Breadth-First Serialisation of Trees and Rational Languages
(Short Paper)

Victor Marsault* and Jacques Sakarovitch

Telecom-ParisTech and CNRS, 46 rue Barrault 75013 Paris, France
victor.marsault@telecom-paristech.fr

Abstract. We present here the notion of *breadth-first signature* of trees and of prefix-closed languages; and its relationship with numeration system theory. A signature is the serialisation into an *infinite word* of an ordered infinite tree of finite degree. Using a known construction from numeration system theory, we prove that the signature of (prefix-closed) rational languages are substitutive words and conversely that a special subclass of substitutive words define (prefix-closed) rational languages.

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 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 [2], we study trees and languages that have periodic signatures.

* Corresponding author.

A.M. Shur and M.V. Volkov (Eds.): DLT 2014, LNCS 8633, pp. 252–260, 2014.

In the present work, we first introduce the notion of signature of trees (Section 2) and of languages (Section 3). Then, in Section 4, we give with Theorem 1 a characterisation of the signatures of (prefix-closed) rational languages as those whose signature is substitutive. The proof of this result relies on a correspondence between substitutive words and automata due to Maes and Rigo [3] or Dumont and Thomas [4] and whose principle goes back to the work of Cobham [5].

2 Signatures of 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 $\mathcal{T} = (V, \Gamma)$ such that there exists 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 a tree get 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. We draw trees with the root on the left, and arcs rightwards.

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 figures, with the convention that lower children are smaller (according to this order).

The *degree* $\mathsf{d}(x)$ of a node is the number of children of x. A breadth-first traversal of a tree $\mathcal{T}$ eventually meets every node of $\mathcal{T}$ if and only if all degrees are finite. In the following, and as we are interested in trees in relation with infinite languages (over finite alphabets), we deal with infinite trees of bounded degree only. Since trees are ordered, there is a canonical breadth-first traversal for every tree. We may then consider that the set of nodes of a tree is always the set of integers $\mathbb{N}$: 0 is the root and the integer i is the $(i+1)$-th node visited by *the* breadth-first traversal of the tree.

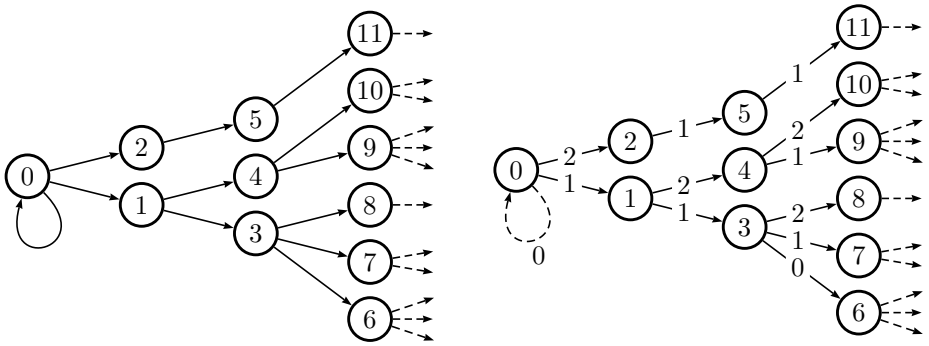
It will prove to be extremely convenient to have a slightly different look at trees and to consider that the root of a tree is also a *child of itself*, that is, bears a loop onto itself. This convention is sometimes taken when implementing tree-like structures (*e.g.* file systems): it makes the father function total. We call such a structure an *i-tree*. It is so close to a tree that we pass from tree to i-tree (or conversely) with no further ado.

We call *signature* any infinite sequence $\boldsymbol{s}$ of non-negative integers. The signature $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ is *valid* if the following holds:

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

Definition 1. *The* breadth-first signature *or, for short, the* signature*, of a tree, or an i-tree, $\mathcal{T}$ is the sequence of the degrees of the nodes of the* i-tree $\mathcal{T}$ *in the order given by the breadth-first traversal of $\mathcal{T}$.*

In other words, $\boldsymbol{s} = s_0 s_1 s_2 \cdots$ is the signature of a tree $\mathcal{T}$ if $s_0 = \mathsf{d}(0) + 1$ and $s_i = \mathsf{d}(i)$ for every node i of $\mathcal{T}$. Note that the definition implies that the signatures of a tree and of the corresponding i-tree are the same.



