

String attractors of some simple-Parry automatic sequences

France Gheeraert^[0000–0002–8970–173X]^{1*},
Giuseppe Romana^[0000–0002–3489–0684]² and
Manon Stipulanti^[0000–0002–2805–2465]³

^{1*}Department of Mathematics, Radboud University, Heyendaalseweg
135, Nijmegen, 6525AJ, Netherlands.

²Dipartimento di Matematica e Informatica, Università di Palermo, Via
Archirafi 34, Palermo, 90123, Italy.

³Department of Mathematics, University of Liège, Allée de la
Découverte 12, Liège, 4000, Belgium.

*Corresponding author(s). E-mail(s): france.gheeraert@ru.nl;
Contributing authors: giuseppe.romana01@unipa.it;
m.stipulanti@uliege.be;

Abstract

Firstly studied by Kempa and Prezza in 2018 as the cement of text compression algorithms, string attractors have become a compelling object of theoretical research within the community of combinatorics on words. In this context, they have been studied for several families of finite and infinite words. In this paper, we obtain string attractors of prefixes of particular infinite words generalizing k -bonacci words (including the famous Fibonacci word) and related to simple Parry numbers. In fact, our description involves the numeration systems classically derived from the considered morphisms. This extends our previous work published in the international conference WORDS 2023.

Keywords: String attractors, Numeration systems, Parry numbers, Automatic sequences, Morphic sequences, Fibonacci word

MSC Classification: Primary: 68R15. Secondary: 05A05 , 11A67 , 68P05 , 68Q45.

1 Introduction

Introduced in the field of data compression by Kempa and Prezza [1] in 2018, the concept of *string attractor* can be conceptualized as follows: within a finite word, it is a set of positions that enables to catch all distinct factors. Since then, questions related to string attractors have drawn the attention of many researchers from various scientific fields. From the point of view of the theory of algorithmic complexity, the problem of finding a smallest string attractor is NP-hard [1]. In parallel, string attractors also have applications in combinatorial pattern matching [2]. In order to understand the best way to measure data compressibility by exploiting repetitiveness in strings, measures have recently been introduced in relation to string attractors [3].

Quickly after, combinatorics-on-words researchers have quite naturally seized the notion and made a systematic topic of research out of it. String attractors have been studied in the context of some (families of) sequences: automatic sequences [4] with some focus on the ubiquitous Thue–Morse word [4–6] and the period-doubling word [4], the famous Sturmian words [7–9] and their extension known as episturmian words [9], the Tribonacci word [4] and more generally the k -bonacci words [10], some binary generalized pseudostandard sequences [11], and bi-infinite words [12]. Besides analyzing important families of words, another classical topic in combinatorics on words is the study of complexity functions for infinite words, such as the distinguished factor complexity function and the so-called abelian [13] and binomial [14] complexity functions. A first complexity function based on string attractors was introduced in [4] and considered for automatic sequences and linearly recurrent infinite words. In addition to further studying such a complexity function, the authors of [8, 10] introduce and examine two other string attractor-based complexity functions.

Historically, the bond between string attractors and numeration systems was observed for the first time in [10]. There, the authors consider generalizations of the Fibonacci word to larger alphabets (on k letters, the corresponding word is called the k -bonacci word) and show that some of their string attractors rely on the well-known k -bonacci numbers. These infinite words are *morphic*, i.e., they are obtained as images, under codings, of fixed points of morphisms. Consequently, we currently believe that the link between string attractors and numeration systems can be adapted to other morphic sequences and we therefore raise the following general question:

Question. Given a morphic sequence \mathbf{z} , does there exist a numeration system \mathcal{S} such that \mathbf{z} is \mathcal{S} -automatic and (minimal) string attractors of the prefixes of \mathbf{z} are easily described using \mathcal{S} ?

In this paper, to support this question, we study a particular family of morphic words. More precisely, given parameters in the shape of a length- k word $c = c_0 \cdots c_{k-1} \in \mathbb{N}^k$, we define the morphism μ_c such that $\mu_c(i) = 0^{c_i} \cdot (i+1)$ for all $i \in \{0, \dots, k-2\}$ and $\mu_c(k-1) = 0^{c_{k-1}}$. When it exists, we then look at the fixed point of μ_c . From a combinatorics-on-words point of view, this family is interesting as it generalizes the k -bonacci morphisms. However, their main interest stems from numeration systems. Given a simple Parry number β (a real number for which the β -expansion of 1 is of the form $u0^\omega$ for some finite word u), Fabre [15] associated with it a morphism μ which is strongly related to the automaton corresponding to the base- β

numeration system (for real numbers), or alternatively to the greedy numeration system associated with the sequence $(|\mu^n(0)|)_{n \in \mathbb{N}}$ (for integers) [16]. This morphism can also be used to understand the base- β integers [15, 17]. It turns out that μ corresponds to μ_c where c is the β -representation of 1. Note that, due to a result of Parry [18], not every c corresponds to a simple Parry number. Combinatorially speaking, the fixed points of these particular morphisms have been studied for instance in [19–22].

