Trees and Languages with Periodic Signature

Victor Marsault[1] and Jacques Sakarovitch[2]

[1] LIAFA, Université Denis Diderot, 8 place Aurélie Nemours, 75013 Paris, France
Victor.Marsault@liafa.univ-paris-diderot.fr
[2] Telecom-ParisTech and CNRS, 46 rue Barrault 75013 Paris, France

Abstract. The *signature* of a labelled tree (and hence of its prefix-closed branch language) is the sequence of the degrees of the nodes of the tree in the *breadth-first traversal*. In a previous work, we have characterised the signatures of the regular languages. Here, the trees and languages that have the simplest possible signatures, namely the periodic ones, are characterised as the sets of representations of the integers in rational base numeration systems.

1 Introduction

Rational base numeration systems were defined in a joint work of the second author together with S. Akiyama and Ch. Frougny [1] and allowed to make some progress in a number theoretic problem, by means of automata theory and combinatorics of words. At the same time, it raised the problem of understanding the structure of the sets of the representations of the integers in these systems from the point of view of formal language theory.

At first sight, these sets look rather chaotic and do not fit in the classical Chomsky hierarchy of languages. They all enjoy a property that makes them defeat, so to speak, any kind of iteration lemma. On the other hand, the most common example given by the set of representations in the base $\frac{3}{2}$ exhibits a remarkable regularity. The set $L_{\frac{3}{2}}$ of representations, which are words written with the three digits $\{0, 1, 2\}$, is prefix-closed and thus naturally represented as a subtree of the full ternary tree which is shown in Fig. 1. It is then easily observed that the *breadth-first* traversal of that tree yields an infinite *periodic* sequence of degrees: $2, 1, 2, 1, 2, 1, \ldots = (2\,1)^\omega$. Moreover, the sequence of labels of the arcs in the same breadth-first search is also a purely periodic sequence $0, 2, 1, 0, 2, 1, \ldots = (0\,2\,1)^\omega$.[3]

Let us call *signature* of a tree (or of the corresponding prefix-closed language) the sequence of degrees in a breadth-first traversal of the tree. With this example, we are confronted with a situation where a regular process, a periodic signature, give birth to the highly non regular language, $L_{\frac{3}{2}}$. This paradox was the incentive to look at the breadth-first traversal description of languages in general. We have

[3] The sequence of degrees observed on the tree in the figure begins indeed with a 1 instead of a 2, the sequence of labels begins at the second term. These discrepancies will be explained later.

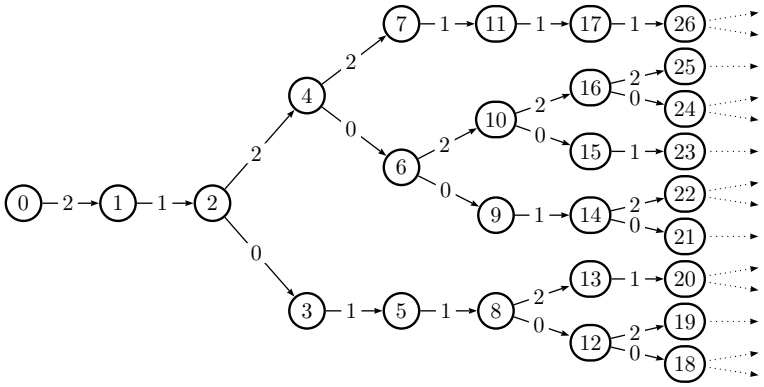


Fig. 1: The tree $\mathcal{T}_{\frac{3}{2}}$, representation of the language $L_{\frac{3}{2}}$

shown in [11] that regular languages are characterised by signatures belonging to a special class of morphic words. The purpose of this paper is to establish that a periodic signature is *characteristic* of the languages of representations of the integers in rational base numeration systems (roughly speaking and up to very simple and rational transformations).

Let us be more specific in order to state more precisely the characterisation results. An ordered tree of finite degree $\mathcal{T}$ is characterised by the infinite sequence of the degrees of its nodes visited in the order given by the breadth-first search, which we call the *signature* $\boldsymbol{s}$ of $\mathcal{T}$. Such a signature $\boldsymbol{s}$, together with an infinite sequence $\boldsymbol{\lambda}$ of letters taken in an ordered alphabet form a *labelled signature* $(\boldsymbol{s}, \boldsymbol{\lambda})$ and characterises then a *labelled* tree $\mathcal{T}$. The breadth-first search of $\mathcal{T}$ corresponds to the enumeration in the *radix order* of the prefix-closed language $L_{\mathcal{T}}$ of branches of $\mathcal{T}$.

We call *rhythm* of *directing parameter* (q, p) a q-tuple $\mathbf{r}$ of integers whose sum is p: $\mathbf{r} = (r_0, r_1, \ldots, r_{q-1})$ and $p = r_0 + r_1 + \cdots + r_{q-1}$. With $\mathbf{r}$, we associate sequences $\boldsymbol{\gamma}$ of p letters that meet some consistency conditions. And we consider the languages that are determined by the labelled signature $(\mathbf{r}^\omega, \boldsymbol{\gamma}^\omega)$. The characterisation announced above splits in two parts.

We first determine (Theorem 1) the remarkable labelled signature $(\mathbf{r}^\omega_{\frac{p}{q}}, \boldsymbol{\gamma}^\omega_{\frac{p}{q}})$ of the languages $L_{\frac{p}{q}}$. The rhythm $\mathbf{r}_{\frac{p}{q}}$ of $L_{\frac{p}{q}}$ corresponds roughly to *the most equitable way of partitioning p objects into q parts*. We call it the *Christoffel rhythm* associated with $\frac{p}{q}$, as it can be derived from the more classical notion of Christoffel word of slope $\frac{p}{q}$ (*cf.* [2]), that is, the canonical way to approximate the line of slope $\frac{p}{q}$ on a $\mathbb{Z} \times \mathbb{Z}$ lattice. The labelling $\boldsymbol{\gamma}_{\frac{p}{q}}$ is induced by the generation of $\mathbb{Z}/p\mathbb{Z}$ by q.

The converse is more convoluted but its complexity is confined in the definition of a *special labelling* $\boldsymbol{\gamma}_{\mathbf{r}}$ associated with every rhythm $\mathbf{r}$ (Definition 5). It is then established (Theorem 2) that the language $L_{\mathbf{r}}$ generated by the labelled signature $(\mathbf{r}^\omega, \boldsymbol{\gamma}^\omega_{\mathbf{r}})$ is a non-canonical representation of the integers in the base

which is the growth ratio of the rhythm $\mathbf{r}$. The properties of alphabet conversion in rational base numeration systems (*cf.* [1] or [5]) allow to conclude that for every rhythm $\mathbf{r}$, the language $L_{\mathbf{r}}$ is as complicated (or as simple, in the degenerate case where the growth ratio happens to be an integer) as these langages $L_{\frac{p}{q}}$.

