

A COMBINATORIAL PROOF OF S -ADICITY FOR SEQUENCES WITH LINEAR COMPLEXITY

Julien Leroy¹

LAMFA, CNRS UMR 6410, Univ. de Picardie Jules Verne, 33 Rue Saint Leu,
80039 cedex 01, France
julien.leroy@uni.lu

Gwenaël Richomme

LIRMM (CNRS, Univ. Montpellier 2), UMR 5506 - CC 477, 161 rue Ada, 34095
Montpellier Cedex 5 **and** Univ. Montpellier 3, Route de Mende, 34199 Montpellier
Cedex 5, France
gwenael.richomme@lirmm.fr

Received: , Revised: , Accepted: , Published:

Abstract

Using Rauzy graphs, Ferenczi proved that if a symbolic dynamical system has linear complexity then it is S -adic. Being more specific, the result can also be proved for infinite words. We provide a new proof of this latter result using the notion of return words to a set of words.

1. Introduction

Since Morse and Hedlund's works [18], due to their numerous properties and applications, Sturmian words were given a lot of attention (see for instance surveys and studies in [1, 3, 8, 15, 20]). In this paper, we are concerned with two of these properties. First, Sturmian words are examples of words with linear complexity [18]. Recall that the complexity of an infinite word is the function that associates to each integer n the number of factors (blocks of consecutive letters) of length n that the infinite word contains. Secondly, Sturmian words have a multi-scale property since each Sturmian word is the image of another Sturmian word by a morphism (a map, also called substitution, defined on letters and compatible with the concatenation operation) taken from a finite set S_{sturm} consisting of 4 morphisms [17]. More precisely, given a Sturmian word \mathbf{s} , one can find a sequence $(f_i)_{i \geq 0}$ of morphisms of S_{sturm} and a letter $(a_i)_{i \geq 0}$, such that $\mathbf{s} = \lim_{n \rightarrow \infty} f_0 \cdots f_n(a_{n+1}^\omega)$. Sturmian words

¹This paper was partially written while the first author was invited to Montpellier in June 2011 thanks to funding from Groupe AMIS from Montpellier 3.

are then called S_{Sturm} -adic. More generally, an infinite word \mathbf{w} is said to be S -adic when S is a finite set of morphisms and the previous equality holds for \mathbf{w} instead of \mathbf{s} and for a suitable sequence of morphisms of S .

Since Pansiot's works [19], it is known that fixed points of morphisms that are S -adic words for a singleton set S do not necessarily have linear complexity. Also, Cassaigne showed [21] that any word $\mathbf{w} = (\mathbf{w}_i)_{i \geq 0}$ (where each \mathbf{w}_i is a letter) over an alphabet A is S -adic, with $S = \{\sigma_a, \varphi_a \mid a \in A\}$, where, using a letter α that does not belong to A , the morphisms σ_a and φ_a are defined by, for all letters a, b in A , $\sigma_a(\alpha) = a$, $\varphi_a(\alpha) = \alpha a$, $\sigma_a(b) = b = \varphi_a(b)$: $\mathbf{w} = \lim_{n \rightarrow \infty} \sigma_{\mathbf{w}_0} \varphi_{\mathbf{w}_1} \cdots \varphi_{\mathbf{w}_n}(\alpha^\omega)$.

Nevertheless, it can be observed that (as far as we know) all infinite words with linear complexity that have been studied are S -adic for a suitable (finite) set S of morphisms. Moreover, an important result due to Ferenczi in the context of symbolic dynamical systems can be reformulated in the context of infinite words as follows.

Theorem 1.1. ([12], see also [13]) *If an aperiodic and uniformly recurrent word has linear complexity then, for a finite set S of non-erasing morphisms, it is an everywhere growing S -adic sequence.*

The *everywhere growing property*² means that there exists a sequence $(\sigma_n)_{n \in \mathbb{N}} \in S^{\mathbb{N}}$ of morphisms of S generating the considered infinite word such that for all letters a in A , the length of $\sigma_0 \sigma_1 \cdots \sigma_n(a)$ tends to infinity with n . This property is not verified by Cassaigne's morphisms. The previous result of Ferenczi and other considerations (as for instance, the existence of an S -adic characterization of the family of linearly recurrent words [10]) are the origin of the S -adic conjecture which states the existence of a strong definition of S -adicity that would be equivalent to having linear complexity for a symbolic dynamical system or an infinite word (see [20, Section 12.1]). Unfortunately the everywhere growing property is not sufficient to solve this conjecture. Some examples [11] tend to show that the strong expected definition (if it exists) should concern not only the set S of morphisms but also the sequence of these morphisms used to define the infinite word.

Ferenczi's proof of Theorem 1.1 generalizes an idea of Arnoux and Rauzy [2]; it is based on the study of Rauzy graphs, a tool independently introduced by Rauzy [23] and Boshernitzan [5] and now often used to study factor complexity of words. The Rauzy graph G_n of order n of an infinite word \mathbf{w} is the directed graph whose vertices are the factors of length n of \mathbf{w} and such that any factor $u = u_1 \cdots u_{n+1}$ of length $n + 1$ of \mathbf{w} induces an edge in G_n from $u_1 \cdots u_n$ to $u_2 \cdots u_{n+1}$. This edge is labelled by the pair (u_1, u_{n+1}) . The label of a path in the graph is the concatenation of labels of edges along the path.

Leroy [13] revisited Ferenczi's proof in order to provide more information on the

²This property was called *ω -growth property* in [13]. However, when S is a singleton, the terminology *everywhere growing* is widely used so it is natural to keep working with it.

morphisms arising in the S -adic representation. He succeeded in going deeper into the study of underlying Rauzy graphs. He observed, without going further, that some important paths (called segments) in the Rauzy graphs used by Ferenczi are related to the notion of return words to a set. More precisely, the first (resp. the second) component of their labels corresponds to left (resp. right) return words to the set of left special factors. In this paper, we use this notion of return words to prove Theorem 1.1. Even though it is an adaptation of Ferenczi's, we hope that our more compact proof will open new ideas to tackle the S -adic conjecture.

An interest of our approach may be to provide non-contextual definitions of short or long return words before the proof when their corresponding terms (short or long segments) in the context of Rauzy graphs are defined in Ferenczi's proof. Furthermore, this concept of return words to a set W generalizes the one of Rauzy graphs since it does not require that the words in W share the same length.

The next two sections are preliminaries for our proof in Section 4. They allow us to present these notions of short and long return words in a general context independently of the complexity of the infinite words considered.

2. Return words to special words

The aim of this section is to recall the main combinatorial notions used in our proof. This is done through the development of an example that will be used in Section 3 to illustrate the existence of short return words that are not constant return words.

2.1. Words and morphisms

We assume that readers are familiar with combinatorics on words; for basic (possibly omitted) definitions, we follow [4, 14, 15].

Given an *alphabet* A , that is, a finite set of symbols called *letters*, we let A^* denote the set of all finite words over A (that is, the set of all finite sequences of elements of A). As usual, the *concatenation* of two words u and v is simply denoted uv . It is well known that the set A^* equipped with the concatenation operation is a free monoid with neutral element ε , the *empty word*.

Given two alphabets A and B , a free monoid morphism, or simply *morphism* σ , is a map from A^* to B^* such that $\sigma(uv) = \sigma(u)\sigma(v)$ for all words u and v over A (observe that this implies $\sigma(\varepsilon) = \varepsilon$). It is well known that a morphism is completely determined by the images of the letters.

