Breadth-first serialisation of trees and rational languages

Victor Marsault*,† and Jacques Sakarovitch†

2014 – 06 – 23

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. We then use this construction to highlight the relationship between Dupont-Thomas numeration systems and abstract numeration systems.

1 Introduction

This work introduces 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 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 characteristic of the tree. Similarly, we call *labelling* the infinite sequence of the labels of the arcs visited by the breadth-first traversal of a labelled tree. The pair signature/labelling is once again characteristic of the labelled tree, hence providing a serialisation of labelled trees, hence of prefix-closed languages.

The serialisation of a (prefix-closed) language is very close 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 constant signature p^ω. 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; and the rational base numeration systems as defined in [1] and whose representation languages have periodic signatures.

In Section 4, we prove that the signatures of (prefix-closed) rational languages all belong to a special subclass of substitutive infinite words that we call *substitutive signatures*. Conversely, we prove that every substitutive signature, paired with an appropriate *subtitutive labelling*, generates a prefix-closed rational language. The proof of these results relies on a correspondence between substitutive words and automata due to Maes and Rigo [11] or Dumont and Thomas [5, 6] and whose principle goes back to the work of Cobham [3].

*Corresponding author, victor.marsault@telecom-paristech.fr
†LTCI, CNRS / Telecom ParisTech

In Section 5, we apply the signature viewpoint to the study of numeration systems. We take the most general settings, that is, the one of Abstract Numeration System (ANS) proposed by Leconte and Rigo [7]: a numeration system is defined by an arbitrary language L (over an ordered alphabet) that will be the integer representations in the new numeration system. An ANS is rational and called ARNS if L is a rational language (indeed Leconte and Rigo consider ARNS only, which they call ANS, but for further development we rather be specific). We prove that an ARNS is essentially determined by its signature (that is, that its labelling has little influence), in the sense that if two ARNS's have the same signature, then there exists a conversion function from one to the other which is realised by a pure sequential and letter-to-letter transducer.

Moreover, we associate with every (prefix-closed) ARNS L a so called Dumont-Thomas numeration system with the same signature. After verifying that every Dumont-Thomas numeration system is indeed a prefix closed ARNS's, the previous paragraph applies and both numeration systems are essentially the same. In this special case, the conversion transducer has a special structure: its underlying input automaton is the prefix automaton of the D.-T numeration system and its underlying output automaton is the automaton accepting L.

2 Serialisation of trees

Classically, a tree is an undirected graph which is acyclic and connected. In this note, however, we call *tree* a directed infinite graph which is

- rooted: there is a special node called the *root*;
- directed outward from the root: there is a unique path from the root to every node.
- ordered: the children of every nodes are ordered;

A tree of this class has a canonical breadth-first traversal: the one starting from the root and following the order of children. We may then consider that the node set of a tree is always $\mathbb{N}$: 0 is the root and the node i is the $(i+1)$-th node visited by this canonical traversal. We draw trees with the root on the left, arcs rightwards and the order of children is implicit with the convention that lower children are smaller.