The same techniques allow to treat the generalisation to *ultimately periodic* which raises no special difficulties and the results readily extend.

The languages with periodic labelled signature keep most of their mystery. But we have at least established that they are all alike, essentially similar to the representation languages of rational base numeration systems, and that variations in the rhythm and labelling do not really matter.

Due to space constraints, some proofs are only sketched and some figures have been removed. A complete version may be found on arXiv [9].

2 Rythmic Trees and Languages

Trees and I-trees Classically, a tree is an undirected graph in which any two vertices are connected by exactly one path (*cf.* [3], for instance). Our point of view differs in two respects (as already discussed in [11]).

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. In the figures, we draw trees with the root on the left, and arcs rightwards.

Second, our trees are *ordered*, that is, the set of children of every node is totally ordered. The order will be implicit in the figures, with the convention that lower children are smaller (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. 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 $\mathcal{I}_x$. Fig. 1 shows a tree and Fig. 2a shows an i-tree.

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. The breadth-first traversal of such a tree defines a total ordering of its nodes. We then consider that the set of nodes of an (i-)tree is always the set of integers $\mathbb{N}$. The root is 0 and n is the $(n+1)$-th node visited by the search. We write $n \underset{\mathcal{T}}{\longrightarrow} m$ if and only if m is a child of n in $\mathcal{T}$.

Let $\mathcal{I}$ be an (infinite) i-tree (of finite degree). The sequence $\boldsymbol{s}$ of the degrees of the nodes of $\mathcal{I}$ visited in the breadth-first search of $\mathcal{I}$ is called the *signature* of $\mathcal{I}$ and is *characteristic* of $\mathcal{I}$, that is, one can compute $\mathcal{I}$ from $\boldsymbol{s}$ (*cf.* Proposition 1).

By convention, *the signature of a tree $\mathcal{T}$ is always that of the corresponding i-tree $\mathcal{I}$.*

In this paper, we are interested in signatures that are purely periodic. We call the period of a periodic signature *a rhythm*.

Rhythms Given two integers n and m such that $m > 0$, we denote by $\frac{n}{m}$ their division in $\mathbb{Q}$; by $n \div m$ and $n \,\%\, m$ respectively the quotient and the remainder of the Euclidean division of n by m, that is verifying $n = (n \div m)\, m + (n \,\%\, m)$ and $0 \leqslant (n \,\%\, m) < m$. We also denote the integer interval $\{n, (n+1), \ldots, m\}$ by $[\![n, m]\!]$.

Definition 1. *Let p and q be two integers with $p > q \geqslant 1$.*

(i) *We call* rhythm *of directing parameter (q, p), a q-*tuple $\mathbf{r}$ *of non-negative integers* whose sum is *p:*

$$\mathbf{r} \;=\; (r_0, r_1, \ldots, r_{q-1}) \quad \textit{and} \quad \sum_{i=0}^{q-1} r_i \;=\; p \;\; .$$

(ii) *We say that a rhythm* $\mathbf{r}$ *is* valid *if it satisfies the following equation:*

$$\forall j \in [\![0, q-1]\!] \qquad \sum_{i=0}^{j} r_i \;>\; j+1 \;\; . \tag{1}$$

(iii) *We call* growth ratio *of* $\mathbf{r}$ *the rational number $z = \frac{p}{q}$, also written $z = \frac{p'}{q'}$ where p' and q' are coprime; it is always greater than* 1.

Examples of rhythms of growth ratio $\frac{5}{3}$ are $(2,2,1)$, $(3,1,1)$, $(1,2,2)$, $(3,0,2)$, $(2,1,3,0,0,4)$; all but the third one are valid; the directing parameter is (3,5) for the first four, and (6,10) for the last one.

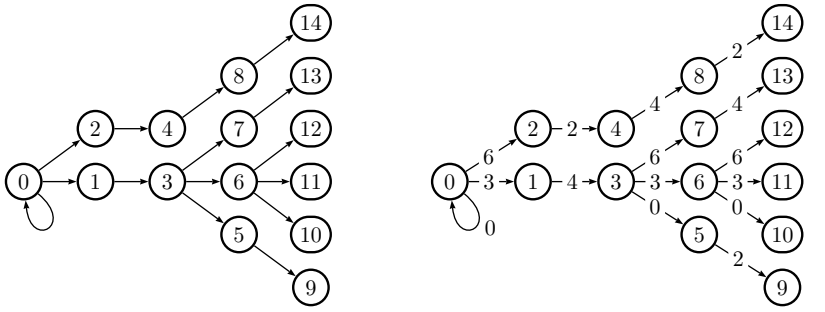
In the following, whenever the reference to a rhythm $\mathbf{r} = (r_0, r_1, \ldots, r_{q-1})$ is clear, we denote simply by R_j the partial sum of the first j components of $\mathbf{r}^\omega$:

$$\forall j \in \mathbb{N} \qquad R_j \;\; = \;\; \sum_{i=0}^{j-1} r_{i \,\%\, q} \qquad \big(= R_{j-1} + r_{(j-1) \,\%\, q} \qquad \text{if } j > 0\big) \;\; .$$

Generating Trees by Rhythm An (i-)tree can be 'reconstructed' from its signature $\boldsymbol{s}$ (*cf.* [11]), hence in the present case, from its rhythm.

Proposition 1. *Let $\mathbf{r} = (r_0, r_1, \ldots, r_{q-1})$ be a (valid) rhythm. Then, there exists a unique i-tree $\mathcal{I}_\mathbf{r}$ whose signature is $\mathbf{r}^\omega$.*

Proof (Sketch). The i-tree $\mathcal{I}_\mathbf{r}$ is built from $\mathbf{r}$ by a kind of procedure which maintains two integers, n and m, both initialised to 0: n is *the node to be processed* and m is *the next node to be created*. At every step of the procedure, $\mathbf{r}_{(n \,\%\, q)}$ nodes are created: the nodes m, $(m+1)$, …, $(m + \mathbf{r}_{(n \,\%\, q)} - 1)$, and the corresponding arcs from n to every new node are created. Then n is incremented by 1, and m by $\mathbf{r}_{(n \,\%\, q)}$. It is verified by induction that at every step, m is equal to R_n. In



