our trees are *ordered*, that is, that there is a total order on the set of children of every node. The order will be implicit in the figure, with the convention that lowermost children are the smallest (according to this order). In one word, the two trees of Figure 1 are different (non-isomorphic).

The class of trees we consider is quite close to the one from [5], but our approach differs greatly. They generate trees through tree automata, a depth-first process while we describe them in a breadth-first manner.

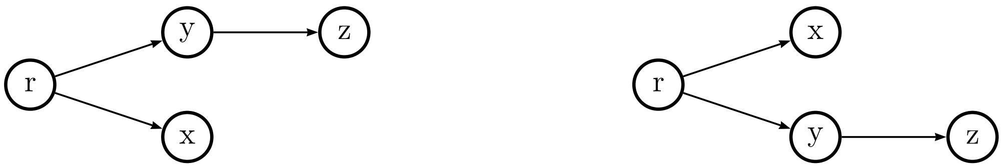


Figure 1: Two non-isomorphic trees

The *degree* of a node is the number of its children; it may be finite or infinite. A tree is of *bounded degree* (resp. *finite degree*) if the degree of every node is bounded (resp. finite). In the following, we deal with infinite trees of bounded degree only. Even though most definitions would still work for infinite trees of finite degree, this more general setting has no use when considering languages, as we will in most of the article.

2.1 Relational definition of trees

Given a particular tree, the breadth-first traversal naturally and uniquely (since the children of every node are ordered) defines a total ordering of its nodes. We may then consider that the set of nodes of a tree is always the set of integers $\mathbb{N}$; the node n of $\mathbb{N}$ being the $(n+1)$-th node visited by the traversal.

Proposition 1. *A directed graph* $(\mathbb{N},\theta)$*, where the relation* $\theta : \mathbb{N} \to \mathbb{N}$ *satisfies the two conditions*
(i) θ *is injective;*
(ii) $\forall n \in \mathbb{N},\ \exists m \in \mathbb{N}, \qquad m > n \quad$ *and* $\quad \theta(\llbracket 0,n \rrbracket) = \llbracket 1,m \rrbracket$*;*
is an infinite ordered tree of finite degree, written T_θ*.*

Proof. In this setting, θ is the child relation, 0 is the root, $\theta(0) = \theta(\llbracket 0,0 \rrbracket) = \llbracket 1,k \rrbracket$, is an interval of $\mathbb{N}$; it is the (finite) ordered set of the $k > 0$ children of the root. Given a *positive* integer n,

$$\theta(n) = \theta(\llbracket 0,n \rrbracket) \smallsetminus \theta(\llbracket 0,n-1 \rrbracket)$$

is the (possibly empty) interval of $\mathbb{N}$ of the children of the node n.
Hence the *father relation* θ^{-1} satisfies the following properties:
1. θ^{-1} is a function — from (i);
2. $\mathrm{Dom}(\theta^{-1}) = \mathbb{N}_+$ — from (ii);
3. $\theta^{-1}(n) < n$ — from (ii).

It then yields a unique path (in θ^*) from the root to every vertex in $\mathbb{N}_+$ □

Computing the relation from the tree. A breath-first traversal of an infinite ordered tree T of finite degree inductively maps the set of nodes of T onto $\mathbb{N}$ and builds a child relation θ by the following procedure whose principle is essential.

The root of T is mapped onto 0, the ordered set of the k children of the root is mapped onto the interval $\llbracket 1,k \rrbracket$, that is $\theta(0) = \llbracket 1,k \rrbracket$ and two integer indices are set: the first one represents *the node to be treated*, call it n, and is set to 1; the second one represents *the last node created*, call it m, and is set to k. At every step of the procedure the node n is considered the ordered set of its k_n children is mapped onto the interval $\llbracket m+1,\ m+k_n \rrbracket$, that is $\theta(n) = \llbracket m+1,\ m+k_n \rrbracket$ (possibly empty if $k_n = 0$); then n is incremented by 1, m by k_n, and the procedure takes on a new step.

Since T is of finite degree, each step is well-defined and since T is infinite, the procedure never ends. Nevertheless, $\theta(n)$ is eventually defined for every integer n. The way it is defined makes θ meet Conditions (i) and (ii) of Proposition 1 and the tree T_θ is isomorphic to T.

On i-trees. It will prove to be extremely 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. It amounts to changing the Condition (ii) of the child relation θ to

(ii') $\forall n \in \mathbb{N},\ \exists m \in \mathbb{N}, \qquad m > n \quad$ and $\quad \theta(\llbracket 0,n \rrbracket) = \llbracket 0,m \rrbracket$;

the difference being that the interval $[\![1,m]\!]$ of (ii) is changed to $[\![0,m]\!]$ in (ii').

It should be noted that this convention is sometimes taken when implementing tree-like structures (for instance the unix/linux file system). It implies that the father relation θ^{-1} is now a *function* $\mathbb{N} \to \mathbb{N}$ and will make the connexion with numeration systems very natural.

Of course, a graph T_θ defined by a relation θ that meets Condition (i) and (ii') is not formally a tree; we call such structures *i-trees*. It is so close to a tree that we pass from tree to i-tree (or conversely) with no further ado.

