The Boundary of Iterated Morphisms on Free Semi-Groups

Philippe Narbel
L.I.T.P, Institut Blaise Pascal, Paris 7
55-56, 1st fl., 2, Place Jussieu
75251 Paris
e-mail: narbel@litp.ibp.fr

As published in Intern. J. of Algebra and Computation Vol 6, No. 2 (1996) pp.229-260.

(with an errata given in appendix (2001)).

Abstract. This paper[1] introduces a generalized way of defining languages of infinite words by topological means. It focuses on languages generated by iterating morphisms on free semi-groups (also called substitutions or D0L-systems). The main result is an effective construction of the boundary of such languages which leads to a bijective mapping of the boundary onto a regular language. Among the obtained properties are the uncountability of the boundary and the strict quasiperiodicity of its words. We also investigate the decidability of the boundary equality problem and the dynamical system induced by the inversed morphism.

Contents

[1] This paper consists of a full enhancement of two preliminary articles: *The Limit Set of Recognizable Substitution Systems*, STACS'93, LNCS 657, pp.226-236, and *The Boundary of Substitution Systems*, MFCS'93, LNCS 711, pp.577-587.

1 Introduction

Morphisms on free semi-groups are simple transformations which replace single generators by words made of generators. In formal language theory, they are either called *substitutions* or *D0L-systems*, i.e. parallel deterministic context-free systems. For instance, with the generators $\{a, b\}$, the so-called Thue-Morse morphism is given by replacing in a word all the a's by ab and all the b's by ba (e.g. the word $abba$ is transformed into $abbabaab$).

Iterating morphisms is a classical way of generating infinite words. The first who investigated this possibility seemed to be Thue [31]. Later, Morse [22, 23] made use of it to obtain bi-infinite words which represented geodesics on surfaces of negative curvature. This was the founding to what was called *symbolic dynamics* [24]. Indeed, sets of asymptotic words coming from infinite iterations of morphisms were shown as one of the simplest way of constructing *minimal sets* and *strict ergodic systems* by the orbit-closure of a shift (see for instance [14, 4, 20, 27]). In formal language theory, one-way infinite words were generated as fixed points of morphisms (see [29, 30, 6]), and analysis of the subword structure of such words led to a distinct field of research in word combinatorics. From a different point of view, as soon as extensions of *regularity* were given for infinite words, topological means of getting infinite words were devised. In particular, *boundaries* [2] of languages were introduced [2, 17, 18]. Originally, these asymptotic languages contained only words going to infinity to the right, but extensions to include both ways have more recently been made [13, 10].

The aim of this paper is to introduce a generalized and effective way of defining the boundary of a language. The main new structural feature about this boundary is that it inherently contains both one-way and bi-infinite words. We shall focus here on studying such boundaries for languages of finite words generated by iterating morphisms. For that purpose, the *boundary of a morphism* is defined as the boundary of the language obtained by finitely iterating this morphism starting from each generator.

The first result to be presented is that the construction of this boundary can be made effective. This is different from the just cited dynamical-oriented works where non-constructive closure operators were directly used. The point we shall exploit here is that, since the space of all words over a finite alphabet endowed with the bi-infinite product topology is compact, the closure of its subspaces can be obtained by applying a *completion*. The idea is therefore to find a procedure which generates all the needed Cauchy sequences. This is implemented by what we call the *embedding map* which systematically embeds finite words into larger ones. This map has been inspired by an idea coming from *tiling theory* by R. Robinson [28, 16], and formally developed by N.G. De Bruijn [7, 8]. With the assumption that morphisms are *expansive*, i.e. morphism iterations starting from every generator lead to arbitrarily long words, the first theorem is the following:

Theorem. *The boundary of an expansive morphism on a free semi-group can be effectively constructed by the embedding map.*

Since the embedding map can be represented as a forest of infinite trees with a systematic branching structure, the boundary can be mapped onto a regular language of right-infinite words. An important case is when this map is injective, since this leads to a

[2] In the language theory literature, they are called *adherence* sets, although the topological concept of *boundary* has been actually considered.

regular coding of the boundary. This property is shown to hold whenever the morphism is *circular*, i.e. locally invertible. This concept was first introduced by Mignosi and Séébold [21], and is closely related to *recognizability* [20, 27, 25]. A consequence is the second main result which is:

THEOREM. *The boundary of an expansive and circular morphism on a free semi-group can be bijectively mapped onto a regular language of right-infinite words.*

In other words, the completion of a language generated by iterating a morphism can be done *automatically*. According to this theorem, the boundary can be studied through its regular coding language. In particular, we show that this set can be uncountable, and that it only contains *strict quasiperiodic words*. Also, the study of the dynamical system on the boundary induced by the inversed morphism is made much more easier: for instance periodic coding words represent periodic orbits. Finally, by making a strong use of a result from Culik and Harju [5], the problem of knowing whether or not two boundaries are equal is shown decidable for primitive morphisms.

The first part of this paper introduces the topology on pointed and unpointed words, as well as the definition of the boundary of a language, and the on-focus languages generated by iterating morphisms. The second part introduces the embedding map and its representation as a forest of infinite trees whose ends consist of boundary words. This leads us to the first theorem. We also discuss how the asymptotic sets of words in dynamical systems and in $D0L$-systems theories can be related to our way of defining the boundary. Finally, the third part presents how the boundary can be coded into a regular language: First, the forest of trees is labeled and shown regular. Next the circularity property is shown sufficient and necessary so that each path of the embedding trees leads to a distinct boundary word. This gives the second theorem. Its consequences are then discussed.

2 NOTATIONS, DEFINITIONS

2.1 WORDS

$\mathbb{N}$ denotes the positive integers with zero,
$\mathbb{N}^-$ denotes the negative integers with zero,
$\mathbb{Z}$ denotes the integers,
ω is the cardinality of the countable,
A denotes a finite alphabet of generators or symbols,
A^+ denotes the free semi-group generated by A,

Definition 1 *A* **pointed word** *on A is a map $\widehat{w} : I \to A$, where I is any interval of $\mathbb{Z}$ which includes zero.*

$\widehat{w}(n)$ is abbreviated w_n,
w_0 is called **the origin** or the **base point** of $\widehat{w}$,
$|\widehat{w}|$ is the **length** of $\widehat{w}$, and is equal to the cardinality of its source set.
If $\widehat{w} : I \to A$ is a pointed word over A, then,
- if I is finite, then $\widehat{w}$ is said **finite**,
- if $\mathbb{N} \subset I$, $I \neq \mathbb{Z}$, then $\widehat{w}$ is said **right infinite**,
- if $\mathbb{N}^- \subset I$, $I \neq \mathbb{Z}$, then $\widehat{w}$ is said **left infinite**,
- if $I = \mathbb{Z}$, then $\widehat{w}$ is said **bi-infinite**.

${}^{\infty}\widehat{A}^{\infty}$ denotes the language of all the pointed words over A,
$\widehat{A}^{\omega}$ denotes the language of all the pointed right-infinite words over A.
Any subset $\widehat{L}$ of ${}^{\infty}\widehat{A}^{\infty}$ is called a **pointed language**,
$|\widehat{L}|$ is the **cardinality** of the language $\widehat{L}$.
A boldface character indicates the origin of a pointed word. For example, the finite pointed word $\widehat{w} = aa\mathbf{a}bbabbb$ is not equal to $\widehat{v} = aaabb\mathbf{a}bbb$.
Let $\widehat{w} : I \to A$, then $\widehat{v} \subset \widehat{w}$ means that $\widehat{v} : I' \subseteq I \to A$ where I' is an interval of $\mathbb{Z}$ included in I. In this case, the word $\widehat{v}$ is said to be a **block** or a **factor** of $\widehat{w}$. Moreover,

if $I = \{k, .., l\}$, $k \in \mathbb{Z}$, $l \in \mathbb{Z} \cup \{\omega\}$, and $I' = \{k, .., l'\}, l' \leq l$, $\widehat{v}$ is a **left factor**.

if $I = \{k, .., l\}$, $k \in \mathbb{Z} \cup \{-\omega\}$, $l \in \mathbb{Z}$, and $I' = \{k', .., l\}, k \leq k'$, $\widehat{v}$ is a **right factor**.

If I' contains zero, then $\widehat{v}$ is a **pointed factor**.

To homogenize ${}^{\infty}\widehat{A}^{\infty}$, all the words which are not yet bi-infinite are padded to both infinities with some dummy symbol not already in A. This will be useful for the metric definition. Also, this allows us to consider the **shift operator** σ on ${}^{\infty}\widehat{A}^{\infty}$:

$$\sigma(\widehat{w}_n) = \widehat{w}_{n+1}, \quad \forall n \in \mathbb{Z}. \tag{1}$$

Because of the padding, the n-fold shift operation $\sigma^n(\widehat{w})$, $n > 0$, is **well-defined** iff $w_{-n} \in A$, i.e. the word cannot be pointed on its padded part. The shift induces an equivalence relation on ${}^{\infty}\widehat{A}^{\infty}$:

$$\widehat{v} \sim_{\sigma} \widehat{w} \text{ iff there exists } n \in \mathbb{Z} \text{ such that } \widehat{v} = \sigma^n(\widehat{w}).$$

Definition **2** *An* **unpointed word** *is an equivalence class in* ${}^{\infty}\widehat{A}^{\infty}/\sim_{\sigma}$.

The set ${}^{\infty}A^{\infty}$ consists of all unpointed words. If $\widehat{w}$ is a pointed word, then w denotes its equivalence class (up to now, symbols with hats will always indicate pointed words and pointed languages). Any subset L of ${}^{\infty}A^{\infty}$ is called a **language**. If $L \subset {}^{\infty}A^{\infty}$ is considered, its **pointed counterpart**, $\widehat{L}$, consists of all the different pointings of the words in L, so that $L = \widehat{L}/\sim_{\sigma}$ holds.

Let us remark that the finite and the one-way infinite pointed words have a canonical writing for their equivalence classes. This comes from the fact that they have at least one end, which may be used as a distinctive point. For example, the word aab is the equivalence class of $\{\mathbf{a}ab, a\mathbf{a}b, aa\mathbf{b}\}$, and $ab\mathbf{a}bbaaa....$ is in the equivalence class $ababbaaa.....$ This is not the case for bi-infinite words which cannot in general be written down.

2.2 The Metric Structure

The interest in using pointed words is that there exists a convenient metric, i.e. there is an easy way of structuring ${}^{\infty}\widehat{A}^{\infty}$ into a topological space.

To compare two pointed words, the following metric relies on the length of the longest common factor around their origins: let $\widehat{u}, \widehat{v}$ be in ${}^{\infty}\widehat{A}^{\infty}$, then

$$d(\widehat{u}, \widehat{v}) = \begin{cases} 0 & \text{iff } \widehat{u} = \widehat{v} \\ 2^{-n} & \text{otherwise} \end{cases} \tag{2}$$

where n is the largest nonnegative integer such that $u_k = w_k$, $|k| \leq n$. Note that the padding with the dummy symbol allows finite and infinite words to be compared. For instance, if $\widehat{u} = aaa\mathbf{a}bbabbb$ and $\widehat{v} = \cdots bbbbaaa\mathbf{a}bbabbb\cdots$, then $d(\widehat{u}, \widehat{v}) = 2^{-3}$. The map

d is the definition of an **ultrametric** since it satisfies the following triangle inequality: let $\widehat{u}, \widehat{v}, \widehat{w}$ be in ${}^{\infty}\widehat{A}^{\infty}$, then

$$d(\widehat{u}, \widehat{w}) \leq max\{d(\widehat{u}, \widehat{v}), d(\widehat{v}, \widehat{w})\}.$$

Here are the main properties of the metric space $({}^{\infty}\widehat{A}^{\infty}, d)$:

- Since d is an ultrametric, the space $({}^{\infty}\widehat{A}^{\infty}, d)$ is **totally disconnected** in $\mathbb{R}$ (see for instance [12]), i.e. the connected component of each point contains only this point. As a consequence, it is **zero-dimensional**.
- Since A is finite, the space $({}^{\infty}\widehat{A}^{\infty}, d)$ is **compact**.
- The finite words in ${}^{\infty}\widehat{A}^{\infty}$ are **isolated points**.
- The set of infinite words in ${}^{\infty}\widehat{A}^{\infty}$ is **perfect** [24], i.e. each word is an accumulation point, and therefore is homeomorphic to the **Cantor set**.

2.3 The Boundary

The **boundary** of a language is defined as the familiar topological notion, that is, if $\widehat{L} \subset {}^{\infty}\widehat{A}^{\infty}$,

$$\partial\widehat{L} = Closure(\widehat{L}) \ \cap \ Closure\,(Complement \ \ (\widehat{L}) \ in \ \ {}^{\infty}\widehat{A}^{\infty})$$

Since $({}^{\infty}\widehat{A}^{\infty}, d)$ is compact, the closure of its subsets can be obtained by a completion, i.e. by considering every limit of every Cauchy sequence. Recall that such a sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ in ${}^{\infty}\widehat{A}^{\infty}$, denoted by (α_n), is such that:

$$\forall \epsilon > 0\ , \ \ \exists k \quad \text{such that} \quad d(\alpha_n, \alpha_m) < \epsilon, \quad \forall n, m > k. \tag{3}$$

Note also that since $({}^{\infty}\widehat{A}^{\infty}, d)$ is ultrametric, Cauchy sequences have a simple form which is such that $d(\alpha_n, \alpha_{n+1}) \to 0$. Besides, we can use the convention that α_0 is a single letter in A so that it corresponds to the origin of all the α_n's. Such a sequence can be described by identifying longer and longer factors of the α_n's around their origins:

Example **2.1** *If $\alpha_0 = \mathbf{b}$ and $\alpha_n = a^{(2^{n-1}-1)}\mathbf{b}a^{(2^n-1)}b$, for $n > 0$, then this sequence can be represented as:*

$$\begin{array}{c} \vdots \\ aaa\mathbf{b}aaaaaaab \\ a\mathbf{b}aaab \\ \mathbf{b}ab \\ \mathbf{b} \end{array}$$

Its limit word is ${}^{\omega}a\mathbf{b}a^{\omega}$.