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

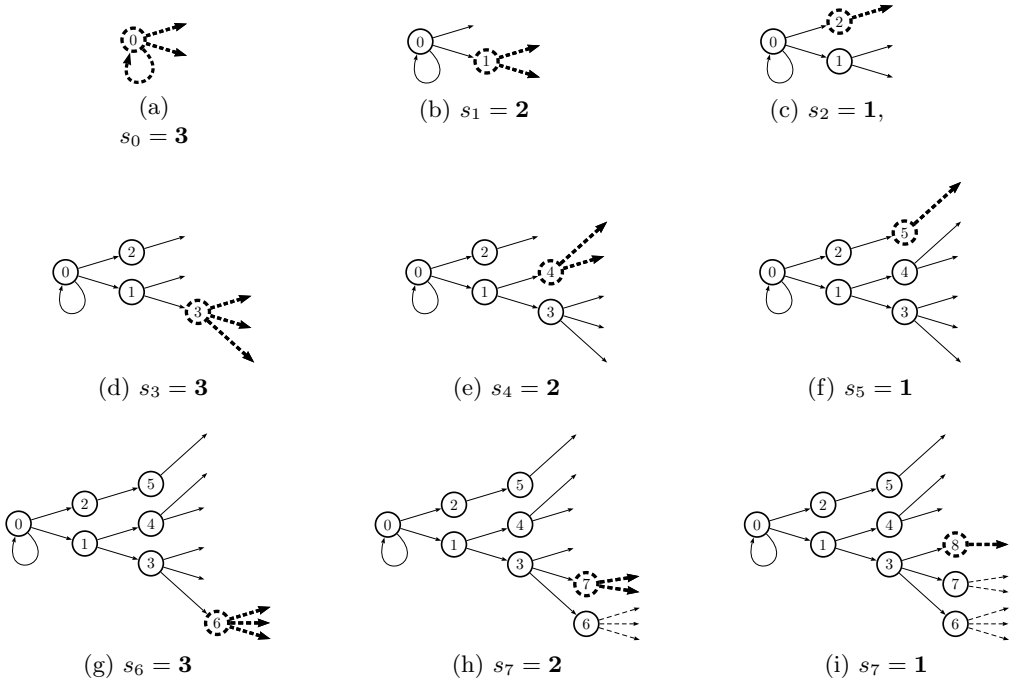
$$\forall j \in \mathbb{N} \qquad \sum_{i=0}^{j} s_i \ > \ j+1 \ . \tag{1}$$

Definition 1. *The* breadth-first signature *or, for short, the* signature*, of a tree, or 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.

Proposition 2. *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 2 takes the form of a procedure generating an i-tree from a valid signature. For instance, Figure 1 shows the first nine steps of the generation of $\mathcal{T}_{\boldsymbol{s}_1}$ by its signature $\boldsymbol{s}_1 = (321)^\omega$.



(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) $s_7 = \mathbf{2}$ (i) $s_7 = \mathbf{1}$

Figure 1: The first nine steps of the generation of $\mathcal{T}_{(321)^\omega}$

3 Serialisation of languages

In the sequel, alphabets are totally ordered. We say that 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 tree $\mathcal{T}$ is a tree whose arcs hold a label taken in an alphabet A. Since both $\mathcal{T}$ and A are ordered, the labels on the 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.

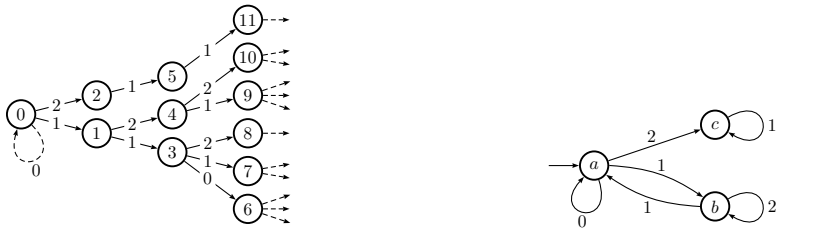
Definition 3. *Let $\boldsymbol{s}$ be a signature. An infinite word $\boldsymbol{\lambda}$ in A^ω is* consistent *with $\boldsymbol{s}$ if the factorisation of $\boldsymbol{\lambda}$ in the infinite sequence $(w_n)_{n\in\mathbb{N}}$ of words in A^*: $\boldsymbol{\lambda} = w_0\, w_1\, w_2 \cdots$ induced by the condition that for every n in $\mathbb{N}$, $|w_n| = s_n$, has the property that w_n is an* increasing word, *for every n in $\mathbb{N}$.*

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 labelled tree $\mathcal{T}$ defines the (prefix-closed) language of the branch labels and conversely, a prefix-closed language L uniquely defines a labelled tree. This identification between prefix closed languages and labelled trees is very similar to the process due to Lecomte and Rigo [7, 8] defining an Abstract Numeration System (ANS, *cf.* Section 5).

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 by the canonical breadth-first search. A simple and formal verification yields the following.