(a) The i-tree $T_{\boldsymbol{s}}$ (b) $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$, *with* $\boldsymbol{\lambda} = (012.12.1)^\omega$

Fig. 1. The i-tree and a language whose signature is $\boldsymbol{s} = (321)^\omega$

Proposition 1. *A tree has a valid signature and conversely a valid signature* $\boldsymbol{s}$ *uniquely defines a tree* $\mathcal{T}_{\boldsymbol{s}}$ *whose signature is* $\boldsymbol{s}$*.*

The proof of Proposition 1 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 starting point n and the end point m of the transition, both initially set to 0. In one step of the procedure, s_n nodes are created, corresponding to the integers $m, m+1, \ldots, (m+s_n-1)$, and s_n edges are created (all from n, and one to each of these new nodes). Then n is incremented by 1, and m by s_n.

The validity of $\boldsymbol{s}$ ensures that at each step of the procedure $n < m$, with the exception of the first step where $n = m = 0$. It follows that every node is strictly larger than its father, excepted for the root, whose father is itself. Figure 1a shows the i-tree whose signature is $(321)^\omega$.

3 Labelled Signatures of Languages

In the sequel, alphabets are totally ordered; and we use implicitly the natural order on digit alphabets (that is $0 < 1 < 2 < \cdots$). A word $w = a_0 a_1 \cdots a_{k-1}$ is *increasing* if $a_0 < a_1 < \cdots < a_{k-1}$. The length of a finite word w is denoted by $|w|$.

A labelled (i-)tree $\mathcal{T}$ is an (i-)tree whose arcs hold a label taken in an alphabet A. Since both $\mathcal{T}$ and A are ordered, the labels on arcs have to be *consistent*, that is, the labels of the arcs to the children of a same node are in the same order as the children: an arc to a smaller child is labelled by a smaller letter.

A labelled (i-)tree $\mathcal{T}$ defines the language of the branch labels. Conversely, a prefix-closed language L (over an ordered alphabet) uniquely defines a labelled (ordered) tree.

The labelling $\boldsymbol{\lambda}$ of a labelled tree $\mathcal{T}$ (labelled in A) is the infinite word in A^ω obtained as the sequence of the arc labels of $\mathcal{T}$ visited in a breadth-first search.

Definition 2. *Let* $\boldsymbol{s}$ *be a signature. An infinite word* $\boldsymbol{\lambda}$ *in* A^ω *is* consistent *with* $\boldsymbol{s}$ *if the factorisation of* $\boldsymbol{\lambda}$ *in the infinite sequence* $(w_n)_{n\in\mathbb{N}}$ *of words in* A^**:* $\boldsymbol{\lambda} = w_0\, w_1\, w_2 \cdots$ *induced by the condition that for every* n *in* $\mathbb{N}$*,* $|w_n| = s_n$*, has the property that for every* n *in* $\mathbb{N}$*,* w_n *is an* increasing word*.*

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

A simple and formal verification yields the following.

Proposition 2. *A prefix-closed language* L *uniquely determines a labelled tree and hence a valid labelled signature,* the *labelled signature of* L *and conversely any valid labelled signature* $(\boldsymbol{s}, \boldsymbol{\lambda})$ *uniquely determines a labelled tree* $\mathcal{T}_{(\boldsymbol{s},\boldsymbol{\lambda})}$ *and hence a prefix-closed language* $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$*, whose signature is precisely* $(\boldsymbol{s}, \boldsymbol{\lambda})$*.*

Figure 1b shows the labelling of the i-tree whose signature is $\boldsymbol{s} = (3\,2\,1)^\omega$ by the infinite word $\boldsymbol{\lambda} = (0\,1\,2\,.\,1\,2\,.\,1)^\omega$. This is of course a very special labelling: labellings consistent with $\boldsymbol{s}$ need not be periodic.