(a) The i-tree $\mathcal{I}_{(3,1,1)}$ (b) The padded language $0^* L_{(3,1,1)}$

Fig. 2: Tree and language generated by the rhythm $(3, 1, 1)$

particular, since R_0 is an empty sum hence equal to 0, the root 0 of $\mathcal{I}_{\mathbf{r}}$ is a child of itself. The next equation then gives an explicit definition of $\mathcal{I}_{\mathbf{r}}$:

$$\forall n, m \in \mathbb{N} \qquad n \underset{\mathcal{I}_{\mathbf{r}}}{\longrightarrow} m \iff R_n \leqslant m < R_{n+1} \ . \tag{2}$$

We denote by $\mathcal{T}_{\mathbf{r}}$ the tree resulting from the removal from $\mathcal{I}_{\mathbf{r}}$ of the loop on its root and call respectively $\mathcal{T}_{\mathbf{r}}$ and $\mathcal{I}_{\mathbf{r}}$ the tree and i-tree *generated by* $\mathbf{r}$. Fig. 2a shows $\mathcal{I}_{(3,1,1)}$ and Fig. 1 shows $\mathcal{T}_{(2,1)}$ (if one forgets the labels on the arcs).

The *validity* of the rhythm is the necessary and sufficient condition for m to always be greater than n in the course of the execution of the procedure, that is, a node is always 'created' before being 'processed', or, equivalently, for the i-tree described in Proposition 1 be infinite.

A direct consequence of the proof is that q consecutive nodes of $\mathcal{I}_{\mathbf{r}}$ (in the breadth-first traversal) have p (consecutive) children, hence the name *growth ratio* given to the *number* $\frac{p}{q}$. More precisely, the following holds.

Lemma 1. *Let $\mathcal{I}_{\mathbf{r}}$ be the i-tree generated by the rhythm $\mathbf{r}$ of directing parameter (q, p). Then, for all n, m in $\mathbb{N}$:* $n \underset{\mathcal{I}_{\mathbf{r}}}{\longrightarrow} m \iff (n+q) \underset{\mathcal{I}_{\mathbf{r}}}{\longrightarrow} (m+p)$.

Generating Languages by Rhythm and Labelling If the arcs of an i-tree $\mathcal{I}$ are labelled then $\mathcal{I}$ also defines the sequence $\boldsymbol{\lambda}$ of the labels of the arcs as they are visited in the breadth-first search; conversely, $\mathcal{I}$ as well as its branch language, will be determined by the pair $(\boldsymbol{s}, \boldsymbol{\lambda})$.

In this paper, labels are digits, that is, integers, hence naturally ordered. The labelling of $\mathcal{I}$ has to be consistent with the order of $\mathcal{I}$, that is, the children of every node are in the same order as the labels of their incoming arcs.

We consider here periodic signatures $\boldsymbol{s} = \mathbf{r}^\omega$ where $\mathbf{r}$ is a rhythm of directing parameter (q, p). We then will consider pairs $(\boldsymbol{s}, \boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = \boldsymbol{\gamma}^\omega$ where $\boldsymbol{\gamma}$ is a sequence of letters (digits) *of length p*.

It follows from Lemma 1 that the labelling is consistent on the whole tree if and only if it is consistent on the first q nodes, hence on the first p arcs.

Let $\boldsymbol{\gamma} = u_0\, u_1 \cdots u_{q-1}$ be the factorisation of $\boldsymbol{\gamma}$ *induced by* $\mathbf{r}$, that is, satisfying $|u_i| = r_i$ for every i, $0 \leqslant i < q$. Note that $u_i = \varepsilon$ if $r_i = 0$. The labelling $\boldsymbol{\gamma}$ is then *consistent with* $\mathbf{r}$ if and only if each u_i is increasing[4] and the pair $(\mathbf{r}, \boldsymbol{\gamma})$ is *valid* if in addition $\mathbf{r}$ is valid.

For instance, the labelling $\boldsymbol{\gamma} = (0, 3, 6, 4, 2)$ is consistent with the rhythm $\mathbf{r} = (3, 1, 1)$ since $u_0 = (0, 3, 6)$, $u_1 = (4)$ and $u_2 = (2)$ are all increasing and $u_0\, u_1\, u_2$ is the factorisation of $\boldsymbol{\gamma}$ induced by $\mathbf{r}$.

We denote by $\mathcal{I}_{(\mathbf{r},\boldsymbol{\gamma})}$ the labelled i-tree *generated by* a rhythm $\mathbf{r}$ of directing parameter (q, p) and a labelling $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1})$ consistent with $\mathbf{r}$. The labels of the arcs of $\mathcal{I}_{(\mathbf{r},\boldsymbol{\gamma})}$ are determined by

$$\forall n, m \in \mathbb{N} \qquad n \xrightarrow[\mathcal{I}_{\mathbf{r}}]{a} m \quad \text{implies } a = \gamma_{(m \,\%\, p)} \text{ which belongs to } u_{(n \,\%\, q)} \ . \tag{3}$$

By convention, we denote by $L_{(\mathbf{r},\boldsymbol{\gamma})}$ the branch language of the **tree** $\mathcal{T}_{(\mathbf{r},\boldsymbol{\gamma})}$ rather than the one of **i-tree** $\mathcal{I}_{(\mathbf{r},\boldsymbol{\gamma})}$, and we call it *the language generated by* $(\mathbf{r}, \boldsymbol{\gamma})$. The branch language of $\mathcal{I}_{(\mathbf{r},\boldsymbol{\gamma})}$ is thus $z^* L_{(\mathbf{r},\boldsymbol{\gamma})}$ where $z = \gamma_0$ is the label of the loop $0 \longrightarrow 0$ in $\mathcal{I}_{(\mathbf{r},\boldsymbol{\gamma})}$ and we call it the *padded* language generated by $(\mathbf{r}, \boldsymbol{\gamma})$.

For instance, the language generated by $\mathbf{r} = (2, 1)$ and $\boldsymbol{\gamma} = (0, 2, 1)$ is shown in Fig. 1 and the padded language generated by $\mathbf{r} = (3, 1, 1)$ and $\boldsymbol{\gamma} = (0, 3, 6, 4, 2)$ in Fig. 2b .