From now on, denoting $\Gamma = \{e_0, e_1, e_2, e_3, b_1, b_2, b_3, b_4, a_1, a_2, 1, 2, c\}$ and $\Sigma = \{e_1, e_2, e_3, b, a, 1, 2, c\}$, we consider the following two morphisms ν and ψ defined respectively from Γ^* to Γ^* and from Γ^* to Σ^* by

$$\begin{array}{ll}
 \nu(e_0) = e_0 1 b_4 1 1 e_0 & \psi(e_0) = e_1 \\
 \nu(e_1) = e_1 c c b_1 1 e_1 & \psi(e_1) = e_1 \\
 \nu(e_2) = e_2 c c b_2 2 e_2 & \psi(e_2) = e_2 \\
 \nu(e_3) = e_3 c c b_3 2 e_3 & \psi(e_3) = e_3 \\
 \nu(b_1) = c c c e_1 a_1 b_1 & \psi(b_1) = b \\
 \nu(b_2) = c c c e_2 a_2 b_2 & \psi(b_2) = b \\
 \nu(b_3) = c c c c e_3 b_3 & \psi(b_3) = b \\
 \nu(b_4) = c c c c c b_4 & \psi(b_4) = b \\
 \nu(1) = c c c c b_1 1 & \psi(1) = 1 \\
 \nu(2) = c c c c b_1 2 & \psi(2) = 2 \\
 \nu(a_1) = c a_1 c c e_0 e_1 & \psi(a_1) = a \\
 \nu(a_2) = c a_2 c c e_2 e_3 & \psi(a_2) = a \\
 \nu(c) = c c c c a_1 a_2 & \psi(c) = c
 \end{array}$$

One can observe that ν is *uniform*, that is, the images of the letters have the same length (let us recall that the *length* of a word u , denoted by $|u|$, is the number of letters of u). The morphism ν is also *primitive*, that is, there exists an integer k (here $k = 5$) such that for all letters x, y in Γ , the letter y occurs in $\nu^k(x)$.

Note also that the morphisms ν and $\psi \circ \nu$ are injective. Moreover, they are synchronizing morphisms; a morphism f from A^* to B^* is *synchronizing* if for all letters x, y, z in the alphabet A and words p, s over A , the equality $f(x)f(y) = pf(z)s$ implies $p = \varepsilon$ or $s = \varepsilon$.

2.2. Infinite words and factors

A (right) *infinite word* (or *sequence*) over an alphabet A is an element of A^ω .

When a morphism is not *erasing*, that is, the images of letters are never the empty word, the notion of morphism extends naturally to infinite words. From now on, \mathbf{u} will be the fixed point of ν ($\mathbf{u} = \nu(\mathbf{u})$) starting with the letter c .

A word v is a *factor* of a finite or infinite word w over an alphabet A if it occurs in w or equivalently, if $w = xvy$ for some words x, y in $A^* \cup A^\omega$. When $x = \varepsilon$ (resp. $y = \varepsilon$), v is a *prefix* (resp. *suffix*) of w . We let $F(w)$ denote the set of all factors of w .

Observe that for all $n \geq 0$, words $\nu^n(1)$ and $\nu^n(2)$ are factors of \mathbf{u} (let us recall that, given a morphism f and an integer $n \geq 0$, f^n is the identity if $n = 0$, and is $f^{n-1} \circ f$ otherwise).

Fact 2.1. *Let $n \geq 0$ be an integer and v be a factor of \mathbf{u} . If $\psi(v) = \psi(\nu^n(1))$ then $v = \nu^n(1)$. If $\psi(v) = \psi(\nu^n(2))$ then $v = \nu^n(2)$.*

Proof. Let $i \in \{1, 2\}$. The result is immediate for $n = 0$ from the definitions of ν and ψ . Assume $n \geq 1$. Since $\psi \circ \nu$ is synchronizing and $\psi(v)$ is a factor of $\psi(\mathbf{u}) = \psi \circ \nu(\mathbf{u})$, there must exist a word v' such that $\psi(v) = \psi \circ \nu(v')$. The injectivity of $\psi \circ \nu$ implies $v' = \nu^{n-1}(i)$. Observe that for all letters α in Γ and for

all words x in Γ^* , the equality $\psi(x) = \psi \circ \nu(\alpha)$ implies $x = \psi(\alpha)$. This extends to α in Γ^* . Hence $v = \nu^n(i)$. \square

As with any fixed point of a primitive morphism [22], the word \mathbf{u} is an example of a *uniformly recurrent* infinite word, *i.e.*, its factors occur infinitely often and with bounded gaps.

The complexity function $p_{\mathbf{w}}$ of an infinite word \mathbf{w} is the function that counts the number of factors of each length (See [7] for a recent survey on this function). An infinite word \mathbf{w} is said to have *linear complexity* if there is a constant C such that $p_{\mathbf{w}}(n) \leq Cn$ for all $n \geq 1$. Again, as with any fixed point of a primitive morphism [22], the word \mathbf{u} is an example of an infinite word with linear complexity.

2.3. Special factors

Let us recall that a factor u of an infinite word \mathbf{w} over an alphabet A is *left special* if there exist at least two different letters a and b in A such that the words au and bu are factors of \mathbf{w} .

For all $n \geq 1$, observing that the words $\nu^n(1)$ and $\nu^n(2)$ differ only by their last letters, we let v_n denote their longest common prefix; $v_0 = \varepsilon$, $v_n = \nu(v_{n-1})ccccb_1$ for $n \geq 1$. We also let u_n denote the word $\psi(v_n)$.

Fact 2.2. *For all $n \geq 1$, the words abu_n , bu_n2 and u_n1e_1 are examples of left special factors of $\psi(\mathbf{u})$.*

Proof. As the words b_11e_1 , b_22e_2 , $11e_0$, and b_32e_3 are factors of \mathbf{u} , for all $n \geq 1$, the following words are also factors of \mathbf{u} : $e_1a_1b_1v_n1e_1c$, $e_2a_2b_2v_n2e_2c$, $1v_n1e_01$, $e_3b_3v_n2e_3$.

Thus, the words $e_1abu_n1e_1c$, $e_2abu_n2e_2c$, $1u_n1e_11$, $e_3bu_n2e_3$ are factors of $\psi(\mathbf{u})$. Hence the fact. \square

As a consequence of the previous fact, the word $\psi(\nu(u))$ has arbitrarily large special factors. Thus it is *aperiodic*, that is, denoting by x_i the letter in position i of $\psi(\nu(u))$, there does not exist any positive integers m, k such that, for all $i \geq m$, $x_i = x_{i+k}$.

We will also need the next property.

Lemma 2.3. *For all $n \geq 1$, the following words are factors of $\psi(\mathbf{u})$ but are not left special factors of $\psi(\mathbf{u})$: bu_n1 , u_n1e_1c , u_n1c , u_n1e_11 , u_n2e_2 , u_n2e_3 .*

Proof. First, let $x = bu_n1 = b\psi(\nu^n(1))$ and assume that, for a letter α , αx is a factor of $\psi(\mathbf{u})$. As $n \geq 1$, u_n starts with ccc . This implies that αb must be a suffix of a word in $\psi(\nu(\Gamma))$ and so $\alpha \in \{a, e_3\}$. If $\alpha = e_3$, then \mathbf{u} must contain as a factor the word $e_3b_3v_n1 = e_3b_3\nu^n(1)$ (see Fact 2.1). But one can observe that \mathbf{u} contains no factor of the form $b_3\nu^n(1)$. Thus $\alpha = a$ and bu_n1 is not a left special factor of \mathbf{u} .

Now let $x = u_n 1e_1c = \psi(\nu^n(1))e_1c$. If αx is a factor of $\psi(\mathbf{u})$ for a letter α , then there exists a letter β such that (we use here Fact 2.1 and the fact that e_0c is not a prefix of a word in $\nu(\Gamma)$) $\beta v_n 1e_1c$ is a factor of \mathbf{u} with $\alpha = \psi(\beta)$. We let the reader verify (for instance by induction) that if, for an integer ℓ and a letter δ , $\delta \nu^\ell(1)e_1$ is a factor of \mathbf{u} , then $\delta = b_1$. Hence $\beta = b_1$ and $\alpha = b$.

The proof for other words is similar. Let us just state the following facts (with the same notation as above).

$$\begin{aligned} x = u_n 1c &\Rightarrow \beta \in \{b_1, b_4\} && \text{and} && \alpha = b \\ x = u_n 1e_1 1 &\Rightarrow \beta = b_1 && \text{and} && \alpha = b \\ x = u_n 2e_2 &\Rightarrow \beta \in \{b_1, b_2\} && \text{and} && \alpha = b \\ x = u_n 2e_3 &\Rightarrow \beta \in \{b_1, b_3\} && \text{and} && \alpha = b \end{aligned}$$

□

2.4. Return words

Given a set L of factors of an infinite word \mathbf{w} , a word r is called a (left) *return word* to L in \mathbf{w} if there exist two words u and v in L such that rv is a factor of \mathbf{w} , rv admits u as a prefix and rv contains no other occurrences of words of L . Clearly if a set L contains a prefix of \mathbf{w} and \mathbf{w} is *recurrent* (all of its factors occur infinitely often), then \mathbf{w} can be decomposed over the set of left return words to L .