Therefore, a Cauchy sequence (α_n) in $({}^{\infty}\widehat{A}^{\infty}, d)$ is determined by an **identified factors sequence** denoted by (α'_n), such that $\alpha'_n \subseteq \alpha_n$, for all $n \in \mathbb{N}$ and such that:

1. $\alpha'_0 = \alpha_0 \in \widehat{A}$,
2. $\alpha'_n \subset \alpha'_{n+1}$, for all $n \in \mathbb{N}$, such that
   $\alpha_{n+1} = \delta_n\gamma_n\alpha'_n\beta_n\eta_n, \ \ \beta_n, \gamma_n, \delta_n, \eta_n \in A^+,$ or are empty,
   $\alpha'_{n+1} = \gamma_n\alpha'_n\beta_n,$

In the last example, the identified factors are given by $\alpha'_n = a^{(2^{n-1}-1)}\mathbf{b}a^{(2^n-1)}$. The next example suggests that limits of Cauchy sequences are not necessarily bi-infinite:

Example **2.2** *If* $\alpha_0 = \mathbf{a}$ *and* $\alpha_n = b\mathbf{a}b^n$*, for* $n > 0$*:*

$$\begin{array}{c} \vdots \\ b\mathbf{a}bbb \\ b\mathbf{a}bb \\ b\mathbf{a}b \\ \mathbf{a} \end{array}$$

Here, its limit word is $b\mathbf{a}b^\omega$.

Clearly the limit words may be either right, left or bi-infinite words.

Now let us denote by $\Im(\widehat{L})$ the set of all distinct limits of Cauchy sequences in $\widehat{L}$. The **completion points** of $\widehat{L}$ consists of every word which is in $\Im(\widehat{L})$ but not in $\widehat{L}$, i.e. the points of $\Im(\widehat{L}) \setminus \widehat{L}$.

Remark **2.3** *Let* $\widehat{L} \subset \widehat{A}^+$*. Then,* $\partial\widehat{L} = \Im(\widehat{L}) \setminus \widehat{L}$.

Proof. Every finite word is isolated in $({}^\infty\widehat{A}^\infty, d)$. Therefore, its complementary language in ${}^\infty\widehat{A}^\infty$ is closed, and a word which belongs to the boundary cannot belong to $\widehat{A}^+$. $\diamondsuit$

Corollary **2.4** *Let* $\widehat{L} \subseteq \widehat{A}^+$ *be of finite cardinality. Then* $\partial\widehat{L} = \emptyset$.

Corollary **2.5** *Let* $\widehat{L} \subseteq \widehat{A}^+$ *be of infinite cardinality. Then,* $\partial\widehat{L}$ *consists of the set of the limits of the Cauchy sequences* (α_n) *on* $\widehat{L}$ *such that* $\lim_{n\to\infty}(\alpha_n)$ *is an infinite word.*

Corollary **2.6** *Let* $\widehat{L} \subseteq \widehat{A}^+$ *be of infinite cardinality. Then,* $\partial\widehat{L} \neq \emptyset$.

In view of the Cauchy sequence examples, the boundary $\partial\widehat{L}$ decomposes into three sets respectively containing only right, left or bi-infinite completion words, and respectively denoted by $L\partial\widehat{L}$, $R\partial\widehat{L}$ and $Bi\partial\widehat{L}$, that is:

$$\partial\widehat{L} = L\partial\widehat{L} \ \cup \ R\partial\widehat{L} \ \cup \ Bi\partial\widehat{L}.$$

Let us define the unpointed boundary sets. First, it is necessary to have a **shift invariance** property:

Remark **2.7** *Let* $L \subseteq {}^\infty A^\infty$ *and* $\widehat{L} \subseteq {}^\infty\widehat{A}^\infty$ *its pointed counterpart. Then*

$$\text{if } \widehat{w} \in \partial\widehat{L} \text{ then } \sigma^m(\widehat{w}) \in \partial\widehat{L}, \ \ \forall m \in \mathbb{Z}.$$

Proof. If $\widehat{w} \in \partial\widehat{L}$, there is a sequence (α_n), such that $\lim_{n\to\infty}(\alpha_n) = \widehat{w}$, and the identified factor sequence (α'_n) must have a growth function $|\alpha'_n|$ which strictly increases with n. Hence, if $\sigma^m(\widehat{w})$ is well-defined, i.e. the origin is a letter in A, there is an index $p > 0$ such that the origin of $\sigma^m(\widehat{w})$ is included in α'_p. Hence, just define another sequence (ζ_n) where ζ_0 is equal to the origin of $\sigma^m(\widehat{w})$ and $\zeta_n = \alpha_{n+p}$, for all $n > 0$. $\diamondsuit$

This allows us to define the respective unpointed boundary sets:

Definition **3** *Let* $L \subseteq {}^\infty A^\infty$ *and* $\widehat{L} \subseteq {}^\infty\widehat{A}^\infty$ *its pointed counterpart. The* **unpointed boundary** ∂L *of* L *is given by* $(\partial\widehat{L}/\sim_\sigma)$*. Equivalently,*

$$L\partial L = L\partial\widehat{L}/\sim_\sigma, \quad R\partial L = R\partial\widehat{L}/\sim_\sigma, \quad Bi\partial L = Bi\partial\widehat{L}/\sim_\sigma.$$

Here are two general remarks about the boundary definitions:

1) These definitions can be related to the familiar ones in formal language theory literature (see for instance [2, 13]) but in a slightly different form. First, they were called "adherences", although this was somewhat misleading since this notion is equivalent to the whole closure. Next, these sets were defined according to their factor sets: they were implicitely unpointed for the one-way boundary sets and pointed for the two-way boundary set.

2) The unpointed sets are not always interesting in regard to the actual metrical quotient spaces: they can be reduced to one point, i.e. to the trivial topology. Indeed, quotient spaces of languages in $({}^\infty\widehat{A}^\infty, d)$ are metrically unseparated whenever their words have comparable sets of factors: they cannot be locally distinguished. For instance, this occurs when languages are *locally isomorphic* [19]. Details about this fact can be found in [3] since these languages are instances for which the C^*-algebra framework is proved more informative than the classical topological view. We shall not get further into this, but just indicate when results about pointed boundaries can readily be extended to unpointed ones.

2.4 Iterated Morphisms on Free Semi-Groups and Substitutions

A map $h : A \to A^+$ is a **morphism** on the semi-group A^+ when extended to $h : A^+ \to A^+$, the following holds for all $w \in A^+$:

$$h(w) = h(s_1...s_n) = h(s_1)...h(s_n), \quad s_j \in A.$$

Applying a morphism on some word is also called a **substitution**. For instance, with the letters $\{p, q\}$, the morphism $h(p) = ppq$, $h(q) = pq$ is such that $h(pqp) = ppqpqppq$. The n-fold composition is denoted by h^n.

Definition 4 *Let A be a finite alphabet and h be a morphism on A^+. The* **letter substitution language** *is given by:*

$$L_h = \{w \in A^+ \mid \quad h^n(s) = w, \ \ n \in \mathbb{N}, \ s \in A\}. \tag{4}$$

If w in L_h is generated by $h^n(s)$, then the letter s is called the **father letter** and n is **the order** of this way of generating w. The **letter substitution sub-language of order** n is defined as:

$$L_h^n = \{w \in A^+ \mid \quad h^n(s) = w, \quad s \in A\}. \tag{5}$$

The **letter substitution sub-language with father** s is defined as:

$$L_h(s) = \{w \in A^+ \mid \quad h^n(s) = w, \quad n \in \mathbb{N}\}. \tag{6}$$

The **image set** of every letter by h, that is L_h^1, is denoted by $h(A)$.

We shall make use of the following kinds of morphisms:

- **primitive** if there exists a finite power n such that all letters of A are included in $h^n(s)$, for all $s \in A$.
- **n-power free**, if each word in L_h does not contain any factor u^n with $u \in A^+$, $n > 1$.

- **expansive**, if for every $s \in A$, $|h^n(s)|$ may be arbitrarily large.

We shall call the boundary of a letter substitution language L_h, the **boundary of the morphism** h. Note that the expansive property of a morphism ensures that $\widehat{L}_h$ is an infinite language included in $\widehat{A}^+$, and therefore that $\partial\widehat{L}_h \neq \emptyset$. Note also that letter substitution languages are directly related to a classical type of substitution languages [29, 30] obtained from a single starting word: A $D0L$-**language** [29, 30] is given by a 3-tuple (A, h, t) such that substitution iterations start from a single word t in A^+, that is:

$$D0L_h(t) = \{w \in A^+ \ | \quad h^n(t) = w, \ \ n \in \mathbb{N}\}.$$

Hence, $L_h = \bigcup_{s\in A} D0L_h(s)$.

2.5 Regularity for Infinite Words

Regular languages will be on use here for finite and right-infinite words. We recall here some basic definitions:

A **finite Büchi automaton** for finite words is a 5-tuple (Q, I, F, E, A) where:

Q denotes the set of **states**,
$I, F \subseteq Q$ are the sets of **initial** and **final** states,
E is the set of **edges** included in the set $Q \times A \times Q$,
A is as usual the set of **labels** or **letters**.

The sets Q, E and A are assumed to be finite. A **path** in the automaton is a sequence of consecutive edges (q_i, g_i, q_{i+1}) of E and its **label** is the word $g = g_0g_1...g_n$ included in A^+ (respect. $g = g_0g_1...$ in A^ω). To **accept a finite word** label, its corresponding path must start from some state of I, i.e. $q_0 \in I$, and must end in a state of F, i.e. $q_n \in F$.

For the infinite case, a new set called a **table set** Tab included in the power set of the final states F is added to the automaton: such a 6-tuple (Q, I, F, E, A, Tab) is called a **Muller automaton** [26]. To **accept a right-infinite word** label, its corresponding path must start from some state of I and the set of states which are infinitely visited must belong to Tab.

The set of finite (respect. infinite) words included in A^+(respect. in A^ω) which are accepted by a given automaton is called the **recognized language**. This language is said **regular**.

3 The Effective Construction of the Boundary

3.1 The Embedding Map

A word of $\partial\widehat{L}_h$ is given by the limit of a Cauchy sequence which is an infinite word. Therefore, if there is a systematic way of defining all these sequences, it would lead to an effective description of the set $\partial\widehat{L}_h$. This will be implemented for most boundary words by a simple process which relies on a constrained way of embedding words in each other. Since $\widehat{w} \in \widehat{L}_h$, there exists a symbol $s \in A$ such that $h^n(s) = w$ (recall that w is the associated $\sim_\sigma$-class of $\widehat{w}$); thus a set of words which contains $\widehat{w}$ as a factor is given, first, by determining all the possible choices of embedding the letter s in the words of $h(A)$, i.e. the morphism image of the single letters, second, by applying h^n to these words. As a consequence, this leads to a map which sends the letter sub-language of order n with father letter s, i.e. $\widehat{L}_h^n(s)$, onto a subset of the letter sub-language of order $n+1$, that is to the power set of $\widehat{L}_h^{n+1}$:

Definition **5** *Let h be a morphism defined on A^+. Its* **embedding map** *is defined for all $n \in \mathbb{N}$ and $s \in A$ as a function*

$$\begin{array}{rcl} Emb_h : \widehat{L}_h^n(s) & \to & \text{Power set of } \widehat{L}_h^{n+1} \\ \widehat{w} & \mapsto & \{\widehat{v} \in \widehat{L}_h^{n+1} \mid \widehat{v} = h^n(u)\widehat{w}\, h^n(u'),\ \ usu' \in h(A)\}. \end{array} \tag{7}$$

The indexation of the factors $h^n(u)$ and $h^n(u')$ is induced by $\widehat{w}$.

Example **3.1** *Consider the morphism defined by $h(p) = ppq$, $h(q) = pq$, then $Emb_h(\mathbf{p}) = \{\mathbf{p}pq, p\mathbf{p}q, \mathbf{p}q\}$ and $Emb_h(\mathbf{q}) = \{pp\mathbf{q}, p\mathbf{q}\}$ (recall that the boldface letters indicate the origins of the words). Consider also the case where $\widehat{w} = p\mathbf{p}q$, where $\widehat{w}$ has father p, i.e. $w = h(p)$, then take all the possibilities of embedding p in $h(A)$ (which is exactly $Emb_h(\mathbf{p})$) to obtain $Emb_h(p\mathbf{p}q) = \{p\mathbf{p}qppqpq, ppqp\mathbf{p}qpq, p\mathbf{p}qpq\}$.*

We stress here that the embedding map is not related to a forward application of the morphism: e.g. the word $p\mathbf{p}qpq$ belongs to $Emb_h(p\mathbf{p}q)$.