Let L be a prefix-closed language over an ordered alphabet A and $\mathcal{T}_L$ its associated labelled tree (whose set of nodes is then $\mathbb{N}$). The enumeration of L in the radix order is then equivalent to the breadth-first traversal of $\mathcal{T}_L$. This ordering of L is precisely the idea underlying the notion of Abstract Numeration System (ANS) as defined by Lecomte and Rigo (*cf.* [7, 8]). An ANS is a language L over an ordered alphabet and in this system every integer n is represented by the $(n+1)$-th word of L in the radix order; this word is denoted by $\langle n \rangle_L$. The integer representations in the ANS L and the nodes of the tree $\mathcal{T}_L$ are thus linked by: $\langle 0 \rangle_L = \varepsilon$ and

$$\forall n \in \mathbb{N}\,,\ \forall m \in \mathbb{N}^+\,,\ \forall a \in A \qquad \langle n \rangle_L\, a \ = \ \langle m \rangle_L \quad \Longleftrightarrow \quad n \xrightarrow[\mathcal{T}_L]{a} m \ . \tag{4}$$

3 From Rational Base Numeration Systems to Rhythms

Integer and Rational Base Numeration Systems Let p be an integer, $p \geqslant 2$, and $A_p = [\![0, p-1]\!]$ the alphabet of the first p digits. Every word $w = a_n\, a_{n-1} \cdots a_0$ of A_p^* is given a value n in $\mathbb{N}$ by the *evaluation function* π_{p}: $\pi_{\mathrm{p}}(a_n\, a_{n-1} \cdots a_0) \ = \ \sum_{i=0}^{n} a_i\, p^i$, and w is a p-development of n. Every n in $\mathbb{N}$ has a unique *p-development* without leading 0's in A_p^*: it is called the *p-representation* of n and is denoted by $\langle n \rangle_p$. The *p-representation* of n can be computed from left-to-right by a greedy algorithm, and also from *right-to-left* by iterating the Euclidean division of n by p, the digits a_i being the successive remainders. The language of the p-representations of the integers is the regular language $L_p = \{\langle n \rangle_p \mid n \in \mathbb{N}\} = (A_p \setminus 0)\, A_p^*$.

[4] A word $a_0\, a_1\, a_2 \cdots a_n$ is *increasing* if $a_0 < a_1 < a_2 < \cdots < a_n$.

Let p and q be two co-prime integers, $p > q > 1$. In [1], these classical statements have been generalised to the case of *numeration system with rational base* $\frac{p}{q}$. The $\frac{p}{q}$-evaluation function $\pi_{\frac{p}{q}}$ is defined by:

$$\forall a_n\, a_{n-1} \cdots a_0 \in A_p^* \qquad \pi_{\frac{p}{q}}(a_n\, a_{n-1} \cdots a_0) \;=\; \sum_{i=0}^{n} \frac{a_i}{q} \left(\frac{p}{q}\right)^i \;,$$

and it is shown that every integer n has a unique $\frac{p}{q}$*-representation* $\langle n \rangle_{\frac{p}{q}}$, that is, a word of A_p^* such that $\pi_{\frac{p}{q}}\left(\langle n \rangle_{\frac{p}{q}}\right) = n$. This representation is computed (from right to left) by the *modified Euclidean division algorithm* as follows: let $N_0 = n$ and, for all $i > 0$,

$$q\, N_i \;=\; p\, N_{(i+1)} + a_i \;, \tag{5}$$

where a_i is the remainder of the Euclidean division of $q\, N_i$ by p, hence belongs to $A_p = [\![0, p-1]\!]$. Since $p > q$, the sequence $(N_i)_{i \in \mathbb{N}}$ is strictly decreasing and eventually stops at $N_{k+1} = 0$. The $\frac{p}{q}$-representation of n is then the word $\langle n \rangle_{\frac{p}{q}} = a_k\, a_{k-1} \cdots a_0$ of A_p^*.

The set $L_{\frac{p}{q}} = \{\langle n \rangle_{\frac{p}{q}} \mid n \in \mathbb{N}\}$ of $\frac{p}{q}$-representations of integers is 'far' from being a regular language. It has a property that we have later called *FLIP* [5] (for *Finite Left Iteration Property*, *cf.* [10]) and which is equivalent (for *prefix-closed* languages) to the fact that it contains no infinite regular subsets (IRS condition of [6]). This implies that $L_{\frac{p}{q}}$ does not meet any kind of iteration lemma and in particular that it is not context-free. It is also shown in [1] that the numeration system with rational base $\frac{p}{q}$ coincide with the ANS $L_{\frac{p}{q}}$.

In many respects, the case of integer base can be seen as a special case of rational base numeration system. The definitions of $\pi_{\frac{p}{q}}$, $\langle n \rangle_{\frac{p}{q}}$ and $L_{\frac{p}{q}}$ coincide with those of π_{p}, $\langle n \rangle_p$ and L_p respectively, when $q = 1$. In the sequel, we consider the base $\frac{p}{q}$ where p and q are two coprime integers verifying $p > q \geqslant 1$, that is, indifferently one numeration system or the other. In particular, the following holds in both integer or rational cases:

$$\forall n \in \mathbb{N}\,,\ \forall m \in \mathbb{N}^+\,,\ \forall a \in A_p \qquad \langle m \rangle_{\frac{p}{q}} = \langle n \rangle_{\frac{p}{q}}\, a \qquad \Longleftrightarrow \qquad a = q\, m - p\, n \;. \tag{6}$$

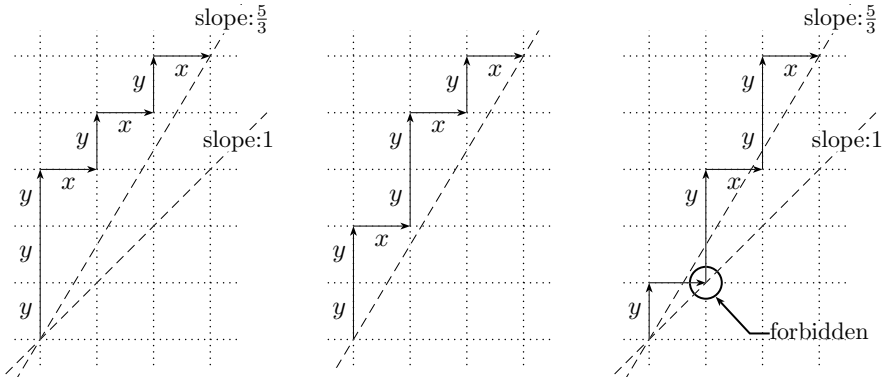
Geometric Representations of Rhythms Rhythms are given a very useful geometric representation as *paths* in the $(\mathbb{Z} \times \mathbb{Z})$-lattice and such paths are coded by *words* of $\{x, y\}^*$ where x denotes an horizontal unit segment and y a vertical unit segment. Hence the name *path* given to a *word* associated with a rhythm.