Proposition 4. *A prefix-closed language L uniquely determines a labelled tree and hence a valid labelled signature,* the *labelled signature of L. 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 $(\boldsymbol{s}, \boldsymbol{\lambda})$.*



(a) The language $L_{(\boldsymbol{s}_1,\boldsymbol{\lambda}_1)}$ (b) The automaton $\mathcal{A}_{(\sigma_1,g_1)}$

Figure 2: The signature $\boldsymbol{s}_1 = (321)^\omega$ and $\boldsymbol{\lambda}_1 = (012.12.1)^\omega$

Figure 2a shows the labelling of the i-tree whose signature is $\boldsymbol{s}_1 = (3\,2\,1)^\omega$ by the infinite word $\boldsymbol{\lambda}_1 = (0\,1\,2\,.\,1\,2\,.\,1)^\omega$. This is of course a very special labelling: labellings consistent with $\boldsymbol{s}_1$ need not be periodic.

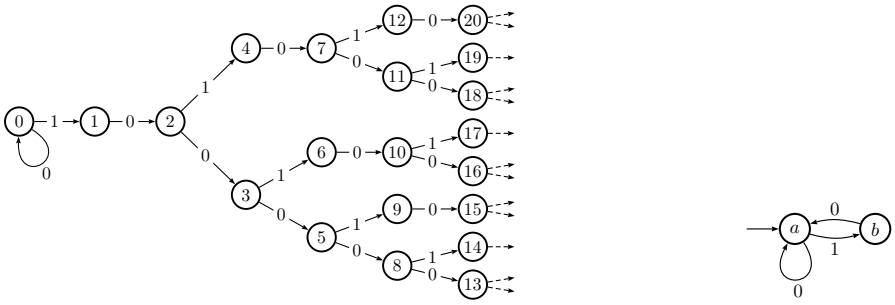
4 Serialisation of rational languages

We follow [2] for the terminology and basic definitions on *substitutions* and [12] for those on finite automata.

Definition 5. *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 $f_\sigma(b) = |\sigma(b)|$, for every b in A. The substitutive word $f_\sigma(\sigma^\omega(a))$ is called a* substitutive signature.

Furthermore, let $g : A^* \to B^*$ be a morphism satisfying $|g(b)| = f_\sigma(b)$, for every b in A. The pair $(f_\sigma(\sigma^\omega(a)),\ g(\sigma^\omega(a)))$ is called a substitutive labelled signature*, and also denoted by (σ, g) for convenience.*

Example 6 (The Fibonacci signature)**.** *Let σ_2 and g_2 be the two morphisms defined by $\sigma_2(a) = ab$, $\sigma_2(b) = a$ and $g_2(a) = 01$, $g_2(b) = 1$. Then (σ_2, g_2) is the substitutive signature of the integer representations in the Fibonacci numeration system, shown at Figure 3a.*



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

Figure 3: The integer representations in the Fibonacci numeration system.

Theorem 7. *A prefix-closed language is rational if and only if its signature is substitutive.*

The proof of this theorem relies on a correspondence between finite automata and substitutive words similar to the one used by Rigo and Maes in [11] to prove the equivalence between two decision problems. Definitions 8 and 9 specify the two directions of this correspondence, that is, two transformations that are the reverse of each other.

Definition 8. *Let (σ, g) be a substitutive labelled signature with $\sigma : A^* \to A^*$ prolongable on a and $g : A^* \to B^*$. We define the* automaton associated with (σ, g)*, denoted by $\mathcal{A}_{(\sigma,g)}$, as follows:*

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

The 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, we write $k = |\sigma(b)| = |g(b)|$. 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 3b shows the automaton $\mathcal{A}_{(\sigma_2,g_2)}$ computed from the Fibonacci substitution and Figure 2b shows $\mathcal{A}_{(\sigma_1,g_1)}$ accepting $L_{(\boldsymbol{s}_1,\boldsymbol{\lambda}_1)}$ (Figure 2a); see Example 11 below for the definitions of (σ_1, g_1).