The embedding map can be recursively iterated: Since each word of $Emb_h^{n-1}(\widehat{w})$ defines its own subset included in the power set of $\widehat{L}_h^{n+1}$, then $Emb_h(Emb_h^{n-1}(\widehat{w}))$ is given by

$$Emb_h^n(\widehat{w}) = \bigcup_{\widehat{u} \in Emb_h^{n-1}(\widehat{w})} \{\widehat{v} \in Emb_h(\widehat{u})\}. \tag{8}$$

3.2 The Embedding Forest

Iterations of the embedding map can be represented as a graph structure:

- The set of nodes is a subset of $\widehat{L}_h \times A \times \mathbb{N}$, where a node $(\widehat{w}, s, n)$ means that $w = h^n(s)$, $s \in A$, $n \in \mathbb{N}$.
- The edges are defined such that $(\widehat{w}, s, n)$ is bound to $(\widehat{v}, t, n+1)$ whenever $\widehat{v} \in Emb_h(\widehat{w})$.

Example **3.2** *Again considering $h(p) = ppq$, $h(q) = pq$, we can see that $(\mathbf{p}, p, 0)$ is bound to $(\mathbf{p}pq, p, 1)$, $(p\mathbf{p}q, p, 1)$, and $(\mathbf{p}q, q, 1)$; also, the node $(p\mathbf{p}q, p, 1)$ is bound to $(p\mathbf{p}qppqpq, p, 2)$, $(ppqp\mathbf{p}qpq, p, 2)$, and $(p\mathbf{p}qpq, q, 2)$. This can be observed in Figure 1.*

Let us prove three remarks about this graph:

Remark **3.3** *Let h be a morphism on A^+. Then, iterations of the corresponding embedding map Emb_h can be represented as a forest of trees.*

Proof. Consider a node $(\widehat{w}, s, n)$ with $n > 0$. The knowledge of the father letter s implies that $\widehat{w}$ has a unique factorization determined by $h(s) \in h(A)$, i.e.

$$\text{if } h(s) = s_1 ... s_m,\ s_i \in A \text{ then } w = h^{n-1}(s_1) ... h^{n-1}(s_m). \tag{9}$$

Let $i \in \{1..m\}$ be such that $h^{n-1}(s_i)$ contains the origin of $\widehat{w}$. Hence,

$$w = h^{n-1}(s_1..s_{i-1}) h^{n-1}(s_i) h^{n-1}(s_{i+1}..s_m),$$

which corresponds to the embedding us_iu', with $u = s_1..s_{i-1}$ and $u' = s_{i+1}..s_m$. Denote by $\widehat{v}$ the pointed occurrence of $h^{n-1}(s_i)$ induced by its embedding in $\widehat{w}$. Therefore, the node $(\widehat{w}, s, n)$ has a unique father node given by $(\widehat{v}, s_i, n-1)$.

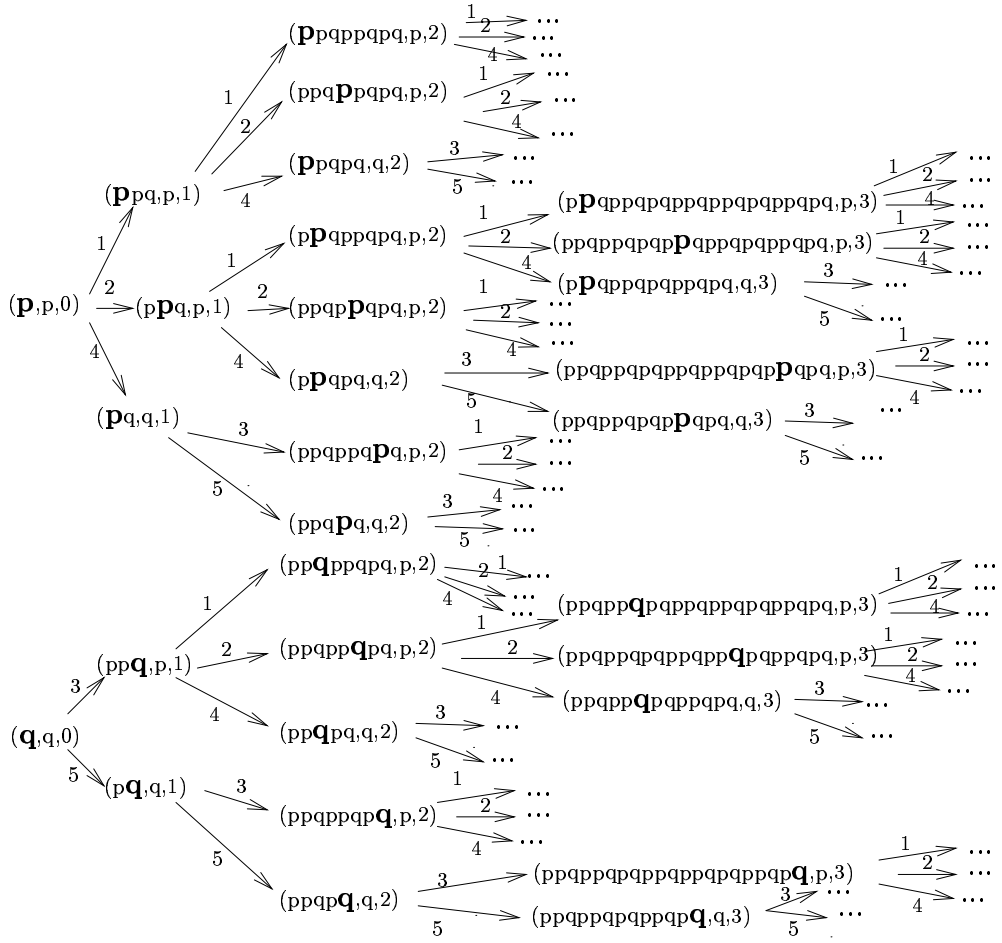

Figure 1: *The beginnings of the embedding forest of $h(p) = ppq$, $h(q) = pq$ (the origins are boldfaced). The labeling is deduced from the connector map given later on in the text.*

Applying recursively this process eventually leads to a node of type $(\mathbf{s}, s, 0)$, $s \in A$. The result follows. $\diamondsuit$

This forest is called the **embedding forest** of the morphism and the number of its trees is just the cardinality of A. In Figures 1 and 2 are illustrated the beginnings of the embedding forests of the morphisms $h(p) = ppq$, $h(q) = pq$ and $h(a) = b$, $h(b) = ab$.

In the embedding forest, the nodes of level n contain exactly the words of the pointed counterpart of the letter substitution sub-language of order n:

Remark **3.4** *Let h be a morphism on A^+. Then, in its corresponding embedding forest, we have for a fixed n:*

$$\widehat{L}_h^n = \{\widehat{w} \in \widehat{L}_h \mid (\widehat{w}, s, n) \in \textit{the embedding forest},\ s \in A\}.$$

Proof. Consider a word $\widehat{w}$ in $\widehat{L}_h^n$, such that $h^n(s) = w$, $s \in A$. From Remark 3.3, one can deduce that there exists an unique path of length n from the node $(\widehat{w}, s, n)$ to the

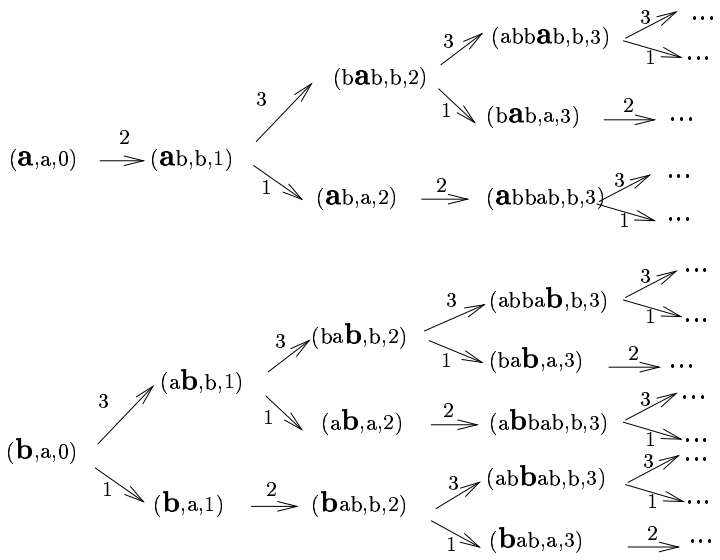

Figure 2: *The beginnings of the embedding forest of $h(a) = b$, $h(b) = ab$ (the origins are boldfaced). The labeling is deduced from the connector map given later on in the text.*

root node $(w_0, w_0, 0)$, where w_0 is the origin of $\widehat{w}$. $\diamondsuit$

Another property of the embedding forest is the **translation property** of its finite paths; a node $(., s, n)$ denotes the set of nodes of type $(\widehat{w}, s, n)$:

Remark **3.5** *Let h be a morphism on A^+. Consider a path in its embedding forest from a node of type $(., s, n)$ to a node of type $(., t, m)$. Then, there is a path from any node of type $(., s, n+k)$ to any node of type $(., t, m+k)$, for every $k > 0$.*

Proof. Let the node $(\widehat{w}, s, n)$ be bound to $(\widehat{v}, t, m)$, with $w = h^n(s)$, $v = h^m(t)$, $m > n$. Hence, according to the definition of the embedding map, there is a sequence of embeddings of $\widehat{w}$ which starts by an embedding of its father letter s into $h(A)$ by usu'. This gives another word $h^n(u)\widehat{w}h^n(u')$ of order $n+1$ with another father letter, say t_1. If $m > n+1$, then again, the letter t_1 is embedded into $h(A)$, say by $u_1 t_1 u_1'$, which gives $h^n(u_1)h^n(u)\widehat{w}h^n(u')h^n(u_1')$. This is carried on $(m-n-2)$ times until we get to $\widehat{v}$ with t as father letter. This sequence of embeddings is only determined by the sequence of the father letters. $\diamondsuit$

Note that father and order informations are necessary to ensure the forest-type structure:

Example **3.6** *For $h(a) = b$, $h(b) = ab$ (see Figure 2): the nodes $(\mathbf{a}b, a, 1)$ and $(\mathbf{a}b, b, 2)$ reflect the fact that there are two different ways of generating $\mathbf{a}b$; thus, the result through the embedding map is not the same. On the other hand, for $h(a) = ab$, $h(b) = c$, $h(c) = b$, the nodes $(\mathbf{b}, b, 2)$ and $(\mathbf{b}, b, 4)$ correspond to two instances of $\mathbf{b}$ which do not give the same result through the embedding map. The last case cannot occur for expansive morphisms.*

3.3 The Ends of the Embedding Forest

According to the definition of Emb_h, the words inside successive nodes in any path of an embedding tree are metrically closer and closer. Recalling that the space $({}^{\infty}\widehat{A}^{\infty}, d)$ is ultrametric and compact, each infinite path in the embedding forest can be looked at as a Cauchy sequence. It is then natural to define the **ends** of the embedding forest as containing the limit pointed words of their corresponding Cauchy sequences. The set $\widehat{\mathcal{E}}nds_h$ consists of all these words which are infinite. Hence, by definition of the boundary of a morphism h, the following holds:

$$\widehat{\mathcal{E}}nds_h \subseteq \partial\widehat{L}_h. \tag{10}$$

The aim of the next sections is to show in what respect the set of ends may be equal to the boundary, justifying the use of the embedding map. In particular, we shall focus on expansive morphisms where just a simple set must be added to $\widehat{\mathcal{E}}nds_h$ in order to get the whole boundary $\partial\widehat{L}_h$.

3.4 The Recovery of the Path by Successive Factorizations

This section solves the problem of recovering the path of the words in $\widehat{\mathcal{E}}nds_h$. The proof of Remark 3.3 uses a backward way which makes a strong use of the information included in each node (the father letter and the order). On the contrary, the words of $\widehat{\mathcal{E}}nds_h$ have not such information, since only the word part of the nodes is ensured to be convergent. Nevertheless, they contain their "embedding history", i.e. their successsive embedding steps around their origins.

First, let us see how such a direct recovery may work for finite words: Consider a word $\widehat{w}$ which belongs to $\widehat{L}_h$. The path in the embedding forest must starts at the node $(w_0, w_0, 0)$ where w_0 is the origin of $\widehat{w}$. Thus the first step of embedding is determined by finding how w_0 is embedded in some word of $h(A)$, i.e. $\widehat{L}_h^1$, which is included into $\widehat{w}$. For that purpose, $\widehat{w}$ is **factorized** into words of L_h^1:

$$\widehat{w} = x_{-m_1,1}...x_{0,1}...x_{n_1,1}, \quad \text{with} \quad x_{i,1} \in L_h^1,\ m_1, n_1 \in \mathbb{N}, \tag{11}$$

where the factor $x_{0,1}$ contains w_0, i.e. $x_{0,1} = uw_0v$, with $u, v \in A^+$ or empty. Since $x_{0,1} \in L_h^1$, there is a letter, say $s_{0,1}$, such that $h(s_{0,1}) = x_{0,1}$. This gives the first step of the corresponding path in the embedding forest as

$$(w_0, w_0, 0) \quad \longrightarrow \quad (x_{0,1}, s_{0,1}, 1).$$

For a second step, the word $\widehat{w}$ must be factorized into words of order 2, i.e. which belongs to L_h^2:

$$\widehat{w} = x_{-m_2,2}...x_{0,2}...x_{n_2,2}, \quad \text{with} \quad x_{i,2} \in L_h^2,\ m_2, n_2 \in \mathbb{N}, \tag{12}$$

where the factor $x_{0,2}$ contains the factor $x_{0,1}$. As well, $x_{0,2}$ has a father letter, say $s_{0,2}$, and this gives a second step of the path as:

$$(x_{0,1}, s_{0,1}, 1) \quad \longrightarrow \quad (x_{0,2}, s_{0,2}, 2).$$

By factorizing $\widehat{w}$ into words in L_h^k, it is possible to recover the next steps of the path for $0 < k < n$ by:

$$(x_{0,k}, s_{0,k}, k) \quad \longrightarrow \quad (x_{0,k+1}, s_{0,k+1}, k+1).$$

The last step is given when $k = n-1$ which gives a node of type $(\widehat{w}, s_{0,n}, n)$. A graphical representation of how the factors $x_{0,k}$ are successively embedded around the origin is shown in Figure 3.