Definition 2. *Let* $\mathbf{r} = (r_0, r_1, \ldots, r_{q-1})$ *be a rhythm of directing parameter* (q, p). *With* $\mathbf{r}$*, we associate the word* $\mathtt{path}(\mathbf{r})$ *of* $\{x, y\}^*$*:*

$$\mathtt{path}(\mathbf{r}) = y^{r_0} x\, y^{r_1} x\, y^{r_2} \cdots x\, y^{r_{q-1}} x$$

which corresponds to a path from $(0, 0)$ *to* (q, p) *in the* $(\mathbb{Z} \times \mathbb{Z})$*-lattice.*

[5] This property was introduced in [10] under the unproper name of *Bounded Left Iteration Property*, or *BLIP* for short.



(a) rhythm (3,1,1) (b) rhythm (2,2,1) (c) rhythm (1,2,2)

Fig. 3: Words and paths associated with rhythms of directing parameter $(3, 5)$

Fig. 3 shows the paths associated with three rhythms of directing parameter $(3, 5)$. It then appears clearly that Definition 1 (ii) can be restated as 'a rhythm is valid if and only if the associated path is strictly above the line of slope 1 passing through the origin'.

Rhythm and Labelling of Rational Base We introduce $\mathbf{r}_{\frac{p}{q}}$, a particular *rhythm* of directing parameter (q, p) associated with a canonical *labelling* $\boldsymbol{\gamma}_{\frac{p}{q}}$. The former relates to the classical notion of *Christoffel words* while the later results from the generation of $\mathbb{Z}/p\mathbb{Z}$ by q. The remarkable fact is then that the representation language in the $\frac{p}{q}$-numeration system is generated by $(\mathbf{r}_{\frac{p}{q}}, \boldsymbol{\gamma}_{\frac{p}{q}})$.

Christoffel words code the 'best (upper) approximation' of segments the $\mathbb{Z}\times\mathbb{Z}$-lattice and have been studied in the field of combinatorics of words (*cf.* [2]).

Definition 3 ([2]). *The (upper)* Christoffel word *of slope $\frac{p}{q}$, denoted by $\mathbf{w}_{\frac{p}{q}}$, is the label of the path from $(0, 0)$ to (q, p) on the $(\mathbb{Z} \times \mathbb{Z})$-lattice, such that*
- *the path is above the line of slope $\frac{p}{q}$ passing through the origin;*
- *the region enclosed by the path and the line contains no point of $\mathbb{Z} \times \mathbb{Z}$.*

We translate then Christoffel words into rhythms.

Definition 4. *The* Christoffel rhythm *associated with $\frac{p}{q}$, and denoted by $\mathbf{r}_{\frac{p}{q}}$, is the rhythm whose path is $\mathbf{w}_{\frac{p}{q}}$: $\mathtt{path}(\mathbf{r}_{\frac{p}{q}}) = \mathbf{w}_{\frac{p}{q}}$, hence its directing parameter is (q, p).*

Fig. 3b shows the path of $\mathbf{w}_{\frac{5}{3}} = \overline{yy}\,x\,\overline{yy}\,x\,\overline{y}\,x$, the Christoffel word associated with $\frac{5}{3}$; then, $\mathbf{r}_{\frac{5}{3}} = (2, 2, 1)$. Other instances of Christoffel rhythms are $\mathbf{r}_{\frac{3}{2}} = (2, 1)$, $\mathbf{r}_{\frac{4}{3}} = (2, 1, 1)$ and $\mathbf{r}_{\frac{12}{5}} = (3, 2, 3, 2, 2)$. The definition of Christoffel words yields the following proposition on rhythms.

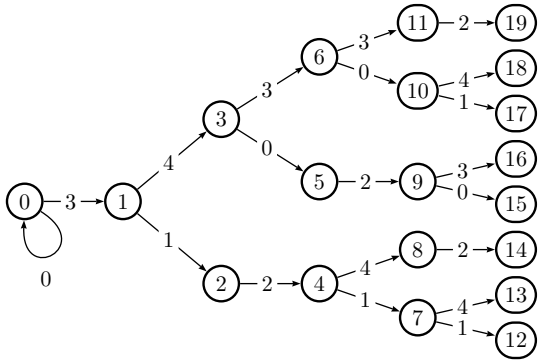


Fig. 4: The padded language $0^*L_{\frac{5}{3}}$ of the representation of integers in base $\frac{5}{3}$

Proposition 2. *Given a base $\frac{p}{q}$ of rhythm $\mathbf{r}_{\frac{p}{q}} = (r_0, r_1, \cdots, r_{q-1})$, for every integer $k \leqslant q$, the partial sum R_k of the first k components of $\mathbf{r}$ is equal to the smallest integer greater than $k\frac{p}{q}$.*

Since p and q are coprime integers, q is a generator of the group $\mathbb{Z}/p\mathbb{Z}$ (additive). We denote by $\boldsymbol{\gamma}_{\frac{p}{q}}$ the sequence induced by this generation process:

$$\boldsymbol{\gamma}_{\frac{p}{q}} \;=\; (\,0,\,(q\,\%\,p),\,(2\,q\,\%\,p),\ldots,\,((p-1)\,q\,\%\,p)\,)\;\;.$$

Theorem 1. *Let p and q be two coprime integers, $p > q \geqslant 1$. The language $L_{\frac{p}{q}}$ of the $\frac{p}{q}$-representations of the integers is generated by the rhythm $\mathbf{r}_{\frac{p}{q}}$ and the labelling $\boldsymbol{\gamma}_{\frac{p}{q}}$.*

For instance, $L_{\frac{3}{2}}$, shown in Fig. 1, is built with the rhythm $\mathbf{r}_{\frac{3}{2}} = (2,1)$ and the labelling $\boldsymbol{\gamma}_{\frac{3}{2}} = (0,2,1)$ while the padded language $0^*L_{\frac{5}{3}}$, shown in Fig. 4, is built with the rhythm $\mathbf{r}_{\frac{5}{3}} = (2,2,1)$ and the labelling $\boldsymbol{\gamma}_{\frac{5}{3}} = (0,3,1,4,2)$.

The proof of Theorem 1 requires additional definitions and statements. We define the sequence of integers $e_0, e_1, \ldots, e_{q-1}$ such that e_j is the difference between the approximation $R_k = (r_0 + r_1 + \cdots + r_{k-1})$ and the point of the associated line of the respective abscissa, that is $(k\,\frac{p}{q})$. This difference is a rational number smaller than 1 and whose denominator is q, in order to obtain an integer we multiply it by q:

$$\forall k \in [\![0, q-1]\!] \qquad e_k \;\;=\;\; q\,R_k - k\,p\;\;. \tag{7}$$

Below are compiled basic properties of the r_j's and e_j's that follow directly from Proposition 2 and Equation (7).