The identification between a prefix-closed language L and the tree $\mathcal{T}_L$ whose branch language is L (and whose set of nodes is $\mathbb{N}$) is very similar to the processes proposed in the works of Lecomte et Rigo [7,8] for the definition of the *Abstract Numeration Systems* (ANS) — without the assumption that L is rational, and with the restriction that L is prefix-closed. Indeed, the $(n+1)$-th word of L in the radix order is the label of the path from the root 0 to the node n in $\mathcal{T}_L$. (The first word of L is always ε and labels the empty path from the root to itself.)

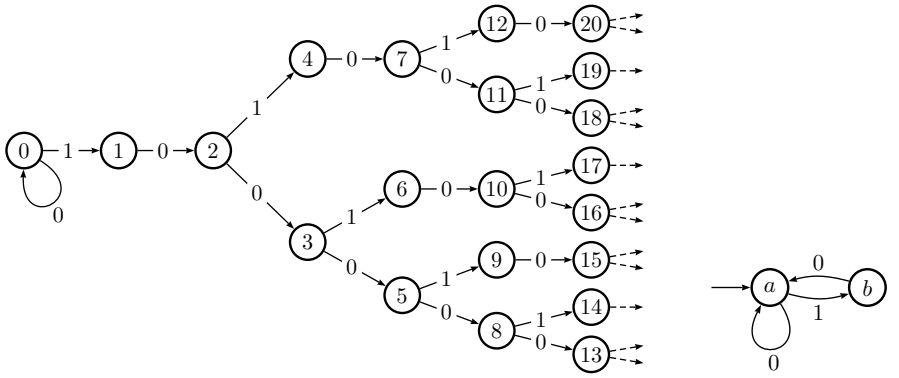
Remark 1. A very simple tree paired with the appropriate labelling may produce an artificially complex language. For instance, the infinite unary tree may be labelled by a non-recursive word. This explains why a result relative to the regularity of languages defined by signatures will always require some restriction on the labelling. The notion of *substitutive labelled signature* defined in the next Section 4 is an example of such a restriction.

4 Substitutive Signature and Rational Languages

We follow [9] for the terminology and basic definitions on *substitutions*. Let A be an alphabet. A morphism $\sigma : A^* \to A^*$ is *prolongable* on a letter a in A if $\sigma(a) = au$ for some word u and moreover $lim_{n\to+\infty}|\sigma^n(a)| = +\infty$. Then, the sequence $(\sigma^n(a))_{n\in\mathbb{N}}$ converges to an infinite word denoted by $\sigma^\omega(a)$; any such word is called *purely substitutive*. The image $f(w)$ of a purely substitutive word w by a letter-to-letter morphism f is called a *substitutive word*.

Definition 3. *Let* $\sigma : A^* \to A^*$ *be a morphism prolongable on a in* A *and let* $f_\sigma : A^* \to D^*$ *be the letter-to-letter morphism defined by* $\forall b \in A$*,* $f_\sigma(b) = |\sigma(b)|$*. The substitutive word* $f_\sigma(\sigma^\omega(a))$ *is called a* substitutive signature*.*

Furthermore, let $g : A^* \to B^*$ *be a morphism satisfying the following condition:* $\forall b \in A,\ |g(b)| = f_\sigma(b)$*. The pair* $(f_\sigma(\sigma^\omega(a)),\ g(\sigma^\omega(a)))$ *is called a* substitutive labelled signature*.*



(a) $L_{(\boldsymbol{r},\mu)}$: integer representations in the Fibonacci numeration system.

(b) $\mathcal{A}_{(\boldsymbol{r},\mu)}$: automaton accepting $L_{(\boldsymbol{r},\mu)}$.