In this paper, given an infinite word \mathbf{w} , we will mostly consider the set $\mathcal{R}_n(\mathbf{w})$ of return words to the set $\text{LS}_n(\mathbf{w})$ composed of the prefix of length $n \geq 1$ of \mathbf{w} and of all left special factors of length n of \mathbf{w} . At the end of this section, we show that a , b and ab are elements of $\mathcal{R}_n(\mathbf{u})$ (where \mathbf{u} is the word defined in the previous sections) with a more precise description of the values of n for which this membership holds.

Lemma 2.4. *For all $n \geq 1$, $\{a, b\} \subseteq \mathcal{R}_n(\mathbf{u})$.*

For all $n \geq 0$, $ab \in \mathcal{R}_{|u_n|+2}(\mathbf{u}) = \mathcal{R}_{6^n+1}(\mathbf{u})$.

Proof. Let $n \geq 1$, $|u_n| = 6^n - 1 \geq n$. Let p_i be the prefix of length i of u_n where $0 \leq i \leq |u_n|$. From Fact 2.2, we know that abp_n is a factor of $\psi(\mathbf{u})$ and that the words abp_{n-2} (when $n \geq 2$), bp_{n-1} and p_n are all left special factors (and all of length n). As p_{n-2} is a prefix of p_{n-1} , itself a prefix of p_n , a and b belong to $\mathcal{R}_n(\mathbf{u})$.

The word $abu_n 1e_1$ is also a factor of $\psi(\mathbf{u})$ with abu_n and $u_n 1e_1$ left special factors. As, by Lemma 2.3, $bu_n 1$ is not a left special factor, $ab \in \mathcal{R}_{|u_n|+2}(\mathbf{u}) = \mathcal{R}_{6^n+1}(\mathbf{u})$. □

The last part of Lemma 2.4 admits a converse.

Lemma 2.5. *Let $n \geq 2$ be an integer. If $ab \in \mathcal{R}_n(\mathbf{u})$ then $n = 6^k + 1$ for some $k \geq 0$.*

Although this lemma will allow the important Remark 3.1, we provide its proof in an appendix (see Section 5) as it is rather long, technical, and a bit far from the main aim of the paper.

3. Short, constant and long return words

In this section, we consider particular return words that will be useful in the next section where we present our proof of Theorem 1.1. Note that, as already said in the introduction, the terms short and long were already used in Ferenczi’s paper [12] but, even though our concepts are related, they are differently defined. For instance the fact that the length of long return words tends to infinity is a consequence of the definition, not part of it.

It is important to observe that all results in this section are independent of the complexity of the words considered.

3.1. Definitions

We will need to distinguish various types of return words to the set $LS_n(\mathbf{w})$. A factor u of \mathbf{w} is called a *short return word* of \mathbf{w} if for all $n \geq 0$, there exists an integer $N \geq n$ such that u is an element of $\mathcal{R}_N(\mathbf{w})$. When a factor u is a return word of \mathbf{w} (i.e. belongs to $\mathcal{R}_n(\mathbf{w})$ for some integer n) without being a short return word, we said that u is a *long return word*. Amongst short return words, we say that a factor u of \mathbf{w} is a *constant return word* of \mathbf{w} if there exists an integer $N \geq 0$ such that for all $n \geq N$, u is an element of $\mathcal{R}_n(\mathbf{w})$. For $n \geq 0$, we let respectively $\mathcal{R}_n^{(L)}(\mathbf{w})$, $\mathcal{R}_n^{(S)}(\mathbf{w})$ and $\mathcal{R}_n^{(C)}(\mathbf{w})$ denote the set of long return words to $LS_n(\mathbf{w})$, the set of short return words to $LS_n(\mathbf{w})$ and the set of constant return words to $LS_n(\mathbf{w})$.

Remark 3.1. Our example \mathbf{u} studied in Section 2 shows that not all short return words are constant return words. More precisely Lemma 2.4 and Lemma 2.5 prove that a , b and ab are examples of short return words with a and b constant return words and ab not a constant return word. Indeed, for all n we have $a, b \in \mathcal{R}_n^{(C)}(\psi(\mathbf{u})) \subseteq \mathcal{R}_n^{(S)}(\psi(\mathbf{u}))$; $ab \in \bigcup_{i=1}^n \mathcal{R}_i^{(S)}(\psi(\mathbf{u})) \setminus \bigcup_{i=1}^n \mathcal{R}_i^{(C)}(\psi(\mathbf{u}))$.

As c , cc and ccc are left special factors of $\psi(\mathbf{u})$, and as $cccc$ is a factor of $\psi(\mathbf{u})$, c is a return word of $\psi(\mathbf{u})$, and more precisely, $c \in \mathcal{R}_1(\psi(\mathbf{u})) \cap \mathcal{R}_2(\psi(\mathbf{u})) \cap \mathcal{R}_3(\psi(\mathbf{u}))$. Observe that for all words x of length at least four such that cx is a factor of $\psi(\mathbf{u})$, x is always preceded in $\psi(\mathbf{u})$ by the letter c . Thus, for all $n \geq 4$, c is not a return word to $LS_n(\psi(\mathbf{u}))$; $c \in (\mathcal{R}_1^{(L)}(\psi(\mathbf{u})) \cap \mathcal{R}_2^{(L)}(\psi(\mathbf{u})) \cap \mathcal{R}_3^{(L)}(\psi(\mathbf{u})) \setminus \mathcal{R}_4^{(L)}(\psi(\mathbf{u}))$.

One would like to have examples of short or constant return words for an infinite word defined over a smaller alphabet than Σ . We let the reader check that this can be obtained by recoding $\psi(\mathbf{u})$ for instance with the following morphism f defined from Σ to $\{a, b\}$ by $f(a) = a^8b$, $f(b) = a^7b^2$, $f(e_1) = a^6b^3$, $f(e_2) = a^5b^4$, $f(e_3) = a^4b^5$, $f(1) = a^3b^6$, $f(2) = a^2b^7$, $f(c) = ab^8$.

We will see that for infinite words with linear complexity the set of short return words is finite. Thus it is convenient to let $\mathcal{R}^{(S)}(\mathbf{w})$ (resp. $\mathcal{R}^{(C)}(\mathbf{w})$) denote the set $\bigcup_{n \geq 1} \mathcal{R}_n^{(S)}(\mathbf{w})$ (resp. $\bigcup_{n \geq 1} \mathcal{R}_n^{(C)}(\mathbf{w})$).

3.2. Coherent decompositions

From our example, we can observe that the sets $\mathcal{R}_n(\mathbf{w})$ may not be codes (a *code* is a set of words such that any word w in C^* has a unique decomposition over C). This leads us to introduce the notion of coherent decomposition.

A sequence $(r_k)_{k \in \mathbb{N}} \in (\mathcal{R}_n(\mathbf{w}))^{\mathbb{N}}$ such that $\mathbf{w} = r_0 r_1 r_2 \cdots$ is called a *decomposition* of \mathbf{w} . The unique decomposition $(r_k)_{k \in \mathbb{N}}$ of \mathbf{w} such that a factor of \mathbf{w} belongs to $\text{LS}_n(\mathbf{w})$ if and only if it is the prefix of length n of $r_i r_{i+1} \cdots$ for some $i \geq 0$ is called *the coherent decomposition* of \mathbf{w} of level n .

Let s_1, \dots, s_k be words. Let n be an integer and let \mathbf{w} be an infinite word. We say that (s_1, \dots, s_k) is a factor of the coherent decomposition of \mathbf{w} of level n if there exists an integer $m \geq 0$ such that for all i in $\{1, \dots, k\}$, $s_i = r_{m+i-1}$. Note that if (s_1, \dots, s_k) is a factor of the coherent decomposition of \mathbf{w} of level n , then $\{s_1, \dots, s_k\} \subseteq \mathcal{R}_n(\mathbf{w})$.

The next result shows a connection between the coherent decompositions of \mathbf{w} for two successive levels.