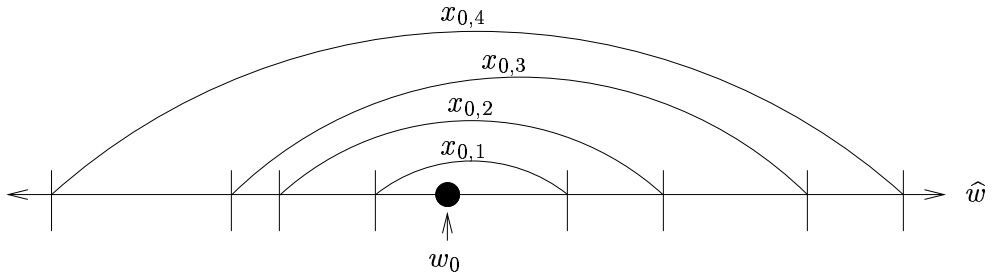

Figure 3: *A graphical representation of the successive embeddings around the origin.*

Example **3.7** *Consider again* $h(p) = ppq$, $h(q) = pq$ *and the word* $ppqppq\mathbf{p}qppqpq$ *of order* 3. *Its factorization into words of* $h(A)$ *is given by* $/ppq/ppq/\mathbf{p}q/ppq/pq/$ *where slashes symbolize separations between factors. Thus,* $x_{0,1} = \mathbf{p}q$ *and* $s_{0,1} = q$, *which gives the first step as* $(\mathbf{p}, p, 0) \longrightarrow (\mathbf{p}q, q, 1)$. *Then, the factorization into words of order two is given as* $/ppqppq\mathbf{p}q/ppqpq/$ *and* $x_{0,2} = ppqppq\mathbf{p}q$ *and* $s_{0,2} = p$ *which generates the next step as:* $(\mathbf{p}q, q, 1) \longrightarrow (ppqppq\mathbf{p}q, p, 2)$. *The last step is then given by:* $(ppqppq\mathbf{p}q, p, 2) \longrightarrow (ppqppq\mathbf{p}qppqpq, q, 3)$.

This way of recovering the path does not make use of the father and the order informations. As a consequence, it can be used on the infinite words belonging to $\widehat{\mathcal{E}}nds_h$: the successive factorizations are just given by:

$$\widehat{w} = ...x_{-n,k}...x_{0,k}...x_{n,k}..., \quad \text{with} \quad x_{i,k} \in L_h^k. \tag{13}$$

Of course, since the factorizations are not necessarily unique, they may lead to ambiguities. This will be handled in a following section. However, at this time, the obvious existence of at least one such sequence of factorizations for every word in $\widehat{\mathcal{E}}nds_h$ is sufficient for the next results.

3.5 The One-Way Boundary Words by the Embedding Map

As a preliminary result about the converse of $\widehat{\mathcal{E}}nds_h \subseteq \partial\widehat{L}_h$, this section will show that the one-way infinite words in the boundary $\partial\widehat{L}_h$ are contained in $\widehat{\mathcal{E}}nds_h$. For this purpose, we shall use the natural decomposition of the ends words by:

$$\widehat{\mathcal{E}}nds_h = R\widehat{\mathcal{E}}nds_h \cup L\widehat{\mathcal{E}}nds_h \cup Bi\widehat{\mathcal{E}}nds_h,$$

where respectively $R\widehat{\mathcal{E}}nds_h$, $L\widehat{\mathcal{E}}nds_h$, and $Bi\widehat{\mathcal{E}}nds_h$ denote the right, left and bi-infinite words of $\widehat{\mathcal{E}}nds_h$.

Now note that the full recursive Definition 2 of a Cauchy sequence is too general when one has to deal *a priori* with one-way infinite words. If a sequence (α_n) is known to converge to an one-way infinite word, it can be simplified to be of the following form (the left case is symmetrically handled): let (α'_n) be the sequence of identified factors,

for which $\alpha'_n \subseteq \alpha_n$, for all $n \in \mathbb{N}$, so that

$$\begin{array}{l} \alpha'_0 = \alpha_0 \in \widehat{A}, \\ \alpha'_n \subset \alpha'_{n+1} \text{ for all } n > 0, \text{ such that} \\ \alpha_{n+1} = \alpha'_n \beta_n \eta_n, \ \ \beta_n \in A^+, \ \eta_n \in A^+ \text{ or is empty}, \\ \alpha'_{n+1} = \alpha'_n \beta_n, \end{array} \tag{14}$$

A letter a is said **right-recurrent** (respect. **left recurrent**) if there exists some integer p such that $h^p(a) = av$ (respect. $h^p(a) = va$), $v \in A^+$. The **power** of a recurrent letter is the least integer giving the last equality.

The next lemma shows that the one-way infinite boundary words are essentially the same as the ones obtained by the fixed-point method of generating one-way infinite words (see for instance [30]):

Lemma **3.8** *Let h be an expansive morphism on A^+. Then, the word $\widehat{w} = aw'$ belongs to $R\partial\widehat{L}_h$, with $a \in A$ (respect. $\widehat{u} = u'a \in L\partial\widehat{L}_h$), iff the letter a is right-recurrent (resp. left-recurrent).*

Proof. Let us prove the result for right-infinite words. First, if the letter a is right-recurrent, then there exists a power p such that $h^p(a) = av$. This means that $h^{2p}(a) = h^p(h^p(a)) = h^p(a)h^p(v)$. More generally, $h^{np}(a) = h^{(n-1)p}(a)h^{(n-1)p}(v)$. Hence, consider a sequence (α_n) such that $\alpha_n = h^{np}(a)$, $n \geq 0$, pointed on the recurrent letter a, so that the identified factor sequence (α'_n) is just equal to (α_n). Since h is expansive, (α_n) converges to an infinite word. Finally, because of the shift invariance property of the boundary (see Remark 2.7), the limit word may be pointed on any letter of the infinite word.

Conversely, take a sequence (α_n) whose limit is in $R\partial\widehat{L}_h$. Since the alphabet A is finite, there is a fixed letter s such that, the sequence (α_n) may be replaced by one of its subsequence so that $\alpha_n = h^{k_n}(s)$, for all $n > 0$ and $k_n \in \mathbb{N}$. Because of the simplified form of Cauchy sequences in Equation (14), every word α_n can be written as au_n, for all $n > 1$, with $u_n \in \widehat{A}^+$, and $a \in A$. Hence, for all $n \in \mathbb{N}$,

$$\begin{array}{rcl} \alpha_{n+1} & = & \alpha_n \beta_n \eta_n, \quad \beta_n \in A^+, \ \eta_n \in A^+ \text{ or is empty}, \\ au_{n+1} & = & au_n \beta_n \eta_n, \end{array}$$

which means that,

$$\begin{array}{l} h^{k_{n+1}}(s) = h^{k_n}(s)\beta_n\eta_n, \\ h^{k_{n+1}}(s) = h^{k_{n+1}-k_n}(h^{k_n}(s))\beta_n\eta_n. \end{array}$$

Therefore, $h^{k_{n+1}-k_n}(a) = av$, $v \in A^+$ and $k_{n+1} - k_n$ is just a multiple of the recurrence power of the letter a, say p. It is then possible to redefine the sequence (α_n) without impairing the limit word as:

$$\alpha_n = h^{(n+l)p}(a), \quad n > 1$$

where l is an integer sufficiently large so that the origin of the original limit word is already included in $h^l(a) = \alpha_1$. The result follows. $\diamondsuit$

A consequence of the last lemma is that the cardinality of the unpointed right (respect. left) boundary words is equal to the number of right (respect. left) recurrent letters in A, i.e.

$$|R\partial L_h| = |\{a \text{ is right-recurrent}, a \in A\}|$$

and

$$|L\partial L_h| = |\{a \text{ is left-recurrent}, a \in A\}|.$$

Let us prove now that the one-way boundary words are contained in the set of ends $\widehat{\mathcal{E}}nds_h$ of the embedding forest:

Proposition **3.9** *Let h be an expansive morphism on A^+. Then,*

$$R\widehat{\mathcal{E}}nds_h = R\partial\widehat{L}_h \quad \textit{and} \quad L\widehat{\mathcal{E}}nds_h = L\partial\widehat{L}_h.$$

Proof. Let us show the result for right boundary sets. By definition of the ends of the embedding forest, we have that $R\widehat{\mathcal{E}}nds_h \subseteq R\partial\widehat{L}_h$. Conversely, consider a word $\widehat{w}$ in $R\partial\widehat{L}_h$. According to Lemma 3.8, the word $\widehat{w}$ must start by a right-recurrent letter, say a, with power p, and can be generated by a sequence defined by $\alpha_0 = w_0$ and $\alpha_1 = h^l(a)$ where l is a power such that $w_0 \subset h^l(a)$ and $\alpha_n = h^{(n+l)p}(a)$, for $n > 1$. Now, we know that

$$\alpha_{n+1} = \alpha_n u_n, \quad u_n \in A^+. \tag{15}$$

Hence, $\widehat{w}$ can factorized into words in $L_h^{(n+l)p}$, where the first factor is α_n. Using the factorization method allows us to conclude that there is a path from the origin w_0 to α_n for all $n > 0$. Moreover, because of the translation property (see Remark 3.5) there is a path from α_n to α_{n+1} for all n. $\diamondsuit$

3.6 The Bi-Infinite Boundary Words and the Pasted Set

Up to now, we learned that $\widehat{\mathcal{E}}nds_h \subseteq \partial\widehat{L}_h$, and the last section proves that $R\partial\widehat{L}_h \subseteq R\widehat{\mathcal{E}}nds_h$ and $L\partial\widehat{L}_h \subseteq L\widehat{\mathcal{E}}nds_h$. We shall show now that we have to add a simple set of bi-infinite words to $Bi\widehat{\mathcal{E}}nds_h$, so that we obtain the equality with $Bi\partial\widehat{L}_h$.

The **pair set** $Pairs(L_h)$ contains all the factors of length two in words of L_h. The **pasted set** of L_h is defined by pasting unpointed one-way limit words:

$$\begin{aligned} PasL_h = \{w \in {}^{\infty}A^{\infty} \mid w = vu, \text{ where } & v = v'x \in L\partial L_h, \\ & u = yu' \in R\partial L_h,\ xy \in Pairs(L_h)\}. \end{aligned} \tag{16}$$

The words $xy \in Pairs(L_h)$ of the last definition, i.e. the words such that x is left-recurrent and y is right-recurrent, are called the **pasting pairs**.

Remark **3.10** *The pasted set $PasL_h$ is a finite language.*

Proof. According to Lemma 3.8, $R\partial L_h$ and $L\partial L_h$ are finite sets. $Pairs(L_h)$ as well. $\diamondsuit$

Note that the pointed counterpart $Pas\widehat{L}_h$ is obtained by setting origins on a letters which are at a finite shift power from the pasting pair.

Lemma **3.11** *Let h be an expansive morphism on A^+. Then, $Pas\widehat{L}_h \subset \partial\widehat{L}_h$.*

Proof. Let $\widehat{w}$ belongs to $Pas\widehat{L}_h$. This means that $\widehat{w}$ has been generated by pasting a right-infinite boundary word, say yu', with $y \in A$, and a left-infinite one, say $v'x$, with $x \in A$, so that the pasting pair of $\widehat{w}$ is xy. The claim is that $\widehat{w} = v'xyu' \in \partial\widehat{L}_h$. From Lemma 3.8, we know that the letter x must be left-recurrent for some power p_1 and that y must be right-recurrent for some other power p_2. Define p as the least common multiple of p_1 and p_2, thus for all $n > 0$, the following holds:

$$h^{np}(xy) \subset h^{(n+1)p}(xy). \tag{17}$$

Thus consider the following Cauchy sequence (α_n): let $\alpha_0 = w_0$, let α_1 be a word of $\widehat{L}_h$ which contains w_0 and which is in the class of $h^{lp}(xy)$ (l is a power such that $w_0 \subset h^{lp}(xy)$). Finally, define $\alpha_{n+1} = h^{(n+l)p}(xy)$ pointed on the same w_0. Since h is expansive, (α_n) converges to $\widehat{w}$. $\diamondsuit$

Now, we can check that $Pas\widehat{L}_h$ is not necessarily included in $\widehat{\mathcal{E}}nds_h$:

Example **3.12** *Consider the following morphism on the alphabet $A = \{a, b, c\}$ where $h(a) = aa$, $h(b) = ab$, $h(c) = abac$, and consider the Cauchy sequence (α_n) where α_n is equal to $h^n(c)$ pointed on the letter a to the right of the first b from the left, i.e.*

$$\begin{array}{c} \vdots \\ aaaaaaab\mathbf{a}aaaaaabaaabac \\ aaab\mathbf{a}aabac \\ \mathbf{a} \end{array}$$

This sequence converges to ${}^{\omega}ab\mathbf{a}a^{\omega}$. By applying the factorization method, the factors $x_{0,k}$, $k > 0$ which contain the origin are stopped by the central letter b and converge to $\mathbf{a}a^{\omega}$. On the other hand, if the origin would have been set to the other side of this "resistant" b, then the factors $x_{0,k}$, $k > 0$, would converge to ${}^{\omega}ab$. Note that the letter a is right-recurrent and b is left-recurrent. In fact, since for every $n > 1$ and $m > 0$, the words ${}^{m}aba^{n}$ do not belong to L_h, the word ${}^{\omega}ab\mathbf{a}a^{\omega}$ strictly belongs to $Pas\widehat{L}_h$.