Property 1. Let $\mathbf{r}_{\frac{p}{q}} = (r_0, r_1, \ldots, r_{q-1})$ be the Christoffel rhythm of slope $\frac{p}{q}$. For every integer k in $[\![0, q-1]\!]$, it holds:

(a) e_k belongs to $[\![0, q-1]\!]$;

(b) r_k is the smallest integer such that $q\,r_k + e_k \geqslant p$;
(c) $e_{k+1} = e_k + q\,r_k - p$.

Lemma 2. *For every integer $n > 0$ (resp. $n = 0$), the smallest letter a of A_p such that $\langle n \rangle_{\frac{p}{q}}\, a$ is in $L_{\frac{p}{q}}$ is $e_{(n \,\%\, q)}$ (resp. $e_0 + q$).*

Proof. Let n be positive integer and k its congruence class modulo q. Letters a such that $\langle n \rangle_{\frac{p}{q}}\, a$ belongs to $L_{\frac{p}{q}}$ are congruent modulo q (*cf.* Equation (6)). Since e_k is in $[\![0, q-1]\!]$ (Property 1a), it is enough to prove that e_k is an outgoing label of n. From Equation (6), it is the case if $(n\,p + e_k)$ is a multiple of q or, equivalently if $(k\,p + e_k)$ is a multiple of q, which follows from the definition of e_k (Equation (7)).

For $n = 0$, $e_0 = 0$ and although the equation $e_0 = q\,m - p\,n$ is verified for some integer m, that integer is $m = 0$. It then follows from Equation (6) that $\langle 0 \rangle_{\frac{p}{q}}\, e_0$ does **not** belong to $L_{\frac{p}{q}}$ (since m is not positive). The reasoning of the previous paragraph then works for $(e_0 + q)$.

Proposition 3. *For every integer $n > 0$ (resp. $n = 0$), there are exacly $r_{(n \,\%\, q)}$ (resp. $(r_0 - 1)$) letters a of A_p such that $\langle n \rangle_{\frac{p}{q}}\, a$ belongs to $L_{\frac{p}{q}}$.*

Proof. Let n be positive integer and k its congruence class modulo q. From Property 1b, r_k is the smallest integer such that $q\,r_k + e_k > p$. It follows that for all k in $[\![0, r_k - 1]\!]$ $(e_k + q\,k) < p$ and that $e_k + q\,r_k > p$.

The set $S = \{e_k, (e_k + q), \ldots, (e_k + q\,(r_k - 1))\}$ contains all the letters of A_p that are congruent to e_k modulo q. Since $\langle n \rangle_{\frac{p}{q}}\, e_k$ belongs to $L_{\frac{p}{q}}$ (Lemma 2), it follows from Equation (6) that

$$S \;=\; \{\, a \in A_p \mid \langle n \rangle_{\frac{p}{q}}\, a \in L_{\frac{p}{q}} \,\} \;\; .$$

The set S is of cardinal r_k, concluding the case $n > 0$.

The proof is similar in the case where $n = 0$, except that the smallest letter is $(e_0 + q)$ instead of e_0 (Lemma 2).

The next proposition follows directly from Equation (6).

Proposition 4. *For every positive integer m, the rightmost letter of $\langle m \rangle_{\frac{p}{q}}$ is equal to $(q\,m) \,\%\, p$.*

Proposition 3 yields that the rhythm of $L_{\frac{p}{q}}$ is indeed $\mathbf{r}_{\frac{p}{q}}$ and Proposition 4 that its labelling is $\boldsymbol{\gamma}_{\frac{p}{q}}$, hence concluding the proof of Theorem 1.

The next statement gives a different way to compute $\boldsymbol{\gamma}_{\frac{p}{q}}$; its generalisation in the next section (Definition 5) to arbitrary rhythms will be instrumental in the proof of Theorem 2.

Proposition 5. *Let $\mathbf{r}_{\frac{p}{q}}$ be a Christoffel rhythm and $\boldsymbol{\gamma}_{\frac{p}{q}} = \gamma_0\gamma_1\cdots\gamma_{(p-1)}$ the associated labelling. We denote by $\boldsymbol{\gamma}_{\frac{p}{q}} = u_0\,u_1\,\cdots\,u_{q-1}$ the factorisation of $\boldsymbol{\gamma}_{\frac{p}{q}}$ induced by $\mathbf{r}_{\frac{p}{q}}$. Then, $\gamma_0 = 0$ and, for all integer i, $0 \leqslant i < (p-1)$,*

– *if the letters* γ_i *and* $\gamma_{(i+1)}$ *belong to the same factor* u_j *then* $\gamma_{(i+1)} = \gamma_i + q$ *;*
– *otherwise,* $\gamma_{(i+1)} = \gamma_i + q - p$ *.*

Proof. We denote by $c_0\, c_1 \cdots c_{p-1}$ the integers computed by the recursive algorithm of the proposition:

$$c_i \;=\; q\, i - p\, j \qquad \text{if } \gamma_i \text{ is a letter of the factor } u_j \;\; .$$

It should be noted that $c_i \equiv i\, q \;[p]$, hence that $c_i \equiv \gamma_i \;[p]$; it is then enough to show that $0 \leqslant c_i < p$ for every integer $i < p$.

Let us take $i, j > 0$ such that $\gamma_0\gamma_1 \cdots \gamma_{i-1} = u_0 u_1 \cdots u_{j-1}$, a word of length $i = R_j$. It follows from Proposition 2 that $i = \lceil j\frac{p}{q} \rceil$, or, in other word, that $j\, p - q \leqslant q\,(i-1) < j\, p$. Since γ_{i-1} is the last letter of u_{j-1}

$$c_{i-1} = q\,(i-1) - p\,(j-1) \qquad \text{hence} \qquad (p-q) \;\leqslant\; c_{(i-1)} \;<\; p \;\; ,$$

and since γ_i is the first letter of u_j

$$c_i \;=\; (c_{(i-1)} + q - p) \qquad \text{hence} \qquad 0 \;\leqslant\; c_i \;<\; (q-1) \;\; .$$

We just proved that the first letter of every factor u_k is non-negative and that its last letter is stricly smaller than p. Since every factor is increasing (each letter being equal to the previous letter plus q), every letter a of every factor satisfies $0 \leqslant a < p$.

4 From Rhythms Back to Rational Bases

We now establish a kind of converse of Theorem 1. With an arbitrary rhythm is associated a rational base (its growth ratio) and a *special* labelling. We consider the language generated by this rhythm and labelling as an abstract numeration system and show that it features a rule much like Equation (6). We finally show that this abtract numeration system is simply a rational base on a non-canonical alphabet (Theorem 2)

In this section, p and q are two integers, $p > q \geqslant 1$, *not necessarily coprime*, and $\mathbf{r}$ is a rhythm of directing parameter (q, p). As in Definition 1, we denote by p' and q' their respective quotient by their gcd.