Lemma 3.2. *Let $(r_i)_{i \geq 0}$ and $(r'_i)_{i \geq 0}$ be the coherent decompositions of an infinite word \mathbf{w} respectively of level n and level $n + 1$ (for an integer $n \geq 1$). There exists a strictly increasing sequence of integers $(i_j)_{j \geq 0}$ such that $i_0 = 0$ and for all $j \geq 0$, $r'_j = r_{i_j} \cdots r_{i_{j+1}-1}$.*

Proof. This is a direct consequence of the definition of coherent decompositions and the observation that the prefix of length n of any element of $\text{LS}_{n+1}(\mathbf{w})$ is an element of $\text{LS}_n(\mathbf{w})$. □

3.3. Short return words are concatenations of constant return words

In this section, we prove the following proposition that claims in a more precise way that any short return word of an infinite word \mathbf{w} is a concatenation of constant return words to $\text{LS}_n(\mathbf{w})$.

Proposition 3.3. *Given an infinite word \mathbf{w} , for any short return word s of \mathbf{w} , there exist infinitely many integers n such that, for the coherent decomposition $(r_i)_{i \geq 0}$ of level n of \mathbf{w} , $s = r_i \cdots r_j$ for some integers i and j such that all words r_ℓ , $i \leq \ell \leq j$, are constant return words.*

Actually Proposition 3.3 is a particular case of the following technical lemma (Proposition 3.3 corresponds to the case $k = 1$ of Lemma 3.4).

Lemma 3.4. *Let \mathbf{w} be an infinite word and let s_1, \dots, s_k be words such that (s_1, \dots, s_k) is a factor of the coherent decomposition of \mathbf{w} of level n for infinitely many integers n (this implies in particular that all words s_1, \dots, s_k are short return words). There exist integers i_0, i_1, \dots, i_k with $i_0 = 0 < i_1 < \dots < i_k$ and*

there exist constant return words r_1, \dots, r_{i_k} such that (r_1, \dots, r_{i_k}) is a factor of the coherent decomposition of \mathbf{w} of level m for infinitely many integers m , and for all $j \in \{1, \dots, k\}$, $s_j = r_{i_{j-1}+1} \cdots r_{i_j}$.

Proof. We start by defining a partial order on tuples of words. Given any words $t_1, \dots, t_\ell, s_1, \dots, s_k$ such that $t_1 \dots t_\ell = s_1 \dots s_k$, we let $(t_1, \dots, t_\ell) < (s_1, \dots, s_k)$ denote the fact that there exists an integer j in $\{1, \dots, \min(\ell, k)\}$ such that for all $i = 1, \dots, j - 1$, $t_i = s_i$ and $|t_j| < |s_j|$. Observe that, given a word s , this partial order is a total order on the set of all tuples (s_1, \dots, s_k) such that $s = s_1 \dots s_k$. This allows us to prove Lemma 3.4 by induction.

Let s_1, \dots, s_k and \mathbf{w} be as in the lemma. Let \mathcal{D} be the set of all integers n such that (s_1, \dots, s_k) is a factor of the coherent decomposition of \mathbf{w} of level n . By hypothesis, this set is infinite.

First, consider the case where for all j , s_j is of length 1. Let j , $1 \leq j \leq k$. For all n in \mathcal{D} , by definition of this set, there exists a word u_n of length n such that both u_n and the prefix of length n of $s_j u_n$ belong to $\text{LS}_n(\mathbf{w})$. As the prefix of length $n - 1$ of any element of $\text{LS}_n(\mathbf{w})$ is an element of $\text{LS}_{n-1}(\mathbf{w})$, we deduce that for all $n \geq 1$, there exists a word v_n of length n such that both v_n and the prefix of length n of $s_j v_n$ is an element of $\text{LS}_n(\mathbf{w})$. As \mathcal{D} is infinite, this shows that s_j is a constant return word.

We now turn to the case where there exists a word s_j (with $j \in \{1, \dots, k\}$) which is not a constant return word. It follows that there exists an infinite subset \mathcal{D}' of \mathcal{D} such that for all n in \mathcal{D}' , (s_1, \dots, s_k) is not a factor of the coherent decomposition of \mathbf{w} of level $n - 1$, whereas (s_1, \dots, s_k) is a factor of the coherent decomposition of \mathbf{w} of level n . By Lemma 3.2, there exist some words $t_1(n), \dots, t_{\ell(n)}(n)$ and some integers $i_0 = 0 < i_1 < \dots < i_k = \ell(n)$ such that for all $j = 1, \dots, k$, we have $s_j = t_{i_{j-1}+1}(n) \cdots t_{i_j}(n)$ and $(t_1(n), \dots, t_{\ell(n)}(n))$ is a factor of the coherent decomposition of \mathbf{w} of level $n - 1$: each word s_j is a composition of return words to $\text{LS}_{n-1}(\mathbf{w})$, and at least one is the concatenation of at least two such words. Moreover, we have $(t_1(n), \dots, t_{\ell(n)}(n)) < (s_1, \dots, s_k)$. Even though the words $t_1(n), \dots, t_{\ell(n)}(n)$ depend on n , as the length of $|t_1(n) \cdots t_{\ell(n)}(n)| = |s_1 \dots s_k|$ is fixed, there must exist an integer ℓ , some words t_1, \dots, t_ℓ , and an infinite set \mathcal{D}'' such that for all n in \mathcal{D}'' , $\ell(n) = \ell$ and for all $i = 1, \dots, \ell$, $t_i(n) = t_i$. The lemma ends by induction applied to (t_1, \dots, t_ℓ) . \square

3.4. More on long return words

The next result justifies *a posteriori* the terminology *long* for long return words.

Proposition 3.5. *For an infinite word \mathbf{w} , if $\mathcal{R}_n^{(L)}(\mathbf{w}) \neq \emptyset$ for infinitely many integers n , then*

$$\lim_{n \rightarrow \infty} \min\{|r| \mid r \in \mathcal{R}_n^{(L)}(\mathbf{w})\} = +\infty.$$

Proof. This result is a direct consequence of the definitions of short and long return words.

Assume by contradiction that $\lim_{n \rightarrow \infty} \min\{|r| \mid r \in \mathcal{R}_n^{(L)}(\mathbf{w})\}$ is finite. Let \mathcal{D} be an infinite subset of \mathbb{N} such that $\mathcal{R}_n^{(L)}(\mathbf{w}) \neq \emptyset$ for all n in \mathcal{D} . For all $i \in \mathcal{D}$, let r_i be a word with minimal length in $\mathcal{R}_i^{(L)}(\mathbf{w})$. As the length of r_i is bounded, the set $\{r_i \mid i \in \mathcal{D}\}$ is finite. In particular, there exists a word r such that $r \in \mathcal{R}_i^{(L)}(\mathbf{w})$ for all i in an infinite subset of \mathcal{D} . Thus, r is a short return word of \mathbf{w} , which is a contradiction as long return words cannot be short by definition. \square

3.5. Return words to a word

In almost the entire paper we are concerned with the set of return words to the set of left special words of an infinite word. The next important and well-known result uses the notion of return word to a single word.

A nonempty word r is a *return word* to a factor u of an infinite word \mathbf{w} if ru is a factor of \mathbf{w} starting with u and containing exactly two occurrences of u as a factor (one as a prefix and one as a suffix).

Proposition 3.6. *Let \mathbf{w} be an aperiodic and uniformly recurrent infinite word. We have*

$$\lim_{n \rightarrow +\infty} \min_{\substack{u \text{ factor of } \mathbf{w} \\ |u|=n}} \{|r| \mid r \text{ return word to } u\} = +\infty.$$

Although the previous result is certainly well known (see for instance [9] for a proof of a similar result), we provide a short proof in order to get a self-contained paper. Let us recall one fundamental lemma of combinatorics on words (see [16] and for instance [14, Prop. 1.3.4]).

Lemma 3.7. *For two nonempty words x and y , there exists a word z such that $xz = zy$ if and only if there exist two words u and v such that $x = uv$, $z \in (uv)^*u$ and $y = vu$.*

Proof of Proposition 3.6. Assume by contradiction that the limit in Proposition 3.6 is finite. This implies that there exists a word r which is the return word to infinitely many words $(u_n)_{n \geq 0}$. This means that there exist some words $(x_n)_{n \geq 0}$ such that $u_n x_n = r u_n$. By Lemma 3.7, this implies that r^k is a factor of \mathbf{w} for arbitrarily large k , which is a contradiction with the fact that \mathbf{w} is aperiodic and uniformly recurrent. \square

4. S -adicity of infinite words with linear complexity

In this section, we provide our proof of Theorem 1.1. Each subsection corresponds to a step or a key element of the proof. We first clarify the notion of S -adicity. Let

A be an alphabet. Recall that with the product topology, the set A^ω of infinite words over A is a compact metric space. As usual, for a finite word u , u^ω denotes the infinite word obtained by concatenating infinitely many copies of u .