Theorem **3.13** *Let h be an expansive morphism on A^+. Then,*

$$\partial\widehat{L}_h = \widehat{\mathcal{E}}nds_h \cup Pas\widehat{L}_h.$$

Proof. Because of Proposition 3.9, it is sufficient to consider only the bi-infinite asymptotic sets. We know that $Bi\widehat{\mathcal{E}}nds_h \subseteq Bi\partial\widehat{L}_h$ holds. Then, according to Lemma 3.11, the remaining thing to prove is that

$$Bi\partial\widehat{L}_h \subseteq Bi\widehat{\mathcal{E}}nds_h \cup Pas\widehat{L}_h.$$

So, assume that $\widehat{w}$ is a word in $Bi\partial\widehat{L}_h$. Apply to $\widehat{w}$ the method of recovering the path by successive factorizations, that is using Equation (13) for all $k > 0$:

$$\widehat{w} = ...x_{-n,k}...x_{0,k}...x_{n,k}..., \quad \text{with} \quad x_{i,1} \in L_h^k. \tag{18}$$

The result is an infinite Cauchy sequence (μ_n) where $\mu_n = x_{0,n}$, $n \geq 0$. Since h is expansive, $lim_{n\to\infty}|\mu_n| = \infty$. Now there are two cases:

If the words of (μ_n) have their lengths increasing to infinity to both directions, then $\lim_{n\to\infty}\mu_n = \widehat{w}$. Hence, according to the definition of the embedding map, $\widehat{w}$ belongs to $Bi\widehat{\mathcal{E}}nds_h$.

Otherwise, (μ_n) has its words going to infinity only to one side, say the right side. In this case, there exists an index $n_0 \geq 0$, such that the words μ_n are identified on a left factor:

$$\mu_{n+1} = \mu_n\beta_n, \quad \beta_n \in A^+, \quad \text{for } n > n_0.$$

This means that $\lim_{n\to\infty}\mu_n$ belongs to $R\partial\widehat{L}_h$. Denote this resulting limit word by $\widehat{v}$. The situation is depicted on the next figure where only one side of $\widehat{w}$ was covered by (μ_n). Now put temporarily the origin on some letter of x_{-1,m_0}, where m_0 is sufficiently large

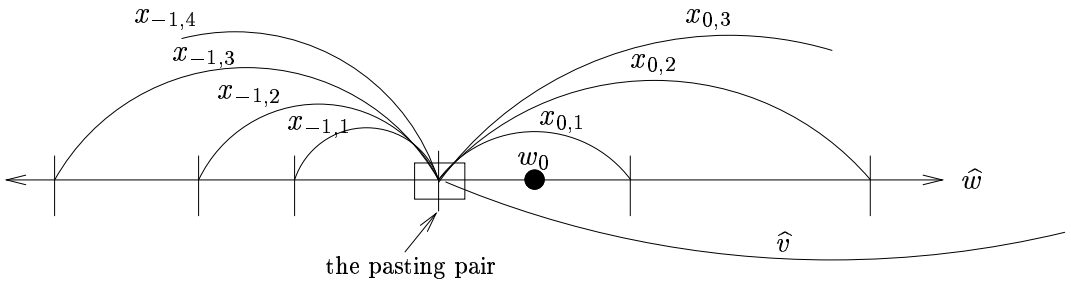

Figure 4: *Graphical representation of the factors embeddings for a pasted word.*

so that this factor is in the other side of $\widehat{v}$. Apply the factorization method of recovery: the result is an infinite Cauchy sequence (ν_n) where $\nu_n = x_{-1,n+m_0}$, $n \geq 0$. Again since h is expansive, $lim_{n\to\infty}|\nu_n| = \infty$, and there exists an index $n_0 \geq 0$ such that:

$$\nu_{n+1} = \gamma_n \nu_n, \quad \gamma_n \in A^+, \quad \text{for } n > n_0 + m_0.$$

This means that $\lim_{n\to\infty} \nu_n$ belongs to $L\partial\widehat{L}_h$. Remove the origin from the limit word of (ν_n), index it according to $\widehat{v}$, and paste it to $\widehat{v}$: this is a word which belongs to $Pas\widehat{L}_h$. $\diamondsuit$

One may generalize the last result for unpointed boundaries:

Corollary **3.14** *Let h be an expansive morphism on A^+. Then,*

$$\partial L_h = (\widehat{\mathcal{E}}nds_h/\sim_\sigma) \cup PasL_h.$$

Proof. Using the factorization method, there is a path in some tree of the embedding forest from the root to every $x_{0,n}$ in Equation 18. This means that if the origin is shifted within these factors, there is still a path in the embedding forest (see Remark 3.4). Hence, the shift invariance property holds also for $\widehat{\mathcal{E}}nds_h$, that is

$$\text{if } \widehat{w} \in \widehat{\mathcal{E}}nds_h \text{ then } \sigma^m(\widehat{w}) \in \widehat{\mathcal{E}}nds_h, \quad \forall m \in \mathbb{Z}.$$

The set $\widehat{\mathcal{E}}nds_h$ can be independently quotiented, and by construction, the pointed pasted set as well. $\diamondsuit$

Note that there are cases where $\widehat{\mathcal{E}}nds_h \cap Pas\widehat{L}_h = \emptyset$ does not hold:

Example **3.15** *The morphism $h(a) = aa$ has the word ${}^\omega a^\omega$ in its boundary, and it belongs to $\widehat{\mathcal{E}}nds_h \cap Pas\widehat{L}_h$.*

Before going for the direct use of the construction of $\partial\widehat{L}_h$, let us investigate some relationships with other frameworks:

3.7 The Boundary and the Closure Operator

Sets of asymptotic bi-infinite words were generated from the beginning of symbolic dynamic theory [22, 23, 24, 15, 14, 4]. One of the ways was by first obtaining a word in

$R\partial\widehat{L}_h \cup L\partial\widehat{L}_h$ (using the fact that iterating a morphism from, say a right-recurrent letter, then prefixes are stable), and second, by applying the shift σ to move the origin. In this section, $Bi\widehat{L}$ denotes the set of the bi-infinite words included in a language $\widehat{L}$. According to the definition of the boundary, we have that if $\widehat{w} \in R\partial\widehat{L}_h \cup L\partial\widehat{L}_h$, then

$$Bi(Closure(\{\sigma^n(\widehat{w}), n \in \mathbb{Z}\})) \subseteq Bi\partial\widehat{L}_h.$$

More generally,

$$Bi(Closure(\{\sigma^n(\widehat{w}),\ \ n \in \mathbb{Z},\ \ \widehat{w} \in R\partial\widehat{L}_h \cup L\partial\widehat{L}_h\})) \subseteq Bi\partial\widehat{L}_h.$$

However this can be a strict inclusion:

Example **3.16** *Consider the morphism $h(a) = ab$, $h(b) = ba$, and $h(c) = aca$: the letter c is not recurrent and so one cannot obtain the whole boundary from $R\partial\widehat{L}_h \cup L\partial\widehat{L}_h$ only.*

One of the important cases where the two sets of bi-infinite words coincide is when h is a primitive morphism (i.e. there exists a finite power n such that all letters of A are included in $h^n(s)$, for all $s \in A$). The interest in $Closure(\{\sigma^n(\widehat{w}), n \in \mathbb{Z}\})$ for the primitive case is that the dynamical system on this set induced by the action of the shift σ has many properties, as for instance *strict ergodicity.*

Remark **3.17** *Let h be a primitive morphism on A^+. Then, for any $\widehat{w} \in R\partial\widehat{L}_h \cup L\partial\widehat{L}_h$,*

$$Bi(Closure(\{\sigma^n(\widehat{w}), n \in \mathbb{Z}\})) = Bi\partial\widehat{L}_h.$$

Proof. If h is primitive, then clearly the set of factors of $\widehat{w}$ is equal to the set of factors of the whole language $\widehat{L}_h$. This is sufficient to generate the bi-infinite boundary words. ♢

Bi-infinite words were also directly generated [15, 14, 4] and this was managed through the pasting pair idea. Indeed, taking a pair of two letters where one is left-recurrent and the other right-recurrent, successive applications of the morphism let stable the suffixes of the former and the prefixes of the last (see the proof of Lemma 3.11). For identical reasons as for the one-way case, the following holds:

$$(Closure(\{\sigma^n(\widehat{w}),\ \ n \in \mathbb{Z},\ \ \widehat{w} \in Pas\widehat{L}_h\}) \subseteq Bi\partial\widehat{L}_h.$$

3.8 The Boundary of D0L-systems

Let us also prove that Theorem 3.13 can be translated into the $D0L$ language framework. Let the subalphabet A_t be the set of letters which appear during the iterations of h on the word $t \in A^+$, that is $A_t = \{s \in A \mid \exists\, n \geq 0,\ s \subset h^n(t)\}$. Let also the morphism h' be the restriction of h to A_t: hence, $D0\widehat{L}_{h'}(t) = D0\widehat{L}_h(t)$.

Now the only important difference from the letter substitution language case lies in the fact that, t being any word in A^+, the set $Pairs(\widehat{L}_{h'})$ is not necessarily equal to $Pairs(\widehat{L}_h)$. Hence, $Pas\widehat{L}_{h'}$ can be strictly included in $PasD0\widehat{L}_{h'}(t)$. However, by the same arguments as in the proof of Theorem 3.13, one can conclude that,

Corollary **3.18** *Let h be an expansive morphism on A^+. Then, for all $t \in A^+$,*

$$\partial D0\widehat{L}_{h'}(t) = \widehat{\mathcal{E}}nds_{h'} \cup PasD0\widehat{L}_{h'}(t).$$

4 The Regular Coding

In this section, we shall use the effective construction of the boundary to bijectively send it onto a regular language.

4.1 The Labelling of the Embedding Forest

Because of the translation property of the embedding forest (see Remark 3.5), the trees must present a systematic branching. Indeed, the translation property was proved by using the fact that, each time Emb_h is applied to some word with s as father letter, the obtained words correspond to all the ways of embedding s into the image of the letters by h, i.e. $h(A)$. So, a distinct letter of some new finite alphabet Γ, called the **connector alphabet**, can be assigned to each kind of embedding. This is implemented by the **connector map** [7], a bijection denoted by f_h, which maps the pointed counterpart of $h(A)$ to Γ:

$$\begin{array}{rcl} f_h : \widehat{h(A)} & \to & \Gamma \\ (u\mathbf{s}u') & \mapsto & g \end{array} \tag{19}$$

where the word $u\mathbf{s}u'$ corresponds to an embedding of $\mathbf{s}$. The connector map induces a labelling of all the edges of the embedding forest: for a node $(\widehat{w}, s, n)$, the arc going to $(\widehat{v}, t, n+1)$ is labeled by $f_h(u\mathbf{s}u')$ if $u\mathbf{s}u'$ is the embedding of the letter s in $h(A)$ such that $h^n(usu') = v$.

Example **4.1** *Consider again the morphism $h(p) = ppq$, $h(q) = pq$, with the connector map can be defined as:*

$$f_h(\mathbf{p}pq) = 1 \quad f_h(p\mathbf{p}q) = 2 \quad f_h(pp\mathbf{q}) = 3 \quad f_h(\mathbf{p}q) = 4 \quad f_h(p\mathbf{q}) = 5. \tag{20}$$

Revisiting Example 3.7, then for example, $(\mathbf{p}, p, 0)$ is bound to $(\mathbf{p}q, q, 1)$ by an edge that is labeled $f_h(\mathbf{p}q) = 4$. Next the node $(\mathbf{p}q, q, 1)$ is bound to $(ppqppq\mathbf{p}q, p, 2)$ by an arc whose label is $f_h(pp\mathbf{q}) = 3$ since this reflects how $\mathbf{p}q$ is embedded into $ppqppq\mathbf{p}q$. Hence, the label path of $(ppqppq\mathbf{p}q, p, 2)$ can be considered to be the word "43" (see Figure 1).

The other example in Figure 2 where $h(a) = b$, $h(b) = ba$ has the following connector map:

$$f_h(\mathbf{b}) = 1 \quad f_h(\mathbf{a}b) = 2 \quad f_h(a\mathbf{b}) = 3 \tag{21}$$

Remark **4.2** *Let h be a morphism on A^+, f_h be its connector map with Γ the connector alphabet. Then, the set of nodes (except the roots) of the embedding forest is injectively mapped onto the semi-group Γ^+.*

Proof. First, consider a node $(\widehat{w}, s, n)$, with $n > 0$ and let us use the same argument as in the proof of Remark 3.3 in order to recover its path label. Its father letter s implies that $\widehat{w}$ has an unique decomposition

$$h(s) = s_1 ... s_m, \quad s_i \in A, \quad \text{then } w = h^{n-1}(s_1) ... h^{n-1}(s_m), \tag{22}$$

with an index $i \in \{1..m\}$ such that $h^{n-1}(s_i)$ contains the origin of $\widehat{w}$, and this corresponds to the embedding us_iu'. Hence, the nth label of the path is recovered, that is $g_n = {f_h}^{-1}(u\mathbf{s_i}u')$. The $(n-1)$th label is obtained by applying this process on $h^{n-1}(s_i)$, and recursively, the whole path label $g_1 ... g_n$ can be recovered.