Definition 9. *Let $\mathcal{A} = \langle\, Q, A, \delta, i, Q \,\rangle$ be a prefix-closed automaton accepting. We denote by $\sigma_{\mathcal{A}}$ the substitution $Q^* \to Q^*$ and by $g_{\mathcal{A}}$ the morphism $Q^* \to A^*$ defined as follows. Let p be a state of Q and let us denote by $p \xrightarrow{a_0} q_0, \ldots,\ p \xrightarrow{a_k} q_k$ all its outgoings transitions. Without loss of generality we assume that $a_0 < a_1 < \cdots < a_k$ and then define*

$$\sigma_{\mathcal{A}}(p) \ = \ q_0\, q_i \cdots q_k \qquad \textit{and} \qquad g_{\mathcal{A}}(p) \ = \ a_0\, a_1 \cdots a_k \ .$$

A word on ultimately periodic signatures Let $\boldsymbol{s} = uv^\omega$ be an ultimately periodic word over the alphabet $\{0, 1, \ldots, k\}$; we call *growth ratio* of v (or, alternatively, the growth ratio of $\boldsymbol{s}$), denoted by $\mathsf{gr}(v)$, the average of the letters of $v = a_0 a_1 \cdots a_{k-1}$:

$$\mathsf{gr}(u\,(a_0\, a_1 \cdots a_{k-1})^\omega) \quad = \quad \mathsf{gr}(a_0\, a_1 \cdots a_{k-1}) \quad = \quad \frac{\sum_{i=0}^{k-1} a_i}{k} \ .$$

Proposition 10. *Let $\boldsymbol{s}$ be an ultimately periodic (valid) signature. The growth ratio of s is an integer if and only if $\boldsymbol{s}$ is a substitutive signature.*

The backward direction is not treated in the complete version of this note, but in another, still in preparation [10]. The proof of the forward direction consists in creating a substitution with $p + m$ letters where p is the period length and m is the pre-period length.

Example 11. *The purely periodic signature $\boldsymbol{s}_1 = (321)^\omega$ is equal to $f_{\sigma_1}({\sigma_1}^\omega(a))$ where σ_1 by $\sigma_1(a) = abc$, $\sigma_1(b) =$ and $\sigma_1(c) = c$. The labelling $\boldsymbol{\lambda}_1 = (012.12.1)^\omega$ (used in Figure 2a) is then equal to $g_1({\sigma_1}^\omega(a))$, where $g_1(a) = 012$, $g_1(b) = 12$ and $g_1(c) = 1$.*

5 Signatures and numeration systems

In this section, we apply a signature viewpoint on to the study of numeration system. We consider first the (prefix-closed) Abstract (Rational) Numeration Systems due to Leconte and Rigo [7].

Definition 12. *Let (σ, g) be a substitutive labelled signature, defining the prefix-closed rational language $L_{(\sigma,g)}$. We denote by $\langle\, . \,\rangle_{(\sigma,g)}$ the* representation function $\mathbb{N} \to L_{(\sigma,g)}$*. This defines the* Abstraction Rational Numeration System *(ARNS) associated with (σ, g).*

Let $L_{(\sigma,g)}$ and $L_{(\tau,h)}$ be two ARNS, the function mapping $\langle n \rangle_{(\sigma,g)}$ to $\langle n \rangle_{(\tau,h)}$ is called the *conversion function*. Intuitively, this function measures the distance between the two numeration systems: if it is realised by a simple structure (*e.g.* a finite transducer) they are closely related.

Proposition 13. *Let (σ, g) and (τ, h) be two substitutive labelled signatures such that $f_\sigma(\sigma^\omega(a))$ and $f_\tau(\tau^\omega(a))$ are equal. The conversion function from $L_{(\sigma,g)}$ to $L_{(\tau,h)}$ is realised by a letter-to-letter and purely sequential transducer.*

We will now use our framework to describe the so called Dumont and Thomas Numeration system (DTNS, [4, 5, 6]). Let $\sigma : A^* \to A^*$ be a substitution prolongable on a. We denote by A_σ an alphabet whose letters are *words* belonging to A^*; it is the set of the strict prefixes of the images of the letters of A by σ: $A_\sigma = \{\, u \mid u \text{ is a strict prefix of } \sigma(b) \text{ for some } b \in A \,\}$.

Let us emphasis that a word of A_σ^* is not a word of A^*, for instance if η denotes the letter of A_σ corresponding to the empty word over A, the words η and $\eta\eta$ are two different words of A_σ^*.