Let \mathbf{w} be an infinite word over A . An *adic representation* of \mathbf{w} is given by a sequence $(A_n)_{n \in \mathbb{N}}$ of alphabets, a sequence $(\sigma_n : A_{n+1}^* \rightarrow A_n^*)_{n \in \mathbb{N}}$ of morphisms and a sequence $(a_n)_{n \in \mathbb{N}}$ of letters such that $a_i \in A_i$ for all $i \geq 0$, $A_0 = A$, $\lim_{n \rightarrow +\infty} |\sigma_0 \sigma_1 \cdots \sigma_n(a_{n+1})| = +\infty$ and

$$\mathbf{w} = \lim_{n \rightarrow +\infty} \sigma_0 \sigma_1 \cdots \sigma_n(a_{n+1}^\omega).$$

When all morphisms $\sigma_n, n \in \mathbb{N}$, belong to a given *finite* set S of morphisms, we say that \mathbf{w} is *S-adic* (often omitting the sequence of morphisms and letters).

4.1. The number of left special factors is bounded

The starting point of the proof of Theorem 1.1 is the following breakthrough result of Cassaigne result on complexity.

Theorem 4.1. [6] *An infinite word \mathbf{w} has linear complexity if and only if the first difference of its complexity $(p_{\mathbf{w}}(n + 1) - p_{\mathbf{w}}(n))$ is bounded.*

For a finite word u and an infinite word \mathbf{w} , let $\delta_{\mathbf{w}}^- u$ denote the number of letters a such that au is a factor of \mathbf{w} . Recall that a word u is a left special factor of \mathbf{w} if and only if $\delta_{\mathbf{w}}^- u \geq 2$. Since Rauzy’s works [23], it is well known that the first difference of complexity is related to special factors by the following Equation (1). For a recurrent infinite word \mathbf{w} ,

$$p_{\mathbf{w}}(n + 1) - p_{\mathbf{w}}(n) = \sum_{\substack{|u|=n, \\ u \text{ left special factor of } \mathbf{w}}} (\delta_{\mathbf{w}}^- u - 1) \tag{1}$$

Thus Theorem 4.1 has the following corollary (let us recall from Section 2.4 that $LS_n(\mathbf{w})$ is the set consisting of the prefix of length n and the left special factors of length n of \mathbf{w}).

Corollary 4.2. *A recurrent infinite word \mathbf{w} has linear complexity if and only if the sets $LS_n(\mathbf{w}), n \geq 0$, have bounded cardinality.*

4.2. The number of return words is bounded

Let us recall that $\mathcal{R}_n(\mathbf{w})$ is the set of return words to $LS_n(\mathbf{w})$ where \mathbf{w} is an infinite word and $n \geq 1$ is an integer.

Proposition 4.3. *If \mathbf{w} is a recurrent infinite word with linear complexity, then the cardinality of the sets $\mathcal{R}_n(\mathbf{w})$ is bounded.*

Proof. Let r be a return word to the set $LS_n(\mathbf{w})$. By definition of this set, there exists a word u in $LS_n(\mathbf{w})$ (u is of length n and is a left special factor or a prefix of \mathbf{w}) such that ru is a factor of \mathbf{w} . Moreover for each factor sp of ru of length n with $1 \leq |s| < |r|$, s suffix of r and p prefix of u , there exists a unique letter γ such that γsp is a factor of \mathbf{w} . Hence r is entirely determined by the pair (u, α) with α its last letter (but the pair may not be unique). Thus, $\text{Card}(\mathcal{R}_n(\mathbf{w})) \leq \sum_{u \in LS_n(\mathbf{w})} \delta^- u = p(n + 1) - p(n) + \text{Card}(LS_n(\mathbf{w}))$. The lemma follows from Theorem 4.1 and Corollary 4.2. \square

The reader will note that the cardinality of a set X is denoted by $\text{Card}(X)$.

4.3. The number of constant return words is finite

Let us recall from Section 3.1 that $\mathcal{R}^{(C)}(\mathbf{w}) = \cup_{n \geq 1} \mathcal{R}_n^{(C)}(\mathbf{w})$ is the set of all constant return words to the sets $LS_n(\mathbf{w})$.

Proposition 4.4. *For any recurrent infinite word \mathbf{w} with linear complexity, the set $\mathcal{R}^{(C)}(\mathbf{w})$ is finite.*

Proof. This is a direct consequence of Proposition 4.3. Let us consider $K = \max_{n \geq 1} \text{Card}(R_n(\mathbf{w}))$. Assume by contradiction that $R^{(C)}(\mathbf{w})$ is not finite. Let r_1, \dots, r_{K+1} be distinct elements of $R^{(C)}(\mathbf{w})$. There are some integers N_1, \dots, N_{K+1} such that for all i in $\{1, \dots, K + 1\}$ and for all $n \geq N_i$, r_i belongs to $R_n^{(C)}(\mathbf{w})$. Thus, for $n \geq \max\{N_i \mid 1 \leq i \leq K + 1\}$, we have $\{r_1, \dots, r_{K+1}\} \subseteq R_n(\mathbf{w})$, which is a contradiction with $\text{Card}(R_n(\mathbf{w})) \leq K$. \square

4.4. The number of short return words is finite

Let us recall from Section 3.1 that $\mathcal{R}^{(S)}(\mathbf{w}) = \cup_{n \geq 1} \mathcal{R}_n^{(S)}(\mathbf{w})$ is the set of all short return words to the sets $LS_n(\mathbf{w})$.

Proposition 4.5. *For any uniformly recurrent infinite word \mathbf{w} with linear complexity, the set $\mathcal{R}^{(S)}(\mathbf{w})$ is finite.*

This result is a consequence of the following technical lemma and Proposition 3.3.

Lemma 4.6. *Let \mathbf{w} be an aperiodic and uniformly recurrent infinite word with linear complexity. Let $(L_i)_{i \geq 1}$ be a family of sets of words such that $L_i \subseteq \mathcal{R}_i(\mathbf{w})$ for all $i \geq 1$ and $R = \cup_{i \geq 1} L_i$ is finite.*

There exist two integers K and N such that for all $n \geq N$ and for any coherent decomposition $(r_i)_{i \in \mathbb{N}}$ of \mathbf{w} into words of $\mathcal{R}_n(\mathbf{w})$, the number of consecutive elements of R in $(r_i)_{i \in \mathbb{N}}$ is bounded by K .

In particular, the concatenation of such return words in a coherent decomposition of level greater than N has length bounded by $K\ell$ where ℓ is the maximal length of words in R .

Proof. We let ℓ be, as defined in the lemma, the maximal length of words in R . By Corollary 4.2 and Proposition 4.3, there exists an integer K such that for all n , $\text{Card}(\mathcal{R}_n(\mathbf{w}) \times \text{LS}_n(\mathbf{w})) \leq K$. By Proposition 3.6, there exists an integer N such that for all factors u of length $n \geq N$ of \mathbf{w} , all return words to u have length greater than $K\ell$. Finally, consider an integer $n \geq N$ and let $(r_i)_{i \geq 0}$ be a coherent decomposition of \mathbf{w} of level n .

Assume that there exist $K + 1$ consecutive elements of L_n in the considered decomposition of \mathbf{w} , say r_m, \dots, r_{m+K} for an integer $m \geq 1$. For $i \geq 0$, let l_i be the prefix of length n of the word $r_{i+1}r_{i+2} \dots$. From the choice of K , there exist two integers i and j such that $m \leq i < j \leq m + K$ and $(r_i, l_i) = (r_j, l_j)$. Consequently, the word $r_{i+1} \dots r_j$ is a return word to l_i and $|r_{i+1} \dots r_j| \leq K\ell$, which contradicts the choice of N . \square

Proof of Proposition 4.5. By Proposition 3.3, any short return word is the concatenation of constant return words r_1, \dots, r_k such that (r_1, \dots, r_k) is a factor of infinitely many coherent decompositions of \mathbf{w} . By Proposition 4.4, the set $\mathcal{R}^{(C)}(\mathbf{w})$ is finite. We can then apply Lemma 4.6 with $L_n = \mathcal{R}^{(C)}(\mathbf{w}) \cap \mathcal{R}_n(\mathbf{w})$ for all $n \geq 0$ to deduce that short return words have bounded length. Consequently, the set of short return words is finite. \square

4.5. About long return words

As an immediate consequence of Lemma 4.6, the next result states that any sufficiently long factor of a coherent decomposition (of sufficiently large level n) contains a long return word.