Special Labelling The next definition is a generalisation of the labelling of rational base for arbitrary rhythms; it is based on the characterisation given by Proposition 5 but is more complicated in order to take into accounts the possible components equal to 0 appearing in the rhythm.

Definition 5. *We call* special labelling *(associated with* $\mathbf{r}$*), and denote by* $\boldsymbol{\gamma}_{\mathbf{r}} = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1})$*, the sequence of digits of length* p *defined as follows. First* $\gamma_0 = 0$*. Second, we denote by* $\boldsymbol{\gamma}_{\mathbf{r}} = \; u_0\, u_1 \cdots u_{q-1}$ *the factorisation of* $\boldsymbol{\gamma}_{\mathbf{r}}$ *induced by* $\mathbf{r}$ *(for all* i*,* $0 \leqslant i < p$*,* $|u_i| = r_i$*). Then, for every* i*,* $0 \leqslant i < p-1$*, if* k *and* j *are the indices such that* γ_i *belongs to* u_k *and* γ_{i+1} *belongs to* u_{k+j}*, then* $\gamma_{i+1} = \gamma_i + q' - j\, p'$*.*

Example 1. Let $\mathbf{r} = (3, 1, 1)$; its directing parameter is $(3, 5)$, hence $p = p' = 5$, $q = q' = 3$ and the computation of $\boldsymbol{\gamma}_{\mathbf{r}}$ is given below, on the left. Within a factor u_i, the difference between two consecutive digits is $3 (= q')$, otherwise it is $-2\,(= q' - p')$.

$$\begin{array}{rl|l}
\mathbf{r} = & (3, \quad 1, \quad 1) & (4, \quad 0, \quad 0, \quad 2) \\
\boldsymbol{\gamma}_{\mathbf{r}} = & (\overbrace{0, 3, 6}^{u_0}, \overbrace{4}^{u_1}, \overbrace{2}^{u_2}) & (\overbrace{0, 2, 4, 6}^{u_0}, \overbrace{\quad}^{u_1}, \overbrace{\quad}^{u_2}, \overbrace{-1, 1}^{u_3})
\end{array}$$

Let now $\mathbf{r} = (4, 0, 0, 2)$; its directing parameter is $(4, 6)$, $p' = 3$, $q' = 2$ and the computation of $\boldsymbol{\gamma}_{\mathbf{r}}$ is given above, on the right. Within a factor u_i, the difference between two consecutive digits is $+2 (= +q')$; the fourth digit belongs to u_0 and the fifth to u_3: the difference between the two is $-7 (= +q' - 3\,p')$.

It directly follows from Definition 5 that $\boldsymbol{\gamma}_{\mathbf{r}}$ *is always consistent with* $\mathbf{r}$.

Notation. We denote by $L_{\mathbf{r}}$ the language generated by a rhythm $\mathbf{r}$ and the associated special labelling $\boldsymbol{\gamma}_{\mathbf{r}}$, that is, $L_{\mathbf{r}} = L_{(\mathbf{r}, \boldsymbol{\gamma}_{\mathbf{r}})}$.

Non-Canonical Representation of Integers If $\mathbf{r}$ happens to be a Christoffel rhythm, then, by Theorem 1, $L_{\mathbf{r}}$ is equal to $L_{\frac{p'}{q'}}$ (which, in this case, is also $L_{\frac{p}{q}}$). The key result of this work is that $L_{\mathbf{r}}$ and $L_{\frac{p'}{q'}}$ are indeed of the same kind.

Theorem 2. *Let* $\mathbf{r}$ *be a rhythm of directing parameter* (q, p) *and* $\frac{p'}{q'}$ *the reduced fraction of* $\frac{p}{q}$*. Then, the language* $L_{\mathbf{r}}$ *is a set of representations of the integers in the rational base* $\frac{p'}{q'}$ *using a non-canonical set of digits.*

The proof of Theorem 2 is sketched below. Let us call $\mathbf{r}$-representation of an integer n, and denote it by $\langle n \rangle_{\mathbf{r}}$, the representation of n in the abstract numeration system $L_{\mathbf{r}}$. We know from Equation (4) that $\langle n \rangle_{\mathbf{r}}$ labels the path from the root 0 to the node n in the labelled tree defined by $L_{\mathbf{r}}$. First we show that the existence of arcs in $L_{\mathbf{r}}$ has a necessary condition similar to those of $L_{\frac{p'}{q'}}$ (*cf.* Equation (6)).

Lemma 3. *Let* $\mathbf{r}$ *be a rhythm of directing parameter* (q, p) *and* $\frac{p'}{q'}$ *the reduced fraction of* $\frac{p}{q}$*. Then, for every integers* n *and* $m > 0$*, it holds:*

$$\langle n \rangle_{\mathbf{r}}\, a \;=\; \langle m \rangle_{\mathbf{r}} \quad \Longrightarrow \quad a \;=\; q'\,m - p'\,n \enspace .$$

The converse of Lemma 3 does not hold in general; it holds only for rhythms (of directing parameter (q, p)) such that p and q are coprime, and for powers of such rhythms. Otherwise, the alphabet of the letters appearing in $\boldsymbol{\gamma}_{\mathbf{r}}$ contains at least two different digits congruent modulo p'; the incoming arc of a given node then depends on its congruence class modulo p (and not only modulo p').

Theorem 2 is then equivalent to the following statement.

Proposition 6. *Let* $\mathbf{r}$ *be a rhythm of directing parameter* (q, p)*,* $\frac{p'}{q'}$ *the reduced fraction of* $\frac{p}{q}$ *and* $\pi_{\frac{p'}{q'}}$ *the evaluation function in the* $\frac{p'}{q'}$*-numeration system. Then, for every integer* n*,* $\pi_{\frac{p'}{q'}}(\langle n \rangle_{\mathbf{r}}) = n$ *holds.*