Furthermore, the techniques used so far to obtain string attractors of infinite words do not apply to the words we consider. Indeed, on the one hand, they are not necessarily episturmian, so we cannot use the approach from [9]. On the other hand, for some parameter $c \in \mathbb{N}^k$, the corresponding numeration system is not *addable*, meaning that the addition within the numeration system is not recognizable by a finite automaton. For example, this is the case of $c = 3203$ [23]. As a consequence, we cannot follow the methods from [4]; in particular, we study words outside the framework needed to use the software `Walnut` [24].

Under some conditions on the parameters, we show that the prefixes of the fixed point admit string attractors strongly related to the associated numeration system. Moreover, they are of size at most one more than the alphabet size, and are therefore nearly optimal. Finally, we provide an infinite family of words for which these string attractors have minimal size.

This work extends and completes some of the results presented in our first exploratory paper on this topic [25]. In this version, we emphasize the link with numeration systems and simple Parry numbers. We also provide new results, and as a consequence, have restructured and simplified most of the proofs.

This paper is organized as follows. In Section 2, we first recall some classical notions of combinatorics on words. In particular, we introduce the family of words that we will focus on. We then show how these words are related to numeration systems in Section 3, and we study some of their properties such as greediness. Section 4 is devoted to the construction of string attractors of the words. More precisely, we give the precise conditions under which every prefix admits a string attractor included in a particular subset, and we exhibit such a string attractor when the conditions are met. Then in Section 5, we analyze these conditions and give alternative formulations in terms of the numeration system and of the parameters c_0, \dots, c_{k-1} . Finally, in Section 6, we discuss the optimality of the string attractors that we obtained and describe an infinite family of words for which they are optimal.

2 Preliminaries

2.1 Words

We start with the bare minimum on words and introduce some notations.

Let A be an alphabet either finite or infinite (for instance, we will consider words over the set of non-negative integers \mathbb{N}). The length of a word is its number of letters and will be denoted with vertical bars $|\cdot|$. We let ε denote the empty word, and A^* denote the set of finite words over A . For any integer $n \geq 0$, we let A^n be the set of length- n words over A . If $w = xyz$ for some $x, y, z \in A^*$, then x is a *prefix*, y is a *factor*, and z is a *suffix* of w . A factor of a word is *proper* if it is not equal to the

initial word. A word v is a *fractional power* of a non-empty word w if there exist $\ell \geq 0$ and x a prefix of w such that $v = w^\ell x$. Infinite words are written in bold and we start indexing them at 0. We use classical notations of intervals to denote portions of words. For a non-empty word $u \in A^*$, we let u^ω denote the concatenation of infinitely many copies of u , that is, $u^\omega = uu\cdots$.

Let \leq be a total order on A . The *lexicographic order* on A^* induced by \leq is defined as follows: for $x, y \in A^*$, we say that x is *lexicographically smaller than* y , and we write $x < y$, if either x is a proper prefix of y , or $x = zax'$ and $y = zby'$ for some letters a, b with $a < b$. We write $x \leq y$ if x is lexicographically smaller than or equal to y . The *genealogical order*, also known as *radix order*, on A^* induced by \leq is defined as follows: for $x, y \in A^*$, we say that x is *genealogically smaller than* y , and we write $x <_{\text{gen}} y$, if either $|x| < |y|$, or $|x| = |y|$ and $x = zax'$ and $y = zby'$ for some letters a, b with $a < b$. We write again $x \leq_{\text{gen}} y$ if x is genealogically smaller than or equal to y .

A non-empty word $w \in A^*$ is *primitive* if $w = u^n$ for $u \in A^* \setminus \{\varepsilon\}$ implies $n = 1$. Two words are *conjugates* if they are cyclic permutation of each other.

A word is *Lyndon* if it is primitive and lexicographically minimal among its conjugates for some given order. Defined in the 50's, Lyndon words are not only classical in combinatorics on words but also of utmost importance. See [26] for a presentation. A celebrated result in combinatorics on words is that Lyndon words form a so-called *complete factorization of the free monoid*.

Theorem 1 (Chen-Fox-Lyndon [27]). *For every non-empty word $w \in A^*$, there exists a unique factorization (ℓ_1, \dots, ℓ_n) of w into Lyndon words over A such that $\ell_1 \geq \ell_2 \geq \dots \geq \ell_n$.*

Several variations of Lyndon words have been considered lately: generalized Lyndon [28], anti-Lyndon [29], inverse Lyndon [30], and Nyldon [31]. In this text, we will use the second.

Definition 1. *Let (A, \leq) be a