Lemma 4.7. *Let \mathbf{w} be a uniformly recurrent infinite word with linear complexity. There exist some integers K and N such that for all $n \geq N$, if $k \geq 1$ and (r_1, \dots, r_k) is a factor of the coherent decomposition of \mathbf{w} of level n with $|r_1 \dots r_k| \geq K$, then at least one of the words r_i ($1 \leq i \leq k$) belongs to $\mathcal{R}_n^{(L)}(\mathbf{w})$.*

Proof. By Lemma 4.6 (taking $L_i = \mathcal{R}_i^{(S)}(\mathbf{w})$ for all $i \geq 0$), there exist some integers K_1 and N such that for all $n \geq N$, at most K_1 short return words can occur successively in the coherent decomposition of \mathbf{w} of level n . Let $\ell = \max\{r \mid r \in \mathcal{R}^{(S)}(\mathbf{w})\}$. Lemma 4.7 follows with $K = K_1\ell + 1$. \square

Remark 4.8. With the same notation as in Lemma 4.7, we let M denote the smallest integer such that $M \geq N$ and $\min\{|u| \mid u \in \mathcal{R}_M^{(L)}(\mathbf{w})\} \geq K$. Due to Lemma 3.2, we have $\min\{|u| \mid u \in \mathcal{R}_n^{(L)}(\mathbf{w})\} \geq K$ for all $n \geq M$.

4.6. Morphisms

Up to the end of Section 4, \mathbf{w} always denotes a recurrent infinite word with linear complexity. For all $n \geq 0$, let $\mathcal{U}_n(\mathbf{w})$ be the set of tuples $((r_1, \ell_1), \dots, (r_k, \ell_k))$, $k \geq 1$,

such that

- there exists an integer j , $1 \leq j \leq k$, such that $r_j \in \mathcal{R}_n^{(L)}(\mathbf{w})$ and for all $i \neq j$, $r_i \in \mathcal{R}_n^{(S)}(\mathbf{w})$ (by abuse of notation, we write $(r_1, \dots, r_k) \in (\mathcal{R}_n^{(S)}(\mathbf{w}))^* \mathcal{R}_n^{(L)}(\mathbf{w}) (\mathcal{R}_n^{(S)}(\mathbf{w}))^*$);
- with $(s_i)_{i \geq 0}$ the coherent decomposition of \mathbf{w} of level n , there exists an integer i such that $(r_1, \dots, r_k) = (s_i, \dots, s_{i+k-1})$ (i.e. (r_1, \dots, r_k) is a factor of the coherent decomposition of \mathbf{w} of level n), and for all $j = 1, \dots, k$, ℓ_j is the prefix of length n of $s_{i+j} \cdots$ ($\ell_j \in \text{LS}_n(\mathbf{w})$).

Lemma 4.9. *For any uniformly recurrent infinite word \mathbf{w} with linear complexity, the cardinality of the sets $(\mathcal{U}_n(\mathbf{w}))_{n \geq 0}$ is ultimately bounded, i.e., there exist some integers K and N such that for all $n \geq N$, $\text{Card}(\mathcal{U}_n(\mathbf{w})) \leq K$.*

Proof. By Lemma 4.6 (taking $L_i = \mathcal{R}_i^{(S)}(\mathbf{w})$ for all $i \geq 0$), there exist some integers K and N such that for all $n \geq N$, at most K short return words can occur successively in the coherent decomposition of \mathbf{w} of level n . Note that, by definition of $\mathcal{U}_n(\mathbf{w})$, for any of its elements $((r_1, \ell_1), \dots, (r_k, \ell_k))$, $r_1 \dots r_k \ell_k$ is a factor of \mathbf{w} and for all $j = 1, \dots, k - 1$, ℓ_j is the prefix of length n of $r_{j+1} \dots r_k \ell_k$ (ℓ_j is uniquely determined by r_{j+1}, \dots, r_k and ℓ_k). By Proposition 4.5 and Proposition 4.3, it follows that, for all $n \geq N$, the cardinality of $\mathcal{U}_n(\mathbf{w})$ admits the upper bound $(\text{Card}(\mathcal{R}_n^{(S)}(\mathbf{w})))^{2(K+1)} K' \text{Card}(\text{LS}_n(\mathbf{w}))$ with $K' = \max_{n \geq 0} \text{Card}(\mathcal{R}_n^{(L)}(\mathbf{w}))$. \square

Let K and M be integers as in Lemma 4.7 and Remark 4.8. Let $n \geq M$ and let $((r_1, \ell_1), \dots, (r_k, \ell_k))$ be an element of $\mathcal{U}_{n+1}(\mathbf{w})$. Let $(s_i)_{i \geq 0}$ be the coherent decomposition of \mathbf{w} of level n . By Lemma 3.2, there exists a strictly increasing sequence of integers $(i_j)_{1 \leq j \leq k+1}$ such that for all $j \geq 0$, $r_j = s_{i_j} \cdots s_{i_{j+1}-1}$ and ℓ_k is a prefix of $s_{i_{k+1}} \cdots$. By the choice of M , as $|s_{i_1} \cdots s_{i_{k+1}-1}| = |r_1 \cdots r_k| \geq K$, at least one of the words s_j is a long return word to $\text{LS}_n(\mathbf{w})$. Let k' be the number of integers i between i_1 and $i_{k+1} - 1$ such that $s_i \in \mathcal{R}_n^{(L)}(\mathbf{w})$ and let $j_1 < j_2 < \dots < j_{k'}$ be these integers. Let ℓ'_j be the prefix of length n of $s_{j+1} s_{j+2} \cdots$. We have just described a way to associate to each element $((r_1, \ell_1), \dots, (r_k, \ell_k))$ of $\mathcal{U}_{n+1}(\mathbf{w})$ a sequence $((s_{i_1}, \ell'_{i_1}), \dots, (s_{j_1}, \ell'_{j_1}))$, $((s_{j_1+1}, \ell'_{j_1+1}), \dots, (s_{j_2}, \ell'_{j_2}))$, \dots , $((s_{j_{k'}-2+1}, \ell'_{j_{k'}-2+1}), \dots, (s_{j_{k'}-1}, \ell'_{j_{k'}-1}))$, $((s_{j_{k'}-1+1}, \ell'_{j_{k'}-1+1}), \dots, (s_{i_k}, \ell'_{i_k}))$ of elements of $\mathcal{U}_n(\mathbf{w})$.

For $n \geq M$, let $\mathcal{A}_n = \{0, 1, \dots, \text{Card}(\mathcal{U}_n(\mathbf{w})) - 1\}$ and let Θ_n be a bijection from \mathcal{A}_n to $\mathcal{U}_n(\mathbf{w})$. Now we define our morphisms from \mathcal{A}_{n+1}^* to \mathcal{A}_n^* by $\sigma_n(a) = a_1 \dots a_k$ whenever $\Theta_n(a)$ can be associated as previously to $\Theta_n(a_1) \dots \Theta_n(a_k)$.

We also let τ be the morphism from \mathcal{A}_n^* to A^* (the alphabet of \mathbf{w}) defined for all $a \in \mathcal{A}_n$ by $\tau(a) = u$ if $\Theta_n(a) = ((r_1, \ell_1), \dots, (r_k, \ell_k))$ and $u = r_1 \cdots r_k$.

Of course our morphisms depend on our construction. Other morphisms can be suitable.

4.7. The number of morphisms is finite

We continue with the notation introduced in the previous section. Let $S = \{\sigma_n \mid n \geq M\}$.