Proof. By induction on the length of $\langle n\rangle_{\mathbf{r}}$. The equality is obviously verified for $\langle 0\rangle_{\mathbf{r}} = \varepsilon$. Let m be a positive integer and $\langle m\rangle_{\mathbf{r}} = a_{k+1}\,a_k\,a_{k-1}\cdots a_1\,a_0$ its $\mathbf{r}$-representation, that is, a word of $L_{\mathbf{r}}$. The word $a_{k+1}\,a_k\,a_{k-1}\cdots a_1$ is also in $L_{\mathbf{r}}$; it is the $\mathbf{r}$-representation of an integer n strictly smaller than m, verifying $\langle n\rangle_{\mathbf{r}}\,a_0 = \langle m\rangle_{\mathbf{r}}$, hence $n \xrightarrow[L_{\mathbf{r}}]{a_0} m$. On the right hand, by induction hypothesis, $n = \pi_{\frac{p'}{q'}}(\langle n\rangle_{\mathbf{r}})$ and on the other hand, it follows from the previous Lemma 3 that $a_0 = q'm - p'n$, or, equivalently, that $m = \frac{np'+a_0}{q'}$, hence

$$m \quad = \quad \frac{p'}{q'}\pi_{\frac{p'}{q'}}(\langle n\rangle_{\mathbf{r}}) + \frac{a_0}{q} \quad = \quad \pi_{\frac{p}{q}}(\langle n\rangle_{\mathbf{r}}\,a_0) \quad = \quad \pi_{\frac{p'}{q'}}(\langle m\rangle_{\mathbf{r}}) \ .$$

It is shown in [1] that in spite of this 'complexity' of $L_{\frac{p}{q}}$, the *conversion* from any digit-alphabet B into the canonical alphabet A_p is realised by a *finite transducer* exactly as in the case of an integer numeration system (*cf.* also [5]). More precisely:

Theorem 3 ([1]). *For all digit alphabets B, the function $\chi\colon B^* \to A_{p'}^*$, which maps every word w of B^* onto the word of $A_{p'}^*$, which has the same value in the $\frac{p'}{q'}$-numeration system — hence $\pi_{\frac{p'}{q'}}(w) = \pi_{\frac{p'}{q'}}(\chi(w))$ — is a (right sequential) rational function.*

If we write B for the set of digits appearing in $\boldsymbol{\gamma}_{\mathbf{r}}$, Theorem 3 implies in particular that $\chi(L_{\mathbf{r}}) = L_{\frac{p'}{q'}}$. Hence, that the complexity of $L_{\frac{p'}{q'}}$ extends to $L_{\mathbf{r}}$.

Corollary 1. *Let $\mathbf{r}$ be a rhythm of directing parameter (q,p) and $L_{\mathbf{r}}$ the language generated by the pair $(\mathbf{r},\boldsymbol{\gamma}_{\mathbf{r}})$. If $\frac{p}{q}$ is an integer, then $L_{\mathbf{r}}$ is a regular language, otherwise, $L_{\mathbf{r}}$ is a FLIP language.*

Example 2. Given a directing parameter (q,p), let $\mathbf{r}$ be the *extreme* rhythm where all components are 0 but one which is p. The validity condition implies that the positive digit is necessarily the first one: $\mathbf{r} = (p,0,\ldots,0)$ and the associated special labelling is then $\boldsymbol{\gamma}_{\mathbf{r}} = (0,q,(2\,q),\ldots,(p-1)\,q)$. Since every letter of $\boldsymbol{\gamma}_{\mathbf{r}}$ is a multiple of q, we perform a component-wise division of $\boldsymbol{\gamma}_{\mathbf{r}}$ by q and obtain $\boldsymbol{\gamma} = (0,1,2,\ldots,(p-1))$.

The language $L_{(\mathbf{r},\boldsymbol{\gamma})}$ generated by $(\mathbf{r},\boldsymbol{\gamma})$ is then the language of the representations of the integers in a variant (that we call FK after its authors) of $\frac{p}{q}$-numeration systems considered in [4]. In the variant FK, the value of a word u, denoted by $\pi_{\mathrm{FK}}(u)$, is q times its standard evaluation: $\pi_{\mathrm{FK}}(u) = q \times \pi_{\frac{p}{q}}(u)$. This is exactly the behaviour described by Proposition 6, since all digits have been divided by q. This example highlights the soundness of the relationship between rational base numeration system and periodic signature.

5 Extension, Future Work and Conclusion

For sake of simplicity, we have considered here purely periodic signatures and the periodic labellings that go with them. The same techniques as the ones developed

in Section 4 allow to treat the generalisation to *ultimately periodic* which raises no special difficulties and the results established here readily extend. One may even generalise these results to every aperiodic signature whose path (as defined in Sect. 2) is confined to a strip between two parallel lines of slope $\frac{p}{q}$.

Using rhythm often sheds light on problems related to rational base. It is the case for the question of representation of the negative integers, tackled in [4], that may be given a new approach in terms of Christoffel words and their properties.

There is certainly still much to be understood on the relationship between the 'high regularity' of periodic signatures and the apparent disorder or complexity of trees that are generated by these periodic signatures. Some questions, such as statistics of labels along infinite branches, are indeed related to identified problems in number theory that are recognised as very difficult.

We have established in this paper that the infinite trees or languages generated by periodic signatures are completely determined (up to very simple transformations — that is, rational sequential functions) by the growth ratio of the period only and independent of the actual components of the period. This first step was somehow unexpected. It makes the scenery simpler but the call for further investigations on the subject even stronger.

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, Aaron Lauve, Christophe Reutenauer, and Franco Saliola. *Combinatorics on Words: Christoffel Words and Repetition in Words*, volume 27 of *CRM monograph series*. American Math. Soc., 2008.
3. Reinhard Diestel. *Graph Theory*. Springer, 1997.
4. Christiane Frougny and Karel Klouda. Rational base number systems for p-adic numbers. *RAIRO - Theor. Inf. and Applic.*, 46(1):87–106, 2012.
5. Christiane Frougny and Jacques Sakarovitch. Number representation and finite automata. In Valérie Berthé and Michel Rigo, editors, *Combinatorics, Automata and Number Theory*. Cambridge University Press, 2010.
6. Sheila A. Greibach. One counter languages and the IRS condition. *J. Comput. Syst. Sci.*, 10(2):237–247, 1975.
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 Valérie Berthé and Michel Rigo, editors, *Combinatorics, Automata and Number Theory*. Cambridge University Press, 2010.
9. Victor Marsault and Jacques Sakarovitch. Rhythmic generation of infinite trees and languages (full version). In preparation. Preprint available at arXiv:1403.5190.
10. Victor Marsault and Jacques Sakarovitch. On sets of numbers rationally represented in a rational base number system. In Traian Muntean, Dimitrios Poulakis, and Robert Rolland, editors, *CAI 2013*, volume 8080 of *LNCS*. Springer, 2013.
11. Victor Marsault and Jacques Sakarovitch. Breadth-first serialisation of trees and rational languages. In Arseny M. Shur and Mikhail V. Volkov, editors, *DLT 2014*, volume 8633 of *LNCS*. Springer, 2014.