2.2 Breadth-first signature of a tree

Definition 2 (Breadth-first signature of a tree)**.** *Given a tree of child relation θ, we call* breadth-first signature *or, for short,* signature *of θ the infinite integer sequence*

$$\boldsymbol{s} = s_0 s_1 \cdots s_k \cdots \qquad \textit{where } s_i = \texttt{Card}(\theta(i)) \qquad \textit{for all integers } i > 0 \tag{1a}$$

$$\textit{and } s_0 = \texttt{Card}(\theta(0)) + 1 \tag{1b}$$

where $\texttt{Card}(X)$ is the cardinal of the set X. It follows directly from this definition that the breadth-first signature is characteristic of its tree, as stated below.

Proposition 3. *Two trees with the same breadth-first signature are equal.*

The special case of 0 (*cf.* Equation (1b)) is an artefact of the non-surjectivity of θ already discussed in the previous Section 2.1. For short, the signature is more canonically associated with an i-tree than with the corresponding tree.

2.3 Generating a tree by its signature

A signature $\mathbf{s} = s_0\, s_1 \cdots s_k \cdots$ is *valid* if it satisfies the following equation

$$\forall j \in \mathbb{N} \qquad \sum_{i=0}^{j} s_i \ > \ j{+}1 \ . \tag{2}$$

This restriction ensures that the sequence is indeed the signature of a tree, as stated below; if it were not the case, one could still apply the procedure hereafter, but the resulting graph would not be connected. [1]

Proposition 4. *For every valid signature $\boldsymbol{s}$, there exists a unique tree whose signature is equal to $\boldsymbol{s}$.*

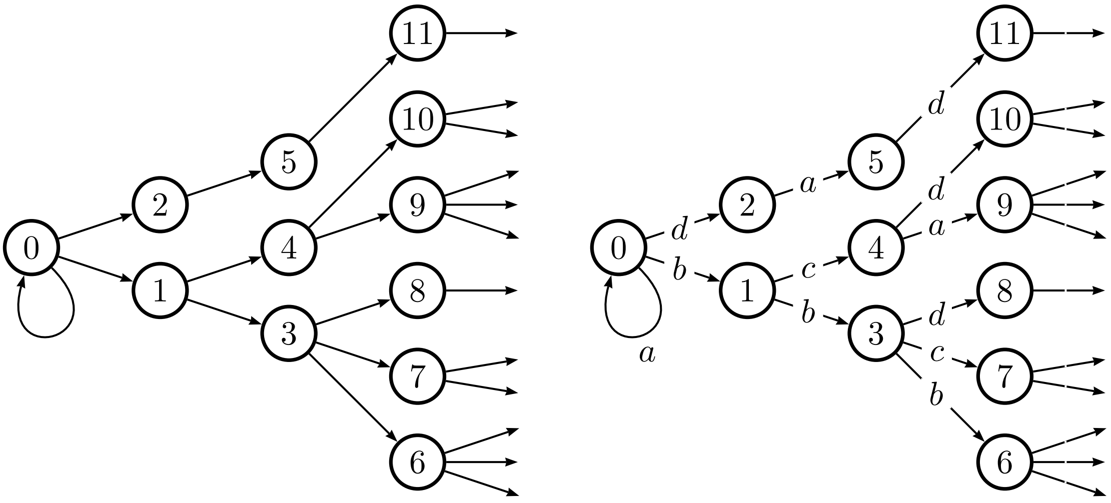
We will describe the tree whose signature is equal to a given signature $\boldsymbol{s}$ by enumerating its edges in the breadth-first order. It is essentially the reverse as the construction of the relation θ from the respective tree, given at Section 2.1

We maintain two integers: the starting point n and the end point m of the transition. In one step of the algorithm, s_n nodes are created, corresponding to

[1] Equation (2) is the counterpart of the '$m > n$' condition in Proposition 1(ii).

the integers $m, m+1, \ldots, (m+s_n-1)$, and s_n edges are created (all from n, and one to each of this new nodes). Then n is incremented by 1, and m by s_n.

The validity of $\boldsymbol{s}$ ensures that, at all point $n < m$, hence that every node has a father smaller than itself. Figure 3, in appendix, shows the first few steps of the procedure for the purely periodic signature $(321)^\omega$, while Figure 2a shows the resulting i-tree.



(a) The i-tree $T_{\boldsymbol{s}}$

(b) $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$, *with* $\boldsymbol{\lambda} = abd\,.\,bc\,.\,a\,.\,bcd\,.\,ad\,.\,d\,\ldots$

Figure 2: I-tree generated by the signature $\boldsymbol{s} = (321)^\omega$

3 Signature for languages

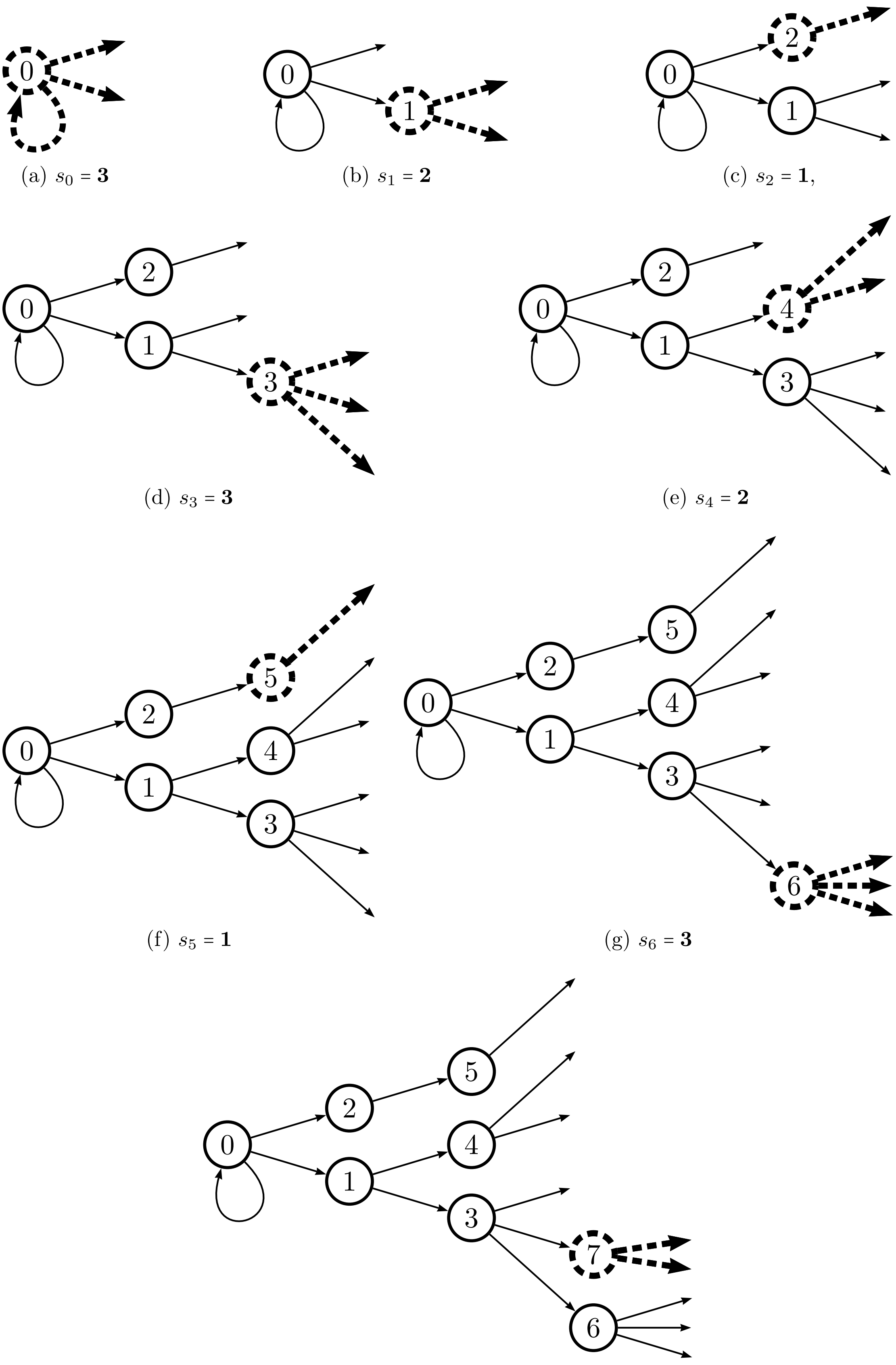
An *alphabet* is a set of *letters* and will always be ordered in the following. Whenever we use a latin or digits alphabet, it will be ordered as usual (that is, $a < b < c < \cdots$ or $0 < 1 < 2 < \cdots$). A *word* w is a finite sequence of letters $a_0\, a_1 \cdots a_{n-1}$ and its length is denoted by $|w| = n$.

3.1 Labelling

A labelling, together with a signature $\boldsymbol{s}$, is the description of a labelled tree (that is, essentially a prefix-closed language). It corresponds to the sequence of the transition labels of the tree, taken in breadth-first order. It follows that a *labelling* is simply a sequence of letters of some alphabet.

However, for a labelled tree to effectively represent a (prefix-closed) language, it must satisfies some properties. For instance, two edges with the same starting point must have distinct labels. More generally, the labels must be consistent with the breadth-first traversal: an edge to a *smaller* child must be labelled by a *smaller* letter. The notion validity for a labelling subdues these issues.

Given a signature $\boldsymbol{s}$, a labelling $\boldsymbol{\lambda}$ over an alphabet A is *valid* (with respect to $\boldsymbol{s}$) if there exists a family $\{w_k\}_{k \in \mathbb{N}}$ of words over A such that



(a) s_0 = **3**

(b) s_1 = **2**

(c) s_2 = **1**,

(d) s_3 = **3**

(e) s_4 = **2**

(f) s_5 = **1**

(g) s_6 = **3**

(h) s_7 = **2**

Figure 3: The first eight steps of the generation of $T_{(321)^\omega}$

1. $\boldsymbol{\lambda}$ is the concatenation of the family $\{w_k\}_{k\in\mathbb{N}}$:
$$\boldsymbol{\lambda} = w_0\, w_1 \cdots w_k \cdots \quad ;$$
2. the length of each word w_k is equal to s_k:
$$\forall k \in \mathbb{N} \qquad |w_k| = s_k \quad ;$$
3. the letters of each word w_k are in strictly increasing order:
$$\forall w_k = a_0 a_1 \cdots a_n \qquad a_0 < a_1 < a_2 < \cdots < a_n \quad .$$

For instance if the signature starts with $3\,2\,1\,3\cdots$, a valid labelling could start with $abd\,.\,bc\,.\,a\,.\,bcd\ldots$; or with $012\,.\,01\,.\,0\,.\,012\ldots$ A pair signature/labelling $(\boldsymbol{s}, \boldsymbol{\lambda})$ is called a *labelled signature*; it is *valid* if both $\boldsymbol{s}$ is valid and $\boldsymbol{\lambda}$ is valid (w.r. to $\boldsymbol{s}$).

A valid labelled signature $(\boldsymbol{s}, \boldsymbol{\lambda})$ uniquely defines a labelled tree, by using a procedure analogous to the one from Section 2.3. Every edge $i \longrightarrow j$ created is labelled by λ_j. For every node n, we denote by $\langle n\rangle_{(\boldsymbol{s},\boldsymbol{\lambda})}$ the word labelling the unique path $0 \longrightarrow n$. We denote by $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ the language of such words: [2]
$$L_{(\boldsymbol{s},\boldsymbol{\lambda})} = \{\langle n\rangle_{(\boldsymbol{s},\boldsymbol{\lambda})} \mid n \in \mathbb{N}\} \quad .$$

Figure 2b, page 5, shows the language whose signature is $\boldsymbol{s} = (321)^\omega$ and labelling starts with $\boldsymbol{\lambda} = abd\,.\,bc\,.\,a\,.\,bcd\,.\,ad\,.\,d\ldots$ The validity of the labelled signature insures that the words $\langle n\rangle_{(\boldsymbol{s},\boldsymbol{\lambda})}$ are all distinct, hence the following lemma.

Lemma 5. *The $(n+1)$-th word of $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ in radix order is $\langle n\rangle_{(\boldsymbol{s},\boldsymbol{\lambda})}$*

Conversely, given any prefix-closed language L over an alphabet A, there is a unique valid labelled signature $(\boldsymbol{s}, \boldsymbol{\lambda})$ generating it; $\boldsymbol{s}$ is defined by the underlying tree of L and $\boldsymbol{\lambda}$ is the sequence of the labels of the edges of the underlying tree of L when taken in breadth-first order. The next statement follows immediately.

Proposition 6. *For every valid labelled signature $(\boldsymbol{s}, \boldsymbol{\lambda})$, there exists a unique language whose signature is equal to $(\boldsymbol{s}, \boldsymbol{\lambda})$.*

3.2 Minimal labelling and rational trees

We call *minimal labelling* of a signature $\boldsymbol{s}$ (or equivalently of a tree $T_{\boldsymbol{s}}$) the labelling induced by the order of children:
$$\boldsymbol{\mu} = w_0\, w_1 \cdots w_k \cdots \qquad \text{where} \quad \forall k \in \mathbb{N} \quad w_k = 0\,1\,2 \cdots n \quad \text{and} \quad n = (s_k - 1) \quad .$$

Intuitively, it corresponds to add labels in the tree such that the transition $n \xrightarrow{a} m$ is labelled by $a = 0$ if m is the *smallest* child of n, and that the transition $n \xrightarrow{b} (m+k)$ is labelled by $b = k$, if it exists. It is always possible to label a tree in such a way and it produces a valid labelled signature. Intuitively, the minimal labelling is the simplest way to label a tree, in the sense that it adds the less possible complexity. The next lemma gives an example of this intuition.

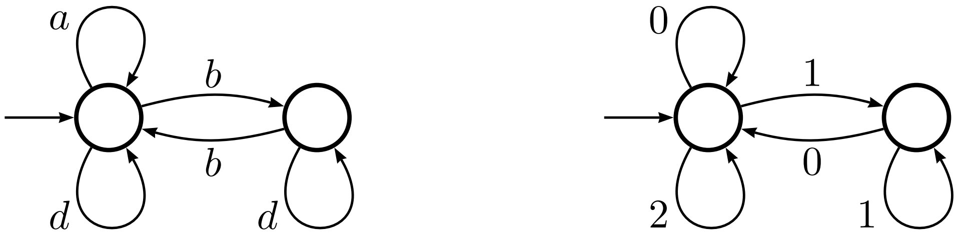
[2] This process closely related to the creation of an abstract numeration systems (*cf.* [8]) which takes a language L (usually assumed to be rational) and set the representation of n in the new numeration system as the $(n+1)$-th word of L in radix order.

Lemma 7. *Let $(\boldsymbol{s}, \boldsymbol{\lambda})$ be a valid labelled signature, and $\boldsymbol{\mu}$ the minimal labelling associated with $\boldsymbol{s}$. If $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ is a rational language, then $L_{(\boldsymbol{s},\boldsymbol{\mu})}$ is rational as well.*

Proof. Given the finite deterministic trim automaton $\mathcal{A} = \langle A, Q, \delta, i \rangle$ accepting $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$, let us consider the automaton $\mathcal{B} = \langle B, Q, \delta', i \rangle$ where

- $B = [\![0, k]\!]$ with $k = \texttt{Card}(A)$
- $p \xrightarrow[\mathcal{B}]{i} q$ iff $p \xrightarrow[\mathcal{A}]{b} q$ and there exists exactly i letters a of A such that
  - $a < b$
  - $p \xrightarrow[\mathcal{A}]{a} q'$ for some state q'

Intuitively, one has to change the labels of the outgoings transitions of every states by the smallest possible (in $[\![0, k]\!]$) without modifying their relative order. For instance if a state p of $\mathcal{A}$ has three outgoings transitions labelled by a, c and d; then in the automaton $\mathcal{B}$, the same state p would have the same transitions but now respectively labelled by 0, 1 and 2 (provided that the order of A is $a < c < d$). See Figure 4, below, for an example. Unfolding automata $\mathcal{A}$ and $\mathcal{B}$ into infinite labelled trees yields the statement. □



(a) An automaton $\mathcal{A}$ (b) The respective automaton $\mathcal{B}$

Figure 4: Minimal labelling

Remark 8. *It should be noted that even if a signature produces a really simple tree (such as the infinite unary tree), one can always choose a labelling in order to produce an artificially complex language (such as the infinite word where the i-th letter is a 1 if the i-th Turing machine stops on the empty word).*

This is why positive results relative to the regularity of a language defined by signature will always require some restriction on the labelling. It usually amounts to ensure that signature and labelling are generated in similar fashions. For instance, it will be the case for substitutive labelled signature *defined in the next Section 4.*

4 Substitutive signature and rational languages

The purpose of this section is to establish a relationship between substitutive sequences and rational languages. Let us first consider the Fibonacci word $\sigma^\omega(0)$ where $\sigma(0) = 01$ and $\sigma(1) = 0$:

$$\sigma^\omega(0) = 0100101001001\cdots$$

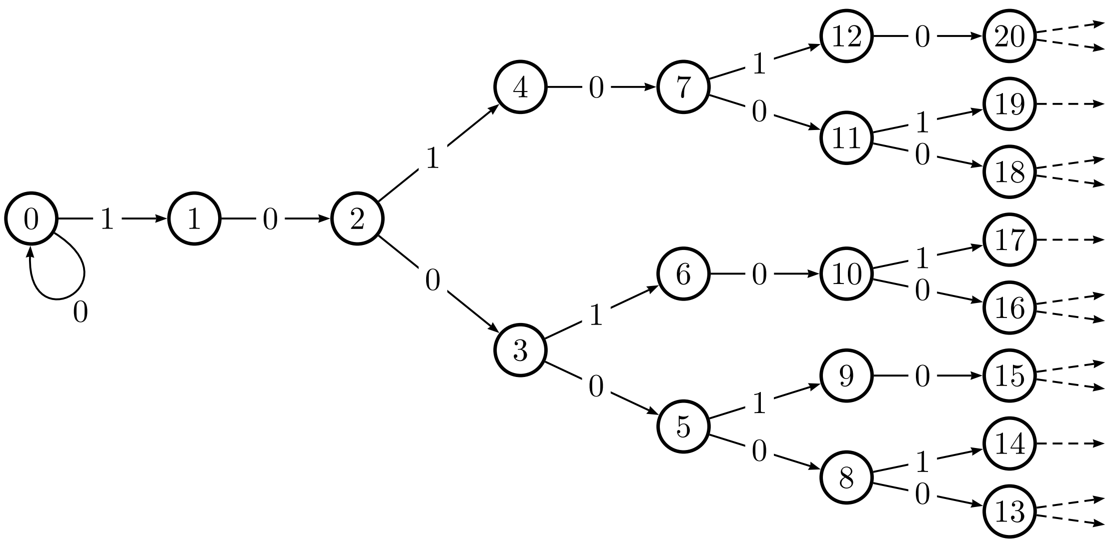


Figure 5: The language of the integer representations in the Fibonacci number system

This word, however, is not valid when considered as a signature. We build a valid signature $\boldsymbol{s}$ by replacing the 0's in the Fibonacci word by 2's:

$$\boldsymbol{s} \quad = \quad 2\,1\,2\,2\,2\,1\,2\,1\,2\,2\,1\,2\,2\,1\cdots$$

It should be noted that the labelling $\boldsymbol{\lambda} = \sigma^{\omega}(0)$ is valid w.r. to $\boldsymbol{s}$: each letter '2' (resp '1') of $\boldsymbol{s}$ is associated to the word '01' (resp '0') of $\boldsymbol{\lambda}$.

The language $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ shown at Figure 5 is then exactly the integer representations in the Fibonacci numeration system (sometimes called Zeckendorf numeration system) that is, the rational language $1\{0,1\}^* \setminus (\{0,1\}^*11\{0,1\}^*)$.

4.1 Substitutive sequences and substitutive signatures

We recall here some basic definition from combinatorics on words; we essentially use the terminology of [3].

Given an alphabet A, we say that an endomorphism $\sigma : A^* \to A^*$ is *prolongable* on $a \in A$ if there exists a word u of A^* such that $\sigma(a) = au$ and that $lim_{n\to+\infty}|\sigma^n(a)| = +\infty$. In this context, the sequence $\sigma^n(a)$ converges (for the usual topology) to an infinite sequence denoted by $\sigma^{\omega}(a)$. Any sequence resulting from the iteration of a prolongable endomorphism (that is, of the form $\sigma^{\omega}(a)$ where σ is prolongable on a) is said to be *purely substitutive*.

A letter-to-letter morphism is called a *coding*[3]. The image $f(w)$ of a purely substitutive sequence w by a morphism f is called an *HD0L* sequence; if furthermore, f is a coding, $f(w)$ is called a *substitutive sequence*.

We will now define particular substitutive sequences and consider them as signatures. Given an endomorphism $\sigma : A^* \to A^*$ prolongable on a letter $a \in A$, we

[3] Note that a *coding does not define a code*, in the sense of [2].

denote by f_σ the following coding entirely defined by σ:

$$\forall a \in A \qquad f_\sigma(a) = |\sigma(a)| \ .$$

We call the substitutive sequence $f_\sigma(\sigma^\omega(a))$ a *substitutive signature.*

Lemma 9. *Every substitutive signature is valid.*

Proof. It amounts to prove that for all prefixes u of $f_\sigma(\sigma^\omega(a))$, the sum of the letters of u is strictly greater than the length of u. Hence, from the definition of f_σ, that for all prefixes v of $\sigma^\omega(a)$, $|\sigma(v)| > |v|$.

Let v be any prefix of $\sigma^\omega(a)$. Since σ is prolongable on the letter a, there is an integer i such that $\sigma^i(a) \sqsubseteq v \sqsubset \sigma^{(i+1)}(a)$; hence $\sigma^{(i+1)}(a) \sqsubseteq \sigma(v) \sqsubset \sigma^{(i+2)}(a)$, hence $|\sigma(v)| \geqslant |\sigma^{(i+1)}(a)| > |v|$.

□

Definition 10. *A labelled signature $(\boldsymbol{s}, \boldsymbol{\lambda})$ is* substitutive *if*

- *$\boldsymbol{s}$ is a substitutive signature $f_\sigma(\sigma^\omega(a))$ and*
- *$\boldsymbol{\lambda}$ is of the form $g(\sigma^\omega(a))$ where $g : A^* \to B^*$ and for all letters $a \in A$, $|g(a)| = |\sigma(a)|$.* [4]

The next lemma follows; its proof is essentially the same as the one of Lemma 9.

Lemma 11. *Every substitutive labelled signature is valid.*

We open now a parenthesis about ultimately periodic signature. Let $\boldsymbol{s} = uv^\omega$ be an ultimately periodic sequence over the alphabet $[\![0, k-1]\!]$; we call *growth ratio* of v, denoted $gr(v)$, the average of the letters of v:

$$gr(v) \quad = \quad \frac{\sum_{i=0}^{|v|-1} v[i]}{|v|} \ .$$

The next proposition states that whenever $gr(v)$ is an integer that is, when the sum of the letters of v is a multiple of the length of v, any signature of the form uv^ω is substitutive.

Proposition 12. *Given an ultimately periodic (valid) signature $\boldsymbol{s} = uv^\omega$ whose growth ratio is an integer then $\boldsymbol{s}$ is a substitutive signature.*

Proof. First, let us consider the case where $u = \varepsilon$. We denote by k the length of v: $k = |v|$; and consider the alphabet $A = [\![0, k-1]\!]$. In the following, any letter (for instance of the form $(j+h)$ for some integers j and h) will be taken in $\mathbb{Z}/k\mathbb{Z}$, hence will belong to A. We define the endomorphism $\sigma : A^* \to A^*$ by

$$\begin{array}{ll} \sigma(0) = 0\,1 \cdots (v[0]-1) & \text{and} \\ \sigma(i) = (j+1)\,(j+2) \cdots (j+v[i]) & \text{where } j \text{ is the last letter of } \sigma(i-1). \end{array}$$

Let us now prove that $\sigma(0\,1 \cdots (k-1)) = (0\,1 \cdots (k-1))^j$, where j is the growth ratio of v. It is quite easy to see that $\sigma(0\,1 \cdots (k-1))$ is a prefix of $(0\,1 \cdots (k-1))^\omega$

[4] A substitutive labelling is then a particular HD0L sequence.

since $\sigma(i+1)$ starts with the letter directly following the last letter of $\sigma(i)$. Since by definition, the length of $\sigma(i)$ is equal to $v[i]$ then

$$|\sigma(0\,1\cdots(k-1))| = \sum_{i=0}^{|v|-1} v[i] = j \times |v| = j \times k$$

yielding the claim.

It follows that $\sigma^\omega(0\,1\cdots(k-1))$ is equal to $(0\,1\cdots(k-1))^\omega$. With a similar reasoning one can prove that $\sigma^\omega(0)$ is also equal to $(0\,1\cdots(k-1))^\omega$. Finally, since for all $i \in [\![0, k-1]\!]$, $|\sigma(i)| = v[i]$, it follows that $f_\sigma(0\,1\cdots(k-1)) = v$ and $f_\sigma(\sigma^\omega(0)) = f_\sigma((0\,1\cdots(k-1))^\omega) = v^\omega$, concluding the special case $u = \varepsilon$.

We no longer assume that $u = \varepsilon$ and then denote by n the length of u. The new alphabet of the morphism is $C = A \uplus B$, where B is of cardinal n, each letter corresponding to a position in u. We denote the letters by :

$$B = \{b_0, b_1, \ldots, b_{(n-1)}\} \qquad \text{and} \qquad A = \{a_0, a_1, \ldots, a_{(k-1)}\}$$

The images of the letters of B by σ are defined inductively: for every integer i, $\sigma(b_0\,b_1\cdots b_i)$ is the prefix of the sequence $(b_0\,b_1\cdots b_{n-1})(a_0\,a_1\cdots a_{(k-1)})^\omega$ such that $|\sigma(b_i)| = u[i]$.

Since the signature uv^ω is valid by hypothesis, the last letter of $\sigma(b_{n-1})$ is some letter of A; the image by σ of the letters of A are then:

$$\begin{array}{rll} \sigma(a_0) &= a_{(j+1)}\,a_{(j+2)}\cdots a_{(j+v[0])} & \text{where } a_j \text{ is the last letter of } \sigma(b_{(n-1)}). \\ \sigma(a_i) &= a_{(j+1)}\,a_{(j+2)}\cdots a_{(j+v[i])} & \text{where } a_j \text{ is the last letter of } \sigma(a_{(i-1)}). \end{array}$$

From here, the proof is the analogous to the case where $u = \varepsilon$.

□

Remark 13. *It should be noted that whenever the growth ratio of an ultimately periodic signature is not an integer, it is never substitutive. The proof of this statement is however convoluted and is the subject of another article in preparation [10].*

It should also be noted that ultimately periodic signatures are, as words, purely substitutive *no matter the growth ratio. For instance the word* $(21)^\omega$ *is equal to* $\sigma^\omega(2)$ *where* $\sigma(2) = \sigma(1) = 21$*. It illustrates the fact that the set of purely substitutive sequences is not included in the set of substitutive signatures.*

4.2 Rational languages and substitutive signatures

Theorem 14. *A prefix-closed language is rational if and only if its labelled signature is a substitutive signature.*

The proof of this theorem relies on a transformation from finite automaton to substitutive word used by Rigo and Maes in [12] (*cf.* also [8, Section 3.4]) to prove the equivalence between two decision problems: 1- the ultimate periodicity in an abstract numeration system (*cf.* [9] or [8]) and 2- the ultimate periodicity problem of an HD0L word (solved independently in [11] and [7]).

Automaton	Substitutive signature
$\langle\, \Sigma, Q, \delta, i \,\rangle$	$\boldsymbol{s} = f_\sigma(\sigma^\omega(a))\ \ \boldsymbol{\lambda} = g(\sigma^\omega(a))$
	$\sigma : A^* \to A^*$
	$g :\ A \to B$
Q	A
i	a
Σ	B
$(b, x, c) \in \delta$	the k-th letter of $\sigma(b)$ is c the k-th letter of $g(b)$ is x

Table 6: Summary of the transformation DFA ↔ Substitutive signature

Proposition 15. *Given a valid substitutive signature $(\boldsymbol{s}, \boldsymbol{\lambda})$, the language $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ is a rational language.*

Proof. We denote by σ the endomorphism $A^* \to A^*$ prolongable on a letter a of A; and by g the projection $A^* \;\to\; B^*$ such that

$$\boldsymbol{s} = f_\sigma(\sigma^\omega(a)) \qquad \text{and} \qquad \boldsymbol{\lambda} = g(\sigma^\omega(a)) \ .$$

Since we are using two alphabets at the same time we will, in this proof, consider that a, b, c are letters of A and x is a letter of B.

We consider the automaton $\mathcal{A} = \langle\, A, B, \delta, a \,\rangle$, whose set of **state** is equal to A; the alphabet is equal to B; the initial state is the letter a, all states are accepting; and the transition function is defined by:

$$b \xrightarrow{x} c \qquad \text{if there exists } i \text{ such that } \begin{cases} c \text{ is the } i\text{-th letter of } \sigma(b) \\ x \text{ is the } i\text{-th letter of } g(b) \end{cases}$$

(*cf.* Table 6 for a summary of this transformation).

Note that there is loop on the initial state a, since the morphism f is prolongable on a; we denote by x the label of this loop, that is, the smallest letter of $g(a)$. This loops corresponds to the usual 0-loop differentiating i-trees from trees and we will consider in the following the language $L = L(\mathcal{A}) \bigcap ((B \smallsetminus x).B^*)$ where the loop is removed *on the root only.*

Let us prove that the labelled signature of L is equal to $(\boldsymbol{s}, \boldsymbol{\lambda})$. We denote by $(u_i)_i$ the following sequence of words (over A) $u_0 = a$, $\quad u_1 = a^{-1}\sigma(a)$ and for all $\;i > 0$, $u_{i+1} = f(u_i)$. It follows that $u_0 u_1 \cdots u_i = \sigma^i(a)$.

Let us fix an i and consider the words of L of length i in radix order; we denote by w_k the $(k+1)$-th word of L of length i. It can be easily proven (by induction over i) that

$$\forall k \in \mathbb{N} \qquad a \xrightarrow[\mathcal{A}]{w_k} c_k \qquad \text{where } c_k \text{ is the } k\text{-th letter of } u_i \ .$$

It follows that the $(k+1)$-th word (of any length) of L in radix order reaches in the automaton $\mathcal{A}$, the k-th letter of $\sigma^\omega(a)$. Since the outgoing transitions of a state $b \in A$ are labelled by the letters of the words $g(b) \in B^*$, it follows that the labelled signature of L is equal to $(\boldsymbol{s}, \boldsymbol{\lambda})$. □

Proposition 16. *The labelled signature of a prefix-closed rational language is a substitutive signature.*

Proof. Let L be a rational language and a finite minimal deterministic trim automaton $\mathcal{A} = \langle \Sigma, Q, \delta, i \rangle$ accepting the language $\#^* L$, '#' being a letter which does not appear in L, and is fixed as smaller than every other letter. Reusing the transformation summed up in Table 6, we define the morphisms:

$$\begin{array}{rccl} \sigma : & Q^* & \longrightarrow & Q^* \\ & p & \longmapsto & q_0\, q_1\, q_2 \cdots q_k \end{array} \qquad \begin{array}{rccl} g : & Q^* & \longrightarrow & \Sigma^* \\ & p & \longmapsto & a_0\, a_1\, a_2 \cdots a_k \end{array}$$

where $a_0 < a_1 < \cdots < a_k$, and for all $i \in [\![0, k]\!]$, $p \xrightarrow[\mathcal{A}]{a_i} q_i$. It follows from definition that σ is prolongable on the letter $i \in Q$, since it corresponds to the initial state of $\mathcal{A}$ on which there is necessarily a loop labelled by $\#$.

From there, the proof is essentially the same as the one from Proposition 15. □

5 Conclusion and future work

In this work, we introduced a way of effectively describing infinite trees and languages by infinite words using a simple breadth-first traversal. Since this transformation is essentially one-to-one, it is natural to wonder which class of words is associated with which class of languages and which properties of the former can be translated into properties of the latter.

In this first work on the subject, we have proved that rational languages are associated with (a particular subclass of) substitutive words.

In a forthcoming paper [10], we study the class of languages associated with periodic signatures and how they are related to the representation language in rational base numeration systems. In both cases these results express the intimate relationship between signatures and numeration systems.

Our aim is to further explore this relationship by means of the notion of direction of a signature, that generalises the notion of growth ratio given at Section 4. For instance, a rational base numeration system has a signature deduced from the Christoffel word associated with that rational number.

References

[1] Shigeki Akiyama, Christiane Frougny, and Jacques Sakarovitch. Powers of rationals modulo 1 and rational base number systems. *Israel J. Math.*, 168:53–91, 2008.

[2] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*, volume 129. Cambridge University Press, 2009.

[3] Valérie Berthé and Michel Rigo. *Combinatorics, Automata and Number Theory.* Number 135 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2010.

[4] Alan Cobham. Uniform tag sequences. *Math. Systems Theory*, 6:164–192, 1972.

[5] H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. Tree automata techniques and applications. Available on: http://www.grappa.univ-lille3.fr/tata, 2007. release October, 12th 2007.

[6] Reinhard Diestel. *Graph Theory.* Springer, 1997.

[7] Fabien Durand. Decidability of the HD0L ultimate periodicity problem. *RAIRO - Theor. Inf. and Applic.*, 47(2):201–214, 2013.

[8] Pierre Lecomte and Michel Rigo. Abstract numeration systems. in *Combinatorics, Automata and Number Theory*, V. Berthé, M. Rigo (Eds), Encyclopedia of Mathematics and its Applications 135, Cambridge Univ. Press (2010) 108–162.

[9] Pierre Lecomte and Michel Rigo. Numeration systems on a regular language. *Theory Comput. Syst.*, 34:27–44, 2001.

[10] Victor Marsault and Jacques Sakarovitch. Rhythmic generation of infinite trees and languages (early version). *arXiv preprint:1403.5190*, 2014.

[11] Ivan Mitrofanov. A proof for the decidability of HD0L ultimate periodicity. *arXiv preprint arXiv:1110.4780*, 2011.

[12] Michel Rigo and Arnaud Maes. More on generalized automatic sequences. *Journal of Automata, Languages and Combinatorics*, 7(3):351–376, 2002.

Appendix: Some More Figures

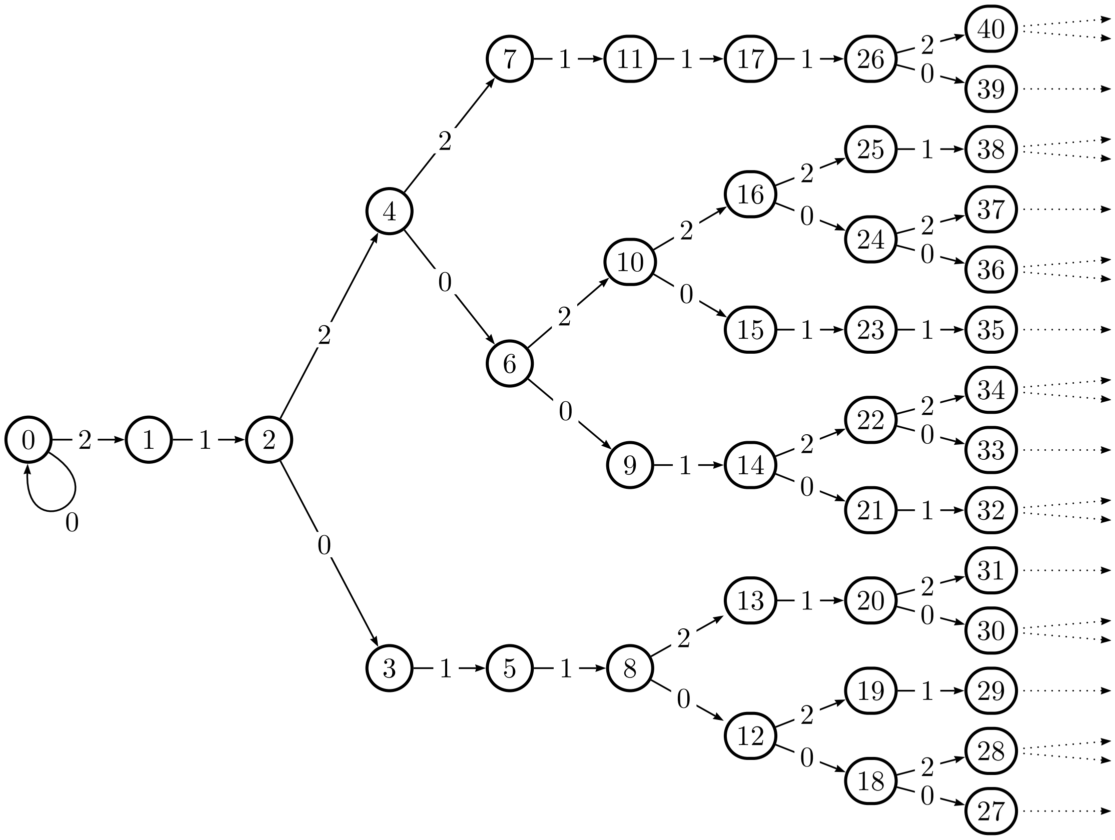


Figure A.7: The language of the representations of integers in base $\frac{3}{2}$, its signature is $(21)^\omega$ and its labelling is $(021)^\omega$.

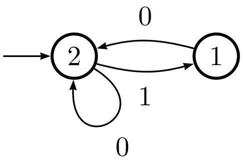


Figure A.8: The automaton accepting the representations of integers in base Fibonnaci