Fact 4.10. *The set S is finite.*

Proof. We start with a technical observation. Let (r, l) in $\mathcal{R}_{n+1}(\mathbf{w}) \times \text{LS}_{n+1}(\mathbf{w})$ such that rl is a factor of \mathbf{w} . Following Lemma 3.2, there exist $(r_1, l_1), (r_2, l_2), \dots, (r_k, l_k)$, elements of $\mathcal{R}_n(\mathbf{w}) \times \text{LS}_n(\mathbf{w})$, such that $rl_{[1,n]} = r_1 r_2 \dots r_k l_k$ (where $l_k = l_{[1,n]}$ denotes the prefix of length n of l) and l_i is a prefix of $r_{i+1} \dots r_k l_k$ for all i in $\{1, \dots, k\}$. We are going to show that, for i, j in $\{2, \dots, k-1\}$, $i \neq j$, we have $(r_i, l_i) \neq (r_j, l_j)$ and, consequently, $k < 2 + \text{Card}(\mathcal{R}_n(\mathbf{w})) \times \text{Card}(\text{LS}_n(\mathbf{w}))$.

Let $l' \in \text{LS}_{n+1}(\mathbf{w})$ be the prefix of rl of length $n+1$. Assume by contradiction that $(r_i, l_i) = (r_j, l_j)$ with $1 < i < j < k$. Since there are no left special factors of length $n+1$ in rl (except its prefix l' and its suffix l), for all integers $h \in \{1, \dots, i\}$, the words $r_h \dots r_i l_i$ can be uniquely extended to the left in $L(\mathbf{w})$ by r_{h-1} . As $r_i l_i = r_j l_j$, this implies that $r_1 \dots r_i l_i$ is also a suffix of $r_1 \dots r_j l_j$. Thus l' has an internal occurrence in rl , which is a contradiction.

We now turn to the proof of the fact. We need to go back to the construction of the morphisms and so we let $((r_1, \ell_1), \dots, (r_k, \ell_k))$ and integers $(i_j)_{1 \leq j \leq k+1}$ be as in this construction. The previous observation shows that for all j , we have $1 \leq j \leq k$, $i_{j+1} - i_j \leq 2 + \text{Card}(\mathcal{R}_n(\mathbf{w})) \times \text{Card}(\text{LS}_n(\mathbf{w}))$, and so $i_{k+1} - i_1 \leq k(2 + \text{Card}(\mathcal{R}_n(\mathbf{w})) \times \text{Card}(\text{LS}_n(\mathbf{w})))$. By Corollary 4.2, Proposition 4.3 and Lemma 4.9, there exists an integer K_1 such that for all $n \geq M$, $\text{Card}(\mathcal{R}_n(\mathbf{w})) \leq K_1$, $\text{Card}(\text{LS}_n(\mathbf{w})) \leq K_1$ and $k \leq K_1$. As the length of $\sigma_n(\Theta_{n+1}^{-1}((r_1, \ell_1), \dots, (r_k, \ell_k)))$ is bounded by $i_{k+1} - i_1$, it is also bounded above by $K_1(2 + K_1^2)$.

In other words, there exists an integer K , such that for all $n \geq M$ and for all letters $a \in \mathcal{A}_{n+1}$, $|\sigma_n(a)| \leq K$. As also $\text{Card}(\mathcal{A}_n) \leq K$, the set S is finite. \square

4.8. Everywhere growing property

We continue with the notation introduced in the two previous sections.

Fact 4.11. *For any sequence $(a_n \in \mathcal{A}_n)_{n \geq M}$, we have*

$$\lim_{n \rightarrow \infty} |\tau \sigma_M \dots \sigma_n(a_{n+1})| = +\infty$$

Proof. By construction, for any $n \geq M$, $a_{n+1} = \Theta_{n+1}^{-1}((r_1, \ell_1), \dots, (r_k, \ell_k))$ for some $(r_1, \ell_1), \dots, (r_k, \ell_k)$ in $\mathcal{U}_{n+1}(\mathbf{w})$ and it can be verified by induction that $|\tau \sigma_M \dots \sigma_n(a_{n+1})| = |r_1 \dots r_k|$. By Proposition 3.5, the length of minimal elements of $\mathcal{R}_n^{(L)}(\mathbf{w})$ grows to infinity. So the length of minimal elements of $\mathcal{U}_n(\mathbf{w})$ also grows to infinity. The lemma follows. \square

4.9. Proof of Theorem 1.1

Let $n \geq 0$. Let $(r_i)_{i \geq 0}$ be the coherent decomposition of \mathbf{w} over $\mathcal{R}_n(\mathbf{w})$ of level n , and for all $i \geq 0$, let ℓ_i be the prefix of $r_i \cdots$. For $n \geq M$, we know that at least one of the elements of $\mathcal{R}_n(\mathbf{w})$ belongs to $\mathcal{R}_n^{(L)}(\mathbf{w})$. Let j be the smallest integer such that $r_j \in \mathcal{R}_n^{(L)}(\mathbf{w})$ and let $a_n = \Theta_n^{-1}((r_0, \ell_0), \dots, (r_j, \ell_j))$. By definition, for $n \geq M$, $\tau\sigma_M \cdots \sigma_n(a_{n+1})$ is a prefix of \mathbf{w} .

By Fact 4.11, it follows that

$$\mathbf{w} = \lim_{n \rightarrow +\infty} \tau\sigma_M \cdots \sigma_n(a_{n+1}^\omega). \quad (2)$$

This, with Fact 4.11, ends the proof of Theorem 1.1.

Acknowledgement

The authors would like to thank Fabien Durand for many fruitful discussions concerning return words and S -adicity. Many thanks also to Narad Rampersad for his careful reading and to the referees for their valuable comments.

References

- [1] J.-P. Allouche and J. Shallit. *Automatic sequences: Theory, applications, generalizations*. Cambridge University Press, Cambridge, 2003.
- [2] P. Arnoux and G. Rauzy. Représentation géométrique de suites de complexité $2n + 1$. *Bull. Soc. Math. France*, 119:199–215, 1991.
- [3] J. Berstel, A. Lauve, C. Reutenauer, and F. Saliola. *Combinatorics on Words: Christoffel Words and Repetitions in Words*, volume 27 of *CRM Monograph Series*. American Mathematical Society, 2008.
- [4] V. Berthé and M. Rigo, editors. *Combinatorics, Automata and Number Theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010.
- [5] M. Boshernitzan. A condition for minimal interval exchange maps to be uniquely ergodic. *Duke Math. J.*, 52(3):723–752, 1985.
- [6] J. Cassaigne. Special factors of sequences with linear subword complexity. In *Developments in language theory, II (Magdeburg, 1995)*, pages 25–34. World Sci. Publ., River Edge, NJ, 1996.

- [7] J. Cassaigne and F. Nicolas. Factor complexity. Chapter 3 in [4].
- [8] A. de Luca. Sturmian words: structure, combinatorics, and their arithmetics. *Theoretical Computer Science*, 183:45–82, 1997.
- [9] F. Durand. A characterization of substitutive sequences using return words. *Discrete Math.*, 179:89–101, 1998.
- [10] F. Durand. Linearly recurrent subshifts have a finite number of non-periodic subshift factors. *Ergodic Theory Dynam. Systems*, 20:1061–1078, 2000.
- [11] F. Durand, J. Leroy, and G. Richomme. Do properties of an S -adic representation determine factor complexity? To appear in *J. Integer Seq.* ArXiv e-prints:1208.6376v1.
- [12] S. Ferenczi. Rank and symbolic complexity. *Ergodic Theory Dynam. Systems*, 16:663–682, 1996.
- [13] J. Leroy. Some improvements of the S -adic conjecture. *Advances in Applied Mathematics*, 48(1):79 – 98, 2012.
- [14] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, UK, 1997.
- [15] M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2002.
- [16] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.*, 9:289–298, 1962.
- [17] F. Mignosi and P. Séébold. Morphismes sturmiens et règles de Rauzy. *J. Théor. Nombres Bordeaux*, 5:221–233, 1993.
- [18] M. Morse and G. A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *Amer. J. Math.*, 62:1–42, 1940.
- [19] J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés. In *Automata, languages and programming (Antwerp, 1984)*, volume 172 of *Lecture Notes in Comput. Sci.*, pages 380–389. Springer, Berlin, 1984.
- [20] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.