Fig. 2. The Fibonacci signature $\boldsymbol{r} = \sigma^\omega(a)$ with $\sigma(a) = ab$ and $\sigma(b) = a$

The next lemma is a direct consequence of the fact that if σ denotes a morphism prolongable on a and if w denotes a prefix of $\sigma^\omega(a)$, then $|\sigma(w)| > |w|$.

Lemma 1. *A substitutive signature is valid.*

Example 1 (The Fibonacci signature). The Fibonacci word is the purely substitutive word $\sigma^\omega(a)$ defined by $\sigma(a) = ab$ and $\sigma(b) = a$:

$$\sigma^\omega(a) \quad = \quad abaababaabaab\cdots$$

Hence the substitutive signature defined by σ is

$$\boldsymbol{r} \quad = \quad f_\sigma(\sigma^\omega(a)) \quad = \quad 2122121221221\cdots$$

Let g be the morphism defined by $g(a) = 01$ and $g(b) = 1$ defining the labelling $\boldsymbol{\mu} = g(\sigma^\omega(a))$ (which is consistent with $\boldsymbol{r}$):

$$\boldsymbol{\mu} \quad = \quad g(\sigma^\omega(a)) \quad = \quad 01.0.01.01.0.01.0.01.01.0.01.01.0\ldots$$

The language $L_{(\boldsymbol{r},\boldsymbol{\mu})}$, as shown at Figure 2a, is the language of integer representations in the Fibonacci numeration system.

Theorem 1. *A prefix-closed language is rational if and only if its labelled signature is substitutive.*

The proof of this theorem relies on a correspondence between finite automata and substitutive words used by Rigo and Maes in [3] (*cf.* also [8, Section 3.4]) to prove the equivalence between two decision problems. A similar construction was used by Dumont and Thomas in [4] to define the prefix-suffix graph.

We give here the proof of one direction in detail, reformulated into the next proposition. The other direction is analogous.

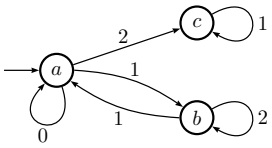


Fig. 3. The automaton $\mathcal{A}_{(\theta,h)}$ accepting the language $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ shown at Figure 1b with $\theta(a) = abc$, $\theta(b) = ab$, $\theta(c) = c$, $h(a) = 012$, $h(b) = 12$, $h(c) = 1$

Proposition 3. *If $(\boldsymbol{s}, \boldsymbol{\lambda})$ is a valid substitutive labelled signature, then $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ is a rational language.*

Proof. Let $\sigma : A^* \to A^*$ such that $\boldsymbol{s} = f_\sigma(\sigma^\omega(a))$ and $g : A^* \to B^*$ such that $\boldsymbol{\lambda} = g(\sigma^\omega(a))$. Since we are using two alphabets at the same time, a, b, c will denote letters of A and x, y letters of B.

Let $\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})} = \langle A, B, \delta, a, A \rangle$ be the automaton whose set of states is A; the alphabet is B; the initial state is a; all states are final; and the transition function is defined as follows. For every b in A, let $k = |\sigma(b)| = |g(b)|$. From b, there are k outgoings transitions and for every i, $1 \leqslant i \leqslant k$, $b \xrightarrow{y} c$, where c is the i-th letter of $\sigma(b)$ and y is the i-th letter of $g(b)$. Figure 2b shows the automaton computed from the Fibonacci signature; see Figure 3 for the signature of Example 2 below.

Note that since the morphism σ is prolongable on a, the automaton $\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})}$ features a loop $a \xrightarrow{x} a$ on the initial state whose label x is the first letter of $g(a)$. This loop induces that $L(\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})})$ is of the form x^*L and leading x's serve the same role as leading 0's in usual numeration systems. We denote by L the language containing the words of $L(\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})})$ that does not start with an x. Proving that $L(\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})})$ has $(\boldsymbol{s}, \boldsymbol{\lambda})$ for signature amounts to prove that if w_i denotes the $(i+1)$-th word of L in the radix order, then w_i reaches the state corresponding to the $(i+1)$ letter of $\sigma^\omega(a)$.