Now, consider two nodes $(\widehat{w}, s, n)$ and $(\widehat{u}, t, m)$. Since the length of the path label is equal to the order, if $m \neq n$, their path labels cannot be equal. Now, since f_h is bijective,

if $s \neq t$, then the recovery of the nth labels must be different. Hence, $w = v$, and if $\widehat{w} \neq \widehat{v}$, they may differ only by their pointings. But, by Equation 22, the factors which contain their origins are reduced until being the origins themselves. Thus, $\widehat{v} = \widehat{w}$. $\diamondsuit$

4.2 The Embedding Forest is Regular

Let $Cod\widehat{L}_h \subset \Gamma^+$ denote the set of all finite path labels of the embedding forest. Recall that a **finite type language of order** n is a language which can be specified by its set of factors of length n: Let X be a subset of $\{w \in \Gamma^+ \mid |w| = n\}$, then the corresponding finite type language is given by $\{v \in \Gamma^+ \mid$ all factors of v are in $X\}$. Such a language is known to be regular.

Remark **4.3** *Let h be a morphism on A^+ and let Γ be the connector alphabet. Then, $Cod\widehat{L}_h$ is a finite type language of order 2 included in Γ^+.*

Proof. First, remind that the edges starting from a specific node $(\widehat{w}, s, n)$ depend only on the father letter s, and this independently to $\widehat{w}$ and n. Each of these differently labeled edges leads to another node, and the following edges correspond to the embeddings of its father letter. Hence, a pair ij may occur in a path label if the embedding of type i leads to a word with a father letter included in the embedding of type j. Moreover, after the embedding i, the embedding j is always possible. Thus $Cod\widehat{L}_h$ is a language which is specified by these pairs. $\diamondsuit$

Thus the paths of the embedding forest can be "folded" into a Büchi automaton which recognizes $Cod\widehat{L}_h$: the set of states is the alphabet A, the sets of initial and final states are all the states, the set of labels is the connector alphabet Γ, the set of edges consists of the edges (s, g, t) such that s is included in $h(t)$, i.e. $h(t) = u\mathbf{s}u'$, and labeled by $f_h(u\mathbf{s}u') = g$. Note that if the set of initial states is restricted to a single state, say s, then the recognized language corresponds to the nodes of the single embedding tree with root $(\mathbf{s}, s, 0)$.

Let us extend this to the set $\widehat{\mathcal{E}}nds_h$. Denote by $Cod\widehat{\mathcal{E}}nds_h$ the set of the corresponding infinite path labels. This is a set included in the right-infinite words over Γ, denoted by Γ^ω:

Lemma **4.4** *Let h be a morphism on A^+ and let Γ be the connector alphabet. Then, $Cod\widehat{\mathcal{E}}nds_h$ is a Muller-recognizable language included in Γ^ω.*

Proof. Let F be the set of all states which corresponds to an embedding for which there is an increasing of the length, i.e. $|h(s)| > 1$, $s \in A$. Consider the automaton which recognizes $Cod\widehat{L}_h$, and constrain it by a table set defined as the power set of F. This gives a Muller automaton which recognizes every path leading to an infinite word, i.e. to a word in $\widehat{\mathcal{E}}nds_h$. $\diamondsuit$

The corresponding Muller automata of the morphisms of Figures 1 and 2 are shown in Figure 5.

Corollary **4.5** *Let h be an expansive morphism on A^+ and let Γ be the connector alphabet. Then, $Cod\widehat{\mathcal{E}}nds_h$ is a finite type language of order 2 included in Γ^ω.*

Proof. If h is expansive, then the table set of the Muller automaton can be defined as the power set of all states. $\diamondsuit$

The aim now is to investigate the cases where the regular language $Cod\widehat{\mathcal{E}}nds_h$ is injectively mapped to $\widehat{\mathcal{E}}nds_h$.

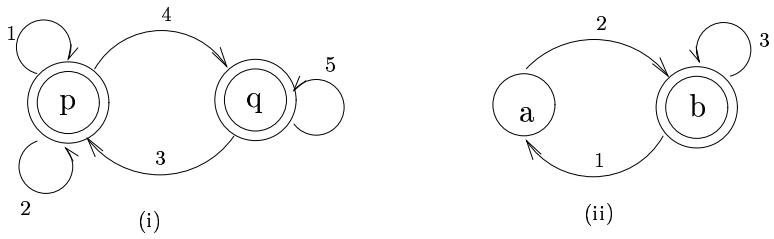

Figure 5: *The respective automata recognizing the ends of the morphisms $h(p) = ppq$, $h(q) = pq$, and $h(a) = b$, $h(b) = ab$. The final states are indicated by a double circle.*

4.3 The Recovery of the Path Label by Successive Factorizations

In this section, we give a final version of the recovering method (after section 3.4) which allows us to obtain the path label of a word in $\widehat{\mathcal{E}}nds_h$.

Recall that the recovery of the path relies on successive factorizations for each $k \geq 0$:

$$\widehat{w} = ...x_{-n,k}...x_{0,k}...x_{n,k}..., \quad \text{with} \quad x_{i,k} \in L_h^k, \tag{23}$$

where the factors $x_{i,k}$ are assumed to include the factors $x_{i,k-1}$, and where $x_{i,0} = w_i$, for all $i \in \mathbb{Z}$. Now each factor $x_{i,k}$ has at least one father letter. Concatenating all these letters, one obtains a word, called **the ancestor**, which belongs to the k-fold **inversion** of h, i.e. h^{-k}: for each $k \geq 0$,

$$...s_{-n,k}...s_{0,k}...s_{n,k}... \in h^{-k}(\widehat{w}) \quad \text{with} \quad s_{i,k} \in A, \text{ and } h^k(s_{i,k}) = x_{i,k}, \tag{24}$$

where $s_{i,0} = w_i$, for all $i \in \mathbb{Z}$. Since $\widehat{w}$ belongs to $\widehat{\mathcal{E}}nds_h$, the image of the inversion is not empty, but can be multiple.

Recovering one letter of the label path is equivalent to knowing how exactly $x_{0,k-1}$ is embedded into $x_{0,k}$ (this is the point we skipped in section 3.4). This can be translated into knowing how $h^{-(k-1)}(x_{0,k-1})$ (which is equal to a letter $s_{0,k-1}$) is embedded into $h^{-(k-1)}(x_{0,k})$ (which is equal to some $us_{0,k-1}u'$ in $h(A)$): the kth letter of the label path is ${f_h}^{-1}(us_{0,k-1}u')$. Summarizing, the recovery algorithm of the label path is:

To recover the kth letter of the path label, factorize $h^{-(k-1)}(\widehat{w})$ into $h(A)$, and find how its origin $s_{0,k-1}$ is embedded into the factor which contains it.

This procedure is equivalent to successively applying $h^{-1}(\widehat{w})$, factorizing it into $h(A)$ and finding how the origin is embedded into the factor of $h(A)$ which contains it.

Example **4.6** *With $h(p) = ppq$, $h(q) = pq$, the instance of Example 21 can be applied to a word $\widehat{w} = ...pqppqppq\mathbf{p}qppqpq...$ in $\widehat{\mathcal{E}}nds_h$. Indeed, its factorization into $h(A)$ is given by $...pq/ppq/ppq/\mathbf{p}q/ppq/pq/...$; so $s_{0,0} = p$, and the first label is given by applying the connector map to $\mathbf{p}q$, i.e. $f_h(\mathbf{p}q)$. Then, $h^{-1}(...pqppqppq\mathbf{p}qppqpq...) = ...qpp\mathbf{q}pq...$, so $s_{0,1} = q$, and its factorization into $h(A)$ is $...q/pp\mathbf{q}/pq/...$; the second label is given by $f_h(pp\mathbf{q})$.*

4.4 The Circularity Property

Injecting $Cod\widehat{\mathcal{E}}nds_h$ into $\widehat{\mathcal{E}}nds_h$ clearly depends on the uniqueness of the inversion, i.e. h^{-1} must be a map.

Lemma **4.7** *When used for the recovery of a path label of a word in $\widehat{\mathcal{E}}nds_h$, the inversion h^{-1} can be restricted to $\widehat{\mathcal{E}}nds_h$.*

Proof. Assume that for some $\widehat{w} \in \widehat{\mathcal{E}}nds_h$, $h^{-1}(\widehat{w})$ has a word $\widehat{v}$ in $({}^{\infty}\widehat{A}^{\infty} \setminus \widehat{\mathcal{E}}nds_h)$. This would mean that $h^{-1}(\widehat{w})$ contains at least one finite factor, say u, which is not in L_h. Now the label path of $\widehat{w}$, except its first letter, may be obtained by applying the recovery method to $\widehat{v}$, i.e.:

$$h^{-1}(\widehat{w}) = \widehat{v} = ...x'_{-n,k}...x'_{0,k}...x'_{n,k}..., \quad \text{with} \quad x'_{i,k} \in L_h^k, \;\; k > 0. \tag{25}$$

Since $\widehat{w} \in \widehat{\mathcal{E}}nds_h$, the sequence $(x'_{0,n})$ must ultimately cover all $\widehat{v}$. Hence, for some $k > 0$, $x'_{0,k}$ covers u. But, this means that we have reached a deadend since the connector map cannot be applied on $x'_{0,k}$. $\diamondsuit$

Thus the uniqueness of the factorization involved in the recovery method is not exactly related to *code theory* [1]. Indeed, the condition that $\widehat{\mathcal{E}}nds_h$ must be stable through h^{-1} is related to the morphism itself. In particular, if h is not injective on the letters of A, then $h(A)$ becomes a multiset, without necessarily impairing the uniqueness of the recovery method.

That is why one must look for a more precise notion. Let us first slightly extend the factorization operation: let w be any finite factor in L_h, then this word admits a factorization $w = ux_1x_2...x_nv$ where $x_i \in h(A)$, and u and v are respectively a right and a left factor of two words x_0, x_{n+1} in $h(A)$. Its ancestor is then given by $h^{-1}(w) = s_0...s_{n+1}$, where $h(s_i) = x_i$, for $i \in \{0..(n+1)\}$. Now we are ready to introduce the following definition due to Mignosi and Séébold [21]:

Definition **6** *A morphism h on A^+ is* **circular** *with* **synchronization delay** *K if for all finite factors w in L_h, if w admits two factorizations, say*

$$w = ux_1x_2...x_nv = u'x'_1x'_2...x'_mv',$$

with $s_0...s_{n+1}$ and $s'_0...s'_{m+1}$ as ancestors which are factors of L_h, then whenever

$$|ux_1x_2...x_{i-1}| > K \quad \text{and} \quad |x_{i+1}...x_nv| > K \tag{26}$$

there exists an index k such that

$$ux_1x_2...x_i = u'x'_1x'_2...x'_k \quad \text{with } s_i = s'_k. \tag{27}$$

Circularity with synchronization delay means that if the word whereto apply the inverse morphism is sufficiently long, then the factorization is uniquely determined excluding the first and last K letters.

Example **4.8** *In the figure 6, we can graphically observe how the circularity works: $|ux_1x_2| > K$ and $|x_6v| > K$, and then, for $i = 3, 4, 5$, $ux_1x_2..x_i = u'x'_1..x'_{i-1}$ such that $s_i = s'_{i-1}$, and $x_i = x'_{i-1}$. In particular, the factor which contains the origin is uniquely determined.*

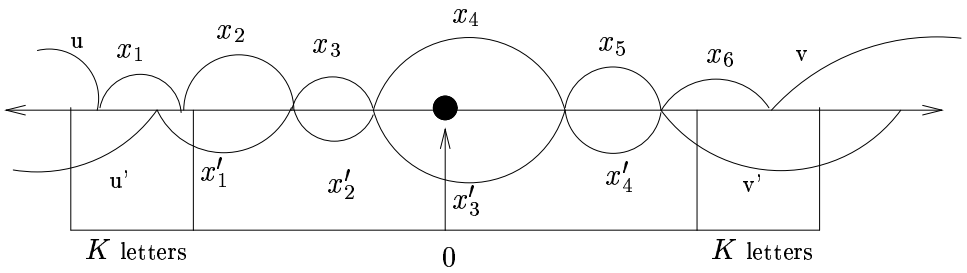

Figure 6: *Graphical representation of an example of two synchronized factorizations of some word whenever circularity holds.*

This property has been derived from the dynamically-oriented literature where it appears in a slightly weaker version called **recognizability** [20, 27, 25]. Circularity is different from recognizability since the last only ensures separation indexes between factors without explicitely determininig them. In particular, circularity do not require injectivity of the morphism for A. This comes from the fact that the ancestor is assumed to be a factor in L_h, an hypothesis in accordance with Lemma 4.7.

Circularity is the notion we shall use to prove properties about unique recovery of the path labels of the boundary words of $\widehat{\mathcal{E}}nds_h$. The next theorem syntesizes all the links we shall need in the sequel. It is mainly a direct consequence of results due to Mignosi and Séébold [21] and to Ehrenfeucht and Rozenberg [11]. These were proved in the $D0L$ languages framework, but recall that letter substitution languages are just finite unions of $D0L$ languages.

Remind that a morphism is said **n-power free** if every word in L_h does not contain any factor u^n with $u \in A^+$, $n > 1$. Recall also that a word $\widehat{w}$ is **periodic** if there exists some $p > 0$ such that $w_i = w_{i+p}$ for all $i \in \mathbb{Z}$ where w_i, w_{i+p} belong to A. If π is a cyclic permutation of A, then a morphism is said **(A, π)-cyclic** if

1. for each $s \in A$, $h(s) = s_1...s_m$, then $s_{i+1} = \pi(s_i)$, with $1 \leq i \leq m-1$,
2. for each pair $s, t \in A$ such that $\pi(s) = t$, then $\pi(last(h(s))) = first(h(t))$, where $first(w)$ extracts the first letter of w, and $last(w)$ its last one.

The important point of this definition is that the word $v = s\pi(s)\pi^2(s)...\pi^{|A|-1}(s)$, with $s \in A$, is such that v^n is a factor of $h^n(v)$, for all $n > 0$ [11](see Lemma 6.11). For instance, $h(a) = ab$, $h(b) = c$, $h(c) = abc$, is cyclic with $\pi(a) = b$, $\pi(b) = c$, $\pi(c) = a$: here, $v = abc$.

Theorem **4.9** *Let h be an expansive morphism on A^+. Then, the following conditions are equivalent:*

1. *h is circular.*
2. *h is n-power free, for some $n > 1$.*
3. *There is no periodic word in $\partial\widehat{L}_h$.*

4. h^{-1} *is a map on* $\partial\widehat{L}_h$.

5. *The recovery of the path label is unique for each word in* $\widehat{\mathcal{E}}nds_h$.

Proof. The fact that (1) <=> (2) has been proved by Mignosi and Séébold in [21](see Corollary 1).

For (3) => (2), let h be not n power-free for some n. Hence, for all $n > 1$, there is a word $u \in A^+$ such that u^n is a factor of L_h. This implies [11](Theorem 2) that there exists a factor v of L_h, such that for any $n > 1$, v^n is also a factor of L_h. Thus ${}^\omega v^\omega$ belongs to $\partial\widehat{L}_h$.

Conversely, (2) => (3) holds since, if $\widehat{w}$ is a periodic word in $\widehat{\mathcal{E}}nds_h$, then its Cauchy sequence must have an identified factor sequence which is a subsequence of $\{v^n\}_{n\in\mathbb{N}}$, for some v which is factor in L_h.

For (1) => (4), one must just consider that if h^{-1} is not a map on $\partial\widehat{L}_h$, then there is at least two different words in its image: there is a place where the factorization into $h(A)$ words is not unique.

Let us then prove that (4) => (3). Again, this is derived from a result in [11]. Indeed, assume that $\widehat{w}$ in $\widehat{\mathcal{E}}nds_h$ is periodic, for instance ${}^\omega v^\omega$. This means that there exists arbitrarily large powers of v as factor in L_h. Let the subalphabet A_v be the set of letters which appear in v. We can conclude [11] (see Lemma 6.12) that there is a restriction of h on s_v which is (A_v, π)-**cyclic**. Thus there is a periodic word ${}^\omega u^\omega$ in $\widehat{\mathcal{E}}nds_h$ (may be different from $\widehat{w}$), where $u = s\pi(s)\pi^2(s)...\pi^{|A_v|-1}(s)$, with $s \in A_v$. Since h is expansive, there is a power k such that u is included in $h^k(s)$, for all $s \in A_v$. This implies that the origin of ${}^\omega u^\omega$ can be enclosed in any $h^k(s)$. Hence, ${}^\omega u^\omega$ cannot factorized uniquely in L_h^k and so h^{-1} cannot be a map on $\widehat{\mathcal{E}}nds_h$.

The claim (4) <=> (5) restricted on $\widehat{\mathcal{E}}nds_h$is clear from the recovery method and from the fact that the connector map f_h is bijective. Since the pasted words of $Pas\widehat{L}_h$ are made of two words which are in $\widehat{\mathcal{E}}nds_h$, and which are independant in their successive factorizations, the claim holds also for h^{-1} acting on all $\partial\widehat{L}_h$.

♢

Corollary **4.10** *Let h be an expansive and circular morphism. Then, its corresponding $\widehat{\mathcal{E}}nds_h$ is bijectively mapped onto a Muller-recognizable language of right-infinite words $Cod\widehat{\mathcal{E}}nds_h$.*

Proof. Just put together the theorem and Lemma 4.4. ♢

Example **4.11** *The simplest example of a non-circular morphism was already given in Example 3.15 and is just $h(a) = aa$. This morphism has a binary embedding tree, although the only word in $Bi\partial\widehat{L}_h$ is clearly ${}^\omega a\mathbf{a}a^\omega$. The two ways of factorizing this word into subwords aa implies that there are infinitely many paths of the embedding forest leading to it.*

Example **4.12** *An other example of a non-circular morphism is given by $h(a) = aba$, $h(b) = bab$, which is clearly $(\{a, b\}, \pi)$-cyclic. The boundary $\partial\widehat{L}_h$ contains the word ${}^\omega(ab)\mathbf{a}b(ab)^\omega$, and there is no way of distinguishing between the factors $\mathbf{a}ba$, $b\mathbf{a}b$ or $ab\mathbf{a}$.*

Corollary **4.13** *Let h be an expansive and circular morphism. Then*

$$\widehat{\mathcal{E}}nds_h \cap Pas\widehat{L}_h = \emptyset.$$

Proof. This is just a consequence of the uniqueness of the factorization implied by circularity: it holds for pasted words too. ♢

4.5 Decidability of Circularity

Sufficient conditions for circularity are given by code theory [1]: if the morphism image of the letters, i.e $h(A)$, is a *synchronous, limited or circular code*, then circularity of the morphism is ensured. The morphisms $h(p) = ppq$, $h(q) = pq$ and $h(a) = b$, $h(b) = ab$ whose embedding forests are shown in Figures 1 and 2 can easily be checked circular by this argument. However,

Example **4.14** *The morphism $h(a) = acb$, $h(b) = acb$, $h(c) = cb$ is circular without being injective.*

Deciding circularity can be nevertheless fully solved for expansive morphisms: this is obtained by the following result given in [11]:

Theorem **4.15** **(Ehrenfeucht, Rozenberg).** *Let h be a morphism on A^+. Then, it is decidable whether or not h is n-power free for some integer n.*

Corollary **4.16** *Let h be an expansive morphism on A^+. Then, it is decidable if h is circular.*

Proof. This is deduced from the result of Mignosi and Séébold in [21] which solved the equivalence (1) <=> (2) in Theorem 4.9. $\diamondsuit$

4.6 The Coding of the Boundary

We have obtained a regular coding of the language $\widehat{\mathcal{E}}nds_h$. Now according to Theorem 3.13, a full coding of the boundary must involve the pasted set as well. Recall that the finite unpointed one-way boundaries $R\partial L_h$ and $L\partial L_h$ have been shown to consist of the words obtained by respectively iterating right-recurrent and left-recurrent letters. Their pointed counterparts $R\partial\widehat{L}_h$ and $L\partial\widehat{L}_h$ are both included in $\widehat{\mathcal{E}}nds_h$, since they were shown equal to $R\widehat{\mathcal{E}}nds_h$ and $L\widehat{\mathcal{E}}nds_h$ (see Proposition 3.9).

Lemma **4.17** *Let h be an expansive and circular morphism on A^+. Then, each pointed counterpart of each word in $R\partial L_h \cup L\partial L_h$ is bijectively mapped to a Muller-recognizable language included in $Cod\widehat{\mathcal{E}}nds_h$.*

Proof. Let us consider the right boundary set. For each letter s which is right-recurrent, define the subset of the connector alphabet Γ by : $\Gamma_s = \{g \in \Gamma |\quad f_h^{-1}(g) = \mathbf{s}u,\ u \in A^+\}$. Consider a right recurrent letter with its corresponding right unpointed infinite word. A pointed occurrence of this word has a label path in the embedding forest which must ultimately have only letters in Γ_s. This is easily translated by constraining the original Muller automaton recognizing $Cod\widehat{\mathcal{E}}nds_h$ by putting its table set equal to the power set of Γ_s. $\diamondsuit$

Corollary **4.18** *Let h be an expansive and circular morphism on A^+. Then, $L\widehat{\mathcal{E}}nds_h$, $R\widehat{\mathcal{E}}nds_h$, and $Pas\widehat{L}_h$ are each bijectively mapped to a Muller-recognizable language included in $Cod\widehat{\mathcal{E}}nds_h$.*

Proof. Since regularity is stable under union [26], the result holds for $L\widehat{\mathcal{E}}nds_h$ and $R\widehat{\mathcal{E}}nds_h$. For $Pas\widehat{L}_h$, each word is made of a word in $L\widehat{\mathcal{E}}nds_h$ or $R\widehat{\mathcal{E}}nds_h$, plus a pasted pair. Recall that the pasting pairs is a subset of $Pairs(L_h)$ which is finite. Let us denote it by $PasPairs(L_h)$. The pasted set is then coded by a regular language made of $PasPairs(L_h) \times (L\widehat{\mathcal{E}}nds_h \cup R\widehat{\mathcal{E}}nds_h)$. $\diamondsuit$

Thus we can go for one of the main results:

Theorem **4.19** *Let h be an expansive and circular morphism on A^+. Then, $\partial\widehat{L}_h$ can be bijectively mapped to a Muller-recognizable language of right-infinite words.*

Proof. This is a direct consequence of Corollaries 4.10 and 4.18 because of the relationship $\partial\widehat{L}_h = \widehat{\mathcal{E}}nds_h \cup Pas\widehat{L}_h$ which was proved in Theorem 3.13. $\diamondsuit$

Note that this result can also be stated in more topological terms, i.e. the closure of a letter substitution language $\widehat{L}_h$ in the metric space $({}^{\infty}\widehat{A}^{\infty}, d)$ can be done automatically.

4.7 Uncountability and Strict Quasiperiodicity of the Boundary Words

The first consequence of the effective construction of the boundary can be now obtained:

Proposition **4.20** *Let h be an expansive and circular morphism on A^+. Then, the boundary $\partial\widehat{L}_h$ is uncountable.*

Proof. We shall prove that the embedding forest of an expansive morphism always contains a tree which has at least a binary tree as subtree. According to the definition of the embedding map, this is equivalent to saying that there is a in A and $k \in \mathbb{N}$ such that $h^k(a)$ contains at least two occurrences of a.

For this purpose, associate to each letter a in A an integer k_a, such that $h^{k_a}(a)$ contains at least two occurrences of some letter, say s_a. Such integer exists for each letter since h is expansive. Then consider the graph where the nodes are the letters of A and where each letter a is bound to s_a. Denote by k the least common multiple of $\{k_a | a \in A\}$. Hence, applying h^k is the same as moving in the graph through one edge. Now since A is finite, the graph must have a cycle, so the claim follows. $\diamondsuit$

Corollary **4.21** *Let h be an expansive and circular morphism on A^+. Then, the unpointed boundary ∂L_h is uncountable.*

Proof. The equivalence classes have a countable number of words. $\diamondsuit$

Another consequence is related to the non-periodicity of the words in the boundary. A word w is said **quasiperiodic** if all its factors occur in bounded gaps, i.e. if v is a factor of w such that $|v| = n$, then there exists $m_n > 0$ so that if u is any factor of w with $|u| \geq m_n$, then u contains v. Clearly a periodic word is aperiodic. Thus we say that w is **strictly quasiperiodic** if it is quasiperiodic but not periodic. The next result relies on a well-known result in dynamical-oriented literature:

Proposition **4.22** *(see [27](Chapter 5)). Let h be a primitive morphism on A^+. Then, every word in $L\partial\widehat{L}_h \cup R\partial\widehat{L}_h$ is quasiperiodic.*

Proposition **4.23** *Let h be a primitive and circular morphism on A^+. Then, every word in $\partial\widehat{L}_h$ is strictly quasiperiodic.*

Proof. First, note that a primitive morphism is expansive. Then, a consequence of Theorem 4.9 is that boundary words cannot be periodic. Now from the last proposition, the result follows since extension of the words to both ways does not impair the quasiperiodicity in the primitive case. $\diamondsuit$

4.8 The Coding of the Unpointed Boundary

The unpointed boundary $(\partial\widehat{L}_h/\sim_\sigma)$ induces an equivalence relation for the regular coding of $\widehat{L}_h$. The following result was hinted at by Robison (see [28, 16]):

Proposition **4.24** *Let h be an expansive and circular morphism on A^+. Let $\widehat{w}$ and $\widehat{v}$ be two words in $\partial\widehat{L}_h$ in the same $\sim_\sigma$-class. Then, their corresponding path labels in the embedding forest differ by a finite prefix.*

Proof. Without loss of generality $\widehat{w}$ and $\widehat{v}$ can be assumed as belonging to $\widehat{\mathcal{E}}nds_h$. Being in the same $\sim_\sigma$-class means they are the same word with two different pointings. Consider the finite factor u between the two origins. Making use of the recovery method, consider the sequence $\{x_{0,k}\}_{k\in\mathbb{N}}$ as in Equation (24) associated to $\widehat{w}$, and the sequence $\{x'_{0,k}\}_{k\in\mathbb{N}}$ associated to $\widehat{v}$. Since h is circular and since $\widehat{w}$ is just a shifted occurrence of $\widehat{v}$ then $x_{0,k} = x'_{i_k,k}$, for some $i_k \in \mathbb{Z}$. The factors $x_{0,k}$ must ultimately cover all $\widehat{w}$, and $\widehat{v}$ too. Hence, as soon as the factor u is included in x_{0,n_0} for some $n_0 \in \mathbb{N}$, then $x_{0,n} = x'_{0,n}$, for all $n > n_0$, and the recovered labels are equal. $\diamondsuit$

This leads to an equivalence relation on the one-way infinite path labels: words with a different prefix of same finite length are identified. More precisely, let u, v be two path labels, then $u \approx v$ iff there is a right-infinite word w and two prefixes $u', v' \in A^+$, $|u'| = |v'|$ such that $u'w = u$ and $v'w = v$.

Thus, the unpointed version of Theorem 4.19 follows from Corollary 3.14 and from the following:

Corollary **4.25** *Let h be an expansive and circular morphism on A^+. Then, there is a bijective map from the unpointed $\mathcal{E}nds_h$ to $(Cod\widehat{\mathcal{E}}nds_h/\approx)$.*

Note however that for primitive morphisms, every factor of a word in $\partial\widehat{L}_h$ must also appear in all the other words of $\partial\widehat{L}_h$ and in bounded gaps. This property is also called **local isomorphism** in quasicrystal theory [19]. This implies that comparison of finite factors cannot generate an interesting metric $(\partial\widehat{L}_h/\sim_\sigma)$. In fact, as it was already discussed in the introduction (Section 2.3), the quotient metric space is metrically unseparated [3] in this case, thus giving the trivial topology.

4.9 A Step to the Decidability of the Boundary Equality

The boundary equality problem can be stated as follows: given two morphisms h_1 and h_2, one must decide whether the corresponding boundaries $\partial\widehat{L}_{h_1}$ and $\partial\widehat{L}_{h_2}$ are equal. Here, we shall give a partial answer to this problem: by using a result of Culik and Harju saying that the $D0L$ language equality problem is decidable [5], the boundary equality is proved decidable for primitive morphisms.

Theorem **4.26** *The boundary equality is decidable for primitive morphisms on A^+.*

Proof. Consider two primitive morphisms h_1 and h_2 over the same alphabet A. The result from [5] says that the equality $D0\widehat{L}_{h1}(a) = D0\widehat{L}_{h2}(a')$, with $a, a' \in A$, is decidable, whenever a and a' are recurrent letters. Hence, the equality

$$R\partial\widehat{L}_{h_1} \cup L\partial\widehat{L}_{h_1} = R\partial\widehat{L}_{h_2} \cup L\partial\widehat{L}_{h_2} \tag{28}$$

is decidable. Remark 3.17 says that each of these unions is sufficient to generate all the bi-infinite words of the boundary. Hence, $\partial\widehat{L}_{h_1} = \partial\widehat{L}_{h_2}$ holds iff Equation (28) holds. $\diamondsuit$

4.10 The Induced Dynamical System

A **dynamical system** is a pair (X, f) where X is a metric space and f is a continuous map from X to X. By definition, circularity implies that h^{-1} lets stable the boundary, that is $h^{-1}(\partial\widehat{L}_h) \subseteq \partial\widehat{L}_h$. Moreover, the map h^{-1} is continuous on $\partial\widehat{L}_h$: if two words are close, then they have long similar factors around their origins and their images by h^{-1} as well. Hence, the pair $(\partial\widehat{L}_h, h^{-1})$ defines a dynamical system. Moreover, the followings hold:

$$h^{-1}(\widehat{\mathcal{E}}nds_h) \subseteq \widehat{\mathcal{E}}nds_h, \quad \text{and} \quad h^{-1}(Pas\widehat{L}_h) \subseteq Pas\widehat{L}_h.$$

We may thus study these dynamical systems by looking at the regular coding of the $\widehat{\mathcal{E}}nds_h$ into the set $Cod\widehat{\mathcal{E}}nds_h$(for the system $(Pas\widehat{L}_h, h^{-1})$, pasting pairs of Pas$Pairs(L_h)$ must be also taken into account (see Corollary 4.18)).

Let us first define the **left shift map** for right infinite words:

$$\tau(w_n) = (w_{n+1}), \quad \forall n \in \mathbb{N}.$$

Its difference from the shift operator σ defined in the introduction (see Equation 1) is its non-reversibility: the first letter w_0 is definitely discarded from $\widehat{w}$ by $\tau(\widehat{w})$. Now according to the recovery method, we have that $\tau(Cod\widehat{\mathcal{E}}nds_h) \subseteq Cod\widehat{\mathcal{E}}nds_h$ (see the proof of Lemma 4.7). The map τ is clearly a continuous function. Recall also that $Cod\widehat{\mathcal{E}}nds_h$ is a finite type language (see Corollary 4.5). This implies that the pair $(Cod\widehat{\mathcal{E}}nds_h, \tau)$ is a classical dynamical system called a **subshift of finite type** (see for instance [9], Chapter 17).

Denote by Cod the map which bijectively sends $\widehat{\mathcal{E}}nds_h$ onto $Cod\widehat{\mathcal{E}}nds_h$. A consequence of the recovery method of the label path is that the following holds for every $\widehat{w} \in \widehat{\mathcal{E}}nds_h$:

$$Cod(h^{-1}(\widehat{w})) = \tau(Cod(\widehat{w}))$$

In other words, this means that the following diagram is commutative:

$$\begin{array}{ccc} \widehat{\mathcal{E}}nds_h & \xrightarrow{Cod} & Cod\widehat{\mathcal{E}}nds_h \\ {\scriptstyle h^{-1}}\downarrow & & \downarrow{\scriptstyle \tau} \\ \widehat{\mathcal{E}}nds_h & \xrightarrow{Cod} & Cod\widehat{\mathcal{E}}nds_h \end{array}$$

Within a system (X, f), a point $x \in X$ is said **p-periodic** if there exists an integer $p > 0$ such that $f^p(x) = x$. A 1-periodic point x is said to be a *fixed point*. A direct consequence of the commutativity of the diagram is that:

Remark **4.27** *The periodic points of $(\widehat{\mathcal{E}}nds_h, h^{-1})$ are in bijection with the periodic words of $Cod\widehat{\mathcal{E}}nds_h$.*

However, note that the diagram does not define a **topological conjugacy**, that is the bijection Cod is not necessarily a homeomorphism. In fact, Cod is continuous only from $\partial\widehat{L}_h$ to $Cod\widehat{\mathcal{E}}nds_h$, but Cod^{-1} may be not continuous.

Note also that the dynamical system $(\widehat{\mathcal{E}}nds_h, h^{-1})$ is a generalization of the fact that the one-way infinite recurrent words obtained in the $D0L$-system theory [30, 6, 27] were called *fixed points*.

References

[1] J. Berstel and D. Perrin, *Theory of codes*, Academic Press, 1985.

[2] L. Boasson and N. Nivat, *Adherence of languages*, J. Comp. Syst. Sc. **20** (1980), 285–309.

[3] A. Connes, *Géométrie non commutative*, Interéditions, 1990.

[4] E.M. Coven and M.S. Keane, *The structure of substitution minimal sets*, Transactions of the American Mathematical Society **162** (1971), 89–102.

[5] K. Culik II and T. Harju, *The $\omega-$sequence equivalence problem for D0L systems is decidable*, Journal of the ACM **31** (1984), no. 2, 282–298.

[6] K. Culik II and A. Salomaa, *On infinite words obtained by iterating morphisms*, Theoretical Computer Science **19** (1982), 29–38.

[7] N.G. De Bruijn, *Updown generation of Beatty sequences*, Indag. Math. **51** (1989), 385–407.

[8] ———, *Updown generation of Penrose patterns*, Indag. Mathem., New Series. **1** (1990), 201–219.

[9] M. Denker, C. Grillenberger, and K. Sigmund, *Ergodic theory on compact spaces*, Lecture Notes in Mathematics, vol. 527, Springer-Verlag, 1976.

[10] J. Devolder and I. Litovsky, *Finitely generated bi$\omega-$languages*, Theoretical Computer Science **85** (1991), 33–52.

[11] A. Ehrenfeucht and G. Rozenberg, *Repetitions of subwords in D0L languages*, Information and Control **59** (1983), 13–35.

[12] R. Engelking, *General topology*, Heldermann Verlag Berlin, 1989.

[13] F. Gire and N. Nivat, *Langages algébriques de mots bi-infinis*, Theoretical Computer Science **86** (1991), 277–323.

[14] W.H. Gottschalk, *Substitution minimal sets*, Transactions of the American Mathematical Society **109** (1963), 467–491.

[15] W.H. Gottschalk and G.A. Hedlund, *Topological dynamics*, American Math. Soc. Colloq. Pub **36** (1955).

[16] B. Grünbaum and G. Shephard, *Tilings and patterns*, Freeman &Co, 1987.

[17] T. Head, *Adherences of D0L languages*, Theoretical Computer Science **31** (1984), 139–149.

[18] ———, *The adherences of languages as topological spaces*, Automata on infinite words, Springer Verlag, 1984, Lecture Notes in Comp. Sci. vol. 192, pp. 147–163.

[19] D. Levine and P.J. Steinhardt, *Quasicrystals (I). Definition and structure*, Physical Review B **(2)34** (1986), 596–615.

[20] J.C. Martin, *Minimal flows arising from substitutions of non-constant length*, Math. Systems Th. **7** (1973), 73–82.

[21] F. Mignosi and P. Séébold, *If a D0L language is k-power free then it is circular*, Proceedings of ICALP'93, Sweden, Springer Verlag, 1993, Lecture Notes in Comp. Sci., 700, pp. 507–518.

[22] M. Morse, *A one-to-one representation of geodesics on a surface of negative curvature*, American Journal of Mathematics **43** (1921), 33–51.

[23] ———, *Recurrent geodesics on a surface of negative curvature*, Transactions of Amer. Math. Soc. **22** (1921), 84–100.

[24] M. Morse and G.A. Hedlund, *Symbolic dynamics I*, American Journal of Mathematics **60** (1938), 815–866.

[25] B. Mossé, *Puissances de mots et reconnaissabilité des points fixes d'une substitution*, Theoretical Computer Science **99** (1992), 327–334.

[26] D. Perrin and J.P. Pin, *Mots infinis*, Tech. Report 93.40, LITP, april 1993.

[27] M. Quéffelec, *Substitution dynamical systems - spectral analysis*, Lecture Notes in Mathematics, vol. 1294, Springer-Verlag, 1987.

[28] R.M. Robinson, *Comments on the Penrose tiles*, Mimeographed notes (1975), 177–209.

[29] G. Rozenberg and A. Salomaa, *The mathematical theory of l systems*, Academic press, 1980.

[30] A. Salomaa, *Jewels of formal language theory*, Computer Science Press, Rockville, MD, 1981.

[31] A. Thue, *Ueber unendliche zeichen reihen*, Selected Mathematical Papers, universitetsforlaget (1977) ed., 1906.

A Errata

- In the proof of Lemma 4.4, the table set of the Muller automaton must be defined as the power set of the set of states Q minus the power set of the complementary of F in Q.

  Note that this has no influence on Corollary 4.5.

- In the proof of Lemma 4.17, dual automata must be used.

  We here explain in details what this means and implies: Strict growth of the words along the paths of the embedding trees can be directly checked as it is done in Lemma 4.4 (non-growing means a letter a such that $h(a)$ is also a single letter, i.e. a state which has a single incoming arc). But this is not anymore the case to handle the way words grow rel. to their endpoints as it is necessary in Lemma 4.17, and one must consider dual automata instead: Let $\mathcal{A} = (Q, I, F, E, A)$ be an automaton; then, its **dual** $\mathcal{A}' = (Q', I', F', E', A')$ is given by:

  $$Q' = A,$$
  $$A' = Q,$$
  $$i' \in I' \subseteq Q' \text{ iff there exists } (i, i', .) \text{ in } E \text{ where } i \in I,$$
  $$f' \in F' \subseteq Q' \text{ iff there exists } (., f', f) \text{ in } E \text{ where } f \in F,$$
  $$(a, q_2, b) \in E' \text{ iff there exists } (., a, q_2) \text{ and } (q_2, b, .) \text{ in } E.$$

  Now, $\Gamma_s = \{g \in \Gamma | f_h^{-1}(g) = \mathbf{s}u,\ u \in A^+\}$ can be defined as a set of states in the dual automaton $\mathcal{A}'$ of the Muller automaton $\mathcal{A}$ recognizing the set of ends. Also define $\Gamma_{finite} = \{g \in \Gamma | f_h^{-1}(g) = \mathbf{s},\ s \in A\}$. Then for the right boundary set, the table set of $\mathcal{A}'$ is equal to the power set of $\Gamma_s \cup \Gamma_{finite}$ minus the power set of Γ_{finite}.

Misprints appearing only in the printed version:

- In Example 3.2 one should read "Fig. 1" instead of "Fig. 2".
- In Lemma 4.17, one should read $R\partial L_h \cup L\partial L_h$ instead of $R\partial \widehat{L}_h \cup L\partial \widehat{L}_h$
- The bibliography didn't appear in a strict alphabetical order.