- [21] N. Pytheas Fogg, Terminologie S -adique et propriétés, Preprint available at <http://tinyurl.com/8opdb8s>, 2011.
- [22] M. Queffélec. *Substitution dynamical systems—spectral analysis*, volume 1294 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1987.
- [23] G. Rauzy. Suites à termes dans un alphabet fini. *Séminaire de Théorie des nombres de Bordeaux*, exposé 25:1–16, 1982-1983.

5. Appendix: proof of Lemma 2.5

This entire section is dedicated to the proof of Lemma 2.5. Two words x and y are said to be *prefix-comparable* if x is a prefix of y or y is a prefix of x .

Fact 5.1. *For any morphism f and any prefix-comparable words x and y , $f(x)$ and $f(y)$ are also prefix-comparable.*

Fact 5.2. *If two words x and y are prefix-comparable and z is a prefix of y , then x and z are also prefix-comparable.*

In what follows, let us recall that \mathbf{u} , introduced in Section 2, is the fixed point of ν starting with the letter c .

Lemma 5.3. *Any factor of \mathbf{u} of length at least two and starting with b_1 is prefix-comparable to one of the words $b_1\nu^\ell(1)e_1c$, $b_1\nu^\ell(1)c$, $b_1\nu^\ell(1)e_01$, $b_1\nu^\ell(2)e_2$ and $b_1\nu^\ell(2)e_3$ for some $\ell \geq 0$.*

Proof. We proceed by induction. For factors of length two, three or four, the result is true with $\ell \in \{0, 1\}$ since

$$\begin{aligned}
 b_1\Gamma \cap F(\mathbf{u}) &= \{b_12, b_1c\}, \\
 b_1\Gamma^2 \cap F(\mathbf{u}) &= \{b_11e_1, b_1cc, b_11c, b_11e_0, b_12e_2, b_12e_3\}, \\
 b_1\Gamma^3 \cap F(\mathbf{u}) &\subseteq \{b_11c, b_12e_2, b_12e_3\}\Gamma \cup \{b_11e_1c, b_11e_01, b_1ccc\}.
 \end{aligned}$$

Let b_1x be a factor of length at least five of \mathbf{u} . The result is immediately true if x starts with $1c$, $2e_2$, $2e_3$, $1e_1c$ or $1e_01$, (with $\ell = 0$). Thus, assume that x starts with ccc . By the definition of ν , this implies the existence of a factor y of \mathbf{u} such that b_1y is a factor of \mathbf{u} , x is a prefix of $\nu(y)$ and $|y| < |x|$. By the inductive hypothesis, b_1y is prefix-comparable to one of the words $b_1\nu^\ell(1)e_1c$, $b_1\nu^\ell(1)c$, $b_1\nu^\ell(1)e_01$, $b_1\nu^\ell(2)e_2$ or $b_1\nu^\ell(2)e_3$ for some $\ell \geq 0$. So, by Fact 5.1 and Fact 5.2, b_1x is prefix-comparable to one of the words $b_1\nu^{\ell+1}(1)e_1c$, $b_1\nu^{\ell+1}(1)c$, $b_1\nu^{\ell+1}(1)e_01$, $b_1\nu^{\ell+1}(2)e_2$ or $b_1\nu^{\ell+1}(2)e_3$. □

Corollary 5.4. *Any factor of \mathbf{u} of length at least three and starting with a_1b_1 is prefix-comparable to one of the words $a_1b_1\nu^\ell(1)e_1c$, $a_1b_1\nu^\ell(1)c$, $a_1b_1\nu^\ell(1)e_01$, $a_1b_1\nu^\ell(2)e_2$ or $a_1b_1\nu^\ell(2)e_3$ for some $\ell \geq 1$.*

Proof. After Lemma 5.3, it is sufficient to observe that the words $a_1b_11e_1c$, a_1b_11c , $a_1b_11e_0$, $a_1b_12e_2$ and $a_1b_12e_3$ are not factors of words in $\nu(\Gamma^*)$. \square

Lemma 5.5. *Any factor a_2b_2m of \mathbf{u} of length at least three is prefix-comparable to $a_2b_2\nu^\ell(2)e_2$ for some $\ell \geq 1$.*

Proof. The proof is similar to that of Corollary 5.4. \square

Proof of Lemma 2.5. Let $n \geq 2$ be an integer such that $ab \in \mathcal{R}_n(\mathbf{u})$. This means that there exists a word v and two letters x, y such that $|v| = n - 2$, $abvxy$ is a factor of $\psi(\mathbf{u})$, both words abv and vxy are left special factors of $\psi(\mathbf{u})$ and bvx is not a left special factor of $\psi(\mathbf{u})$. We aim to prove that, for some integer k , $v = u_k$ (and so $n = 6^k + 1$), $x = 1$ and $y = e_1$ (see before Fact 2.2 for the definition of words u_k).

By construction of $\psi(\mathbf{u})$, two cases can hold (maybe not exclusively).

- Case I: There exists a word w_1 and two letters x_1, y_1 such that $abvxy = \psi(a_1b_1w_1x_1y_1)$ and $a_1b_1w_1x_1y_1$ is a factor of \mathbf{u} .
- Case II: There exists a word w_2 and two letters x_2, y_2 such that $abvxy = \psi(a_2b_2w_2x_2y_2)$ and $a_2b_2w_2x_2y_2$ is a factor of \mathbf{u} .

First, we prove that Case II is impossible. By Lemma 5.5, the word $a_2b_2w_2x_2y_2$ is prefix-comparable to $a_2b_2\nu^k(2)e_2$ for an integer k . If $|a_2b_2w_2x_2y_2| \leq |a_2b_2\nu^k(2)e_2|$, then $b_2w_2x_2$ is a prefix of $b_2\nu^k2$. Thus, the non-left special word $bvx = \psi(bw_2x_2)$ is a prefix of bu_k2 . This contradicts the fact that bu_k2 is a left special factor of $\psi(\mathbf{u})$ (the words $abu_k2 = \psi(a_2b_2\nu^k2)$ and $e_3bu_k2 = \psi(e_3b_3\nu^k2)$ are factors of $\psi(\mathbf{u})$). Hence $|a_2b_2w_2x_2y_2| > |a_2b_2\nu^k(2)e_2|$ and $\nu^k(2)e_2$ is a prefix of $w_2x_2y_2$. Consequently, the non-left special factor u_k2e_2 (see Lemma 2.3) is a prefix of the left special factor $vxy = \psi(w_2x_2y_2)$, which is a contradiction.

Therefore, Case I holds. By Corollary 5.4, the word $a_1b_1w_1x_1y_1$ is prefix-comparable to one of the words $a_1b_1\nu^k(1)e_1c$, $a_1b_1\nu^k(1)c$, $a_1b_1\nu^k(1)e_01$, $a_1b_1\nu^k(2)e_2$ and $a_1b_1\nu^k(2)e_3$ for some $k \geq 1$.

If $|a_1b_1w_1x_1y_1| \leq |a_1b_1\nu^k(1)|$, then b_1w_1x is a prefix of $b_1\nu^k$. Thus the non-left special factor $bvx = \psi(b_1w_1x)$ of $\psi(\mathbf{u})$ is a prefix of the left special factor bu_k , which is a contradiction.

If $|a_1b_1w_1x_1y_1| \geq |a_1b_1\nu^k(1)| + 2$, then one of the words $a_1b_1\nu^k(1)e_1c$, $a_1b_1\nu^k(1)c$, $a_1b_1\nu^k(1)e_01$, $a_1b_1\nu^k(2)e_2$ or $a_1b_1\nu^k(2)e_3$ is a prefix of $a_1b_1w_1x_1y_1$. Thus u_k1e_1c , u_k1c , u_k1e_11 , u_k2e_2 or u_k2e_3 is a prefix of the left special factor vxy of $\psi(\mathbf{u})$. This contradicts Lemma 2.3, which states that these five words are not left special.

Thus, $|a_1 b_1 w_1 x_1 y_1| = |a_1 b_1 \nu^k(1)| + 1$. Consequently, we have $w_1 = v_k$ and $x_1 y_1 \in \{1e_1, 1c, 1e_0, 2e_2, 2e_3\}$, that is, $xy \in \{1e_1, 1c, 2e_2, 2e_3\}$. Noting that $vx_1 y_1$ is left special whereas by Lemma 2.3, the words $u_k 1c$, $u_k 2e_2$ and $u_k 2e_3$ are not left special, we deduce that $xy = 1e_1$. This ends the proof of Lemma 2.5. \square