Let b be a letter of A, hence a state of $\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})}$. The word $\sigma(b)$ is exactly the sequence of the states that are direct successors of b in $\mathcal{A}_{(\boldsymbol{s},\boldsymbol{\lambda})}$ in the right order that is, a successor by a smaller label is before a successor by a larger label. It follows that the word $\sigma(\sigma(b))$ is the sequence of the states that are reachable from b in two steps and once again, in the right order. An easy induction yields that $\sigma^i(b)$ is the sequence of the states reachable in exactly i steps.

If $\sigma(a) = au$, then the words of length 1 of L reach the states of u (and the empty word reaches the state a). An easy induction yields that the words of length i belonging to L reach the sequence of states $\sigma^i(u)$. Hence the words of L taken in the radix order reach the state sequence $a\,u\,\sigma(u)\,\sigma^2(u)\sigma^3(u)\cdots$ which is equal to $\sigma^\omega(a)$.

Remark 2. As we said, our view on a prefix-closed rational language L essentially amounts to considering L as an ANS, in the sense of [7,8]. The consequence of Theorem 1 is to associate with L a substitution σ_L. In [4], Dumont and Thomas described the numeration system associated with a substitution σ. It can be derived from the construction of Theorem 1 that the ANS L may be mapped onto the Dumont-Thomas numeration system associated with σ_L.

5 On Ultimately Periodic Signature

Let $\boldsymbol{s} = uv^\omega$ be an ultimately periodic word over the alphabet $\{0, 1, \ldots, k\}$; we call *growth ratio* of v, denoted by $gr(v)$, the average of the letters of v:

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

We treat here the case where $gr(v)$ is an integer that is, when the sum of the letters of v is a multiple its length. In this case, uv^ω is a substitutive signature.

Proposition 4. *If $\boldsymbol{s}$ denotes an ultimately periodic valid signature whose growth ratio is an integer, then $\boldsymbol{s}$ is a substitutive signature.*

Proof. Let $\boldsymbol{s} = uv^\omega$ be an ultimately periodic signature. We write $k = |u|$, $n = |v|$ and denote by A an alphabet whose $(k + n)$ letters are denoted as follows.

$$A = B \uplus C \quad \text{where} \quad B = \{b_0, b_1, \ldots, b_{(k-1)}\} \quad \textit{and} \quad C = \{c_0, c_1, \ldots, c_{(n-1)}\} \ .$$

The letters of B correspond to positions of u and those of C to positions of v. Let $\sigma : A^* \to A^*$ be a morphism defined implicitly by

$$\sigma(b_0 b_1 \cdots b_{(k-1)} c_0 c_1 \cdots c_{(n-1)}) \text{ is prefix of } b_0 b_1 \cdots b_{(k-1)} (c_0 c_1 \cdots c_{(n-1)})^\omega \quad (2a)$$

$$\forall i < k \qquad |\sigma(b_i)| = u_i \quad (2b)$$

$$\forall i < n \qquad |\sigma(c_i)| = v_i \quad (2c)$$

Let us denote by $\overline{u} = b_0 b_1 \cdots b_{(k-1)}$ and by $\overline{v} = c_0 c_1 \cdots c_{(n-1)}$, hence, respectively from Equations 2b and 2c, $f_\sigma(\overline{u}) = u$ and $f_\sigma(\overline{v}) = v$. Let i and j be the two integers such that $\sigma(\overline{u}) = \overline{u}(\overline{v})^i c_0 \cdots c_{(j-1)}$. Equation 2c implies that $|\sigma(\overline{v})| = n \times gr(v)$ hence, from Equation 2a,

$$\sigma(\overline{v}) = c_j \cdots c_{(n-1)} (\overline{v})^{gr(v)-1} c_0 \cdots c_{(j-1)} \ .$$

It follows that $\overline{u}(\overline{v})^\omega$ is a fixed point of σ.