The prefix automaton associated with σ (originally defined in [6]), is exactly the automaton $\mathcal{A}_{(\sigma,g_\sigma)}$ (from Definition 8) where $g_\sigma : A^* \to A_\sigma^*$ is defined as follows.

$$\forall b \in A \qquad g_\sigma(b) \;=\; x_0\, x_1 \cdots x_k \quad \left| \begin{array}{l} \text{where } k = |\sigma(b)| - 1; \\ \text{and } x_i \text{ is the prefix of } \sigma(a) \text{ of length } i. \end{array} \right.^{1} \qquad (2)$$

We denote by ρ_σ the function $A_\sigma^* \to A^*$ defined by:

$$\rho_\sigma(y_k\, y_{k-1} \;\cdots\; y_0) \;=\; \sigma^k(y_k)\, \sigma^{k-1}(y_{k-1}) \cdots \sigma^0(y_0) \;. \qquad (3)$$

Theorem 14 (Dumont, Thomas [4])**.** *Let σ be a morphism and $n \geqslant 1$ be an integer. There exists a unique word u of A_σ^* accepted by $\mathcal{A}_{(\sigma,g_\sigma)}$ such that $|\rho_\sigma(u)| = n$.*

With this denotation, the word u is called the *Dumont-Thomas representation* of the integer n, defining a Dumont-Thomas numeration system. The next lemma follows from Theorem 14.

Lemma 15. *The integer representations in the DTNS associated with σ is the (prefix-closed) rational language $L_{(\sigma,g_\sigma)}$.*

The next two theorems compare the class of prefix-closed ARNS's with the class of DTNS's. The latter is contained in the former (Theorem 16) and every element of the former is easily convertible into one of the latter (Theorem 17).

Theorem 16. *Every Dumont-Thomas numeration system is the (prefix-closed) ARNS $L_{(\sigma,g_\sigma)}$.*[2]

Theorem 17. *Let $L_{(\sigma,g)}$ be a prefix closed ARNS. The conversion function from $L_{(\sigma,g)}$ to the DTNS associated with σ is realised realised by a letter-to-letter and purely sequential transducer. In addition, this transducer has $\mathcal{A}_{(\sigma,g)}$ as input automaton and $\mathcal{A}_{(\sigma,g_\sigma)}$ as output automaton.*

6 Conclusion and Future Work

We introduced a way of effectively describing infinite trees and languages by infinite words using a simple breadth-first traversal. 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.

We have then applied this framework to study numeration systems and have proved that the so called Dumont-Thomas numeration systems and the prefix-closed abstract rational numeration systems are essentially the same object.

In a forthcoming paper [10], we study the 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, we plan to further explore the relationship between numeration systems and signature by extending the notion of growth ratio to aperiodic signatures.

Acknowledgements The authors are grateful to the referee of DLT2014 who drew their attention to the work of Dumont and Thomas.

[1] Note that x_i is a strict prefix of $\sigma(b)$, hence a *letter* of A_σ.

[2] Theorem 16 is a not trivial: it is not clear that the D.-T. representation of n is the $(n+1)$-th word of $L_{(\sigma,g_\sigma)}$.

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] Valérie Berthé and Michel Rigo. *Combinatorics, Automata and Number Theory.* Cambridge University Press, 2010.

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

[4] Jean-Marie Dumont and Alain Thomas. Systèmes de numération et fonctions fractales relatifs aux substitutions. *Theor. Comput. Sci.*, 65(2):153–169, 1989.

[5] Jean-Marie Dumont and Alain Thomas. Digital sum problems and substitutions on a finite alphabet. *Journal of Number Theory*, 39(3):351–366, 1991.

[6] Jean Marie Dumont and Alain Thomas. Digital sum moments and substitutions. *Acta Arith*, 64(3):205–225, 1993.

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

[8] Pierre Lecomte and Michel Rigo. Abstract numeration systems. In *Combinatorics, Automata and Number Theory* [2].

[9] Victor Marsault and Jacques Sakarovitch. On sets of numbers rationally represented in a rational base number system. In *DLT*, volume 8633 of *LNCS*. Springer, 2014. to appear. Early version accessible at arXiv.org:1404.0798.

[10] Victor Marsault and Jacques Sakarovitch. Rhythmic generation of infinite trees and languages, 2014. In preparation, early version accessible at arXiv:1403.5190.

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

[12] Jacques Sakarovitch. *Elements of Automata Theory.* Cambridge University Press, 2009. Corrected English translation of *Éléments de théorie des automates*, Vuibert, 2003.