It remains to prove that the morphism σ is prolongable on b_0 or, more precisely, that $lim_{n \to +\infty} |\sigma^n(b_0)| = +\infty$. Let us denote by w any prefix of $\overline{u}(\overline{v})^\omega$ and prove that $|\sigma(w)| > |w|$. Since w is a prefix of $\overline{u}(\overline{v})^\omega$, $f_\sigma(w)$ is a prefix of $\boldsymbol{s}$, and since $\boldsymbol{s}$ is valid, the sum of the letters of $f_\sigma(w)$ is strictly greater than $|w|$. From the definition of f_σ, $|\sigma(w)|$ is equal to the sum of the letters of $f_\sigma(w)$, hence is strictly greater than $|w|$.

Example 2. The purely periodic signature $\boldsymbol{s} = (321)^\omega$ is the substitutive signature $f_\theta(\theta^\omega(b_0))$ where θ is defined by $\theta(c_0) = c_0 c_1 c_2$, $\theta(c_1) = c_0 c_1$ and $\theta(c_2) = c_2$. Figure 3 shows the automaton $\mathcal{A}_{(\theta,h)}$ accepting $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ (shown at Figure 1b) where $\boldsymbol{\lambda} = h(\theta^\omega(c_0)))$ with $h(c_0) = 012$, $h(c_1) = 12$ and $h(c_2) = 1$. This language consists of non-canonical representations of the integers in base 2 (that is, the growth ratio of $\boldsymbol{s}$): the $(n + 1)$-th word[1] of $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ in the radix order is a word $d_n d_{n-1} \cdots d_0$ over the alphabet $\{0, 1, 2\}$ and its binary value $\sum_{i=0}^{n} d_i 2^i$ is equal to n.

[1] Recall that we are ignoring leading 0's, hence the $(n + 1)$-th word of $L_{(\boldsymbol{s},\boldsymbol{\lambda})}$ is the one labelling the path $0 \longrightarrow n$ in Figure 1b.

6 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.

In this first work on the subject, we have proved that rational languages are associated with (a particular subclass of) substitutive words. We also proved that ultimately periodic signatures whose growth ratio is an integer are substitutive, and hinted their link to integer base numeration systems.

In a forthcoming paper [2], we study the class of languages associated with periodic signatures whose growth ratio is not an integer and how they are related to the representation language in rational base numeration systems. In the future, our aim is to further explore this relationship by means of the notion of direction, that extends the notion of growth ratio to aperiodic signatures.

Acknowledgements. The authors are grateful to the referee who drew their attention to the work of Dumont and Thomas.

References

1. Akiyama, S., Frougny, C., Sakarovitch, J.: Powers of rationals modulo 1 and rational base number systems. Israel J. Math. 168, 53–91 (2008)
2. Marsault, V., Sakarovitch, J.: Rhythmic generation of infinite trees and languages (2014), In preparation, early version accessible at arXiv:1403.5190
3. Rigo, M., Maes, A.: More on generalized automatic sequences. Journal of Automata, Languages and Combinatorics 7(3), 351–376 (2002)
4. Dumont, J.M., Thomas, A.: Systèmes de numŕation et fonctions fractales relatifs aux substitutions. Theor. Comput. Sci. 65(2), 153–169 (1989)
5. Cobham, A.: Uniform tag sequences. Math. Systems Theory 6, 164–192 (1972)
6. Diestel, R.: Graph Theory. Springer (1997)
7. Lecomte, P., Rigo, M.: Numeration systems on a regular language. Theory Comput. Syst. 34, 27–44 (2001)
8. Lecomte, P., Rigo, M.: Abstract numeration systems. In: Berthé, V., Rigo, M. (eds.) Combinatorics, Automata and Number Theory, pp. 108–162. Cambridge Univ. Press (2010)
9. Berthé, V., Rigo, M.: Combinatorics, Automata and Number Theory. Cambridge University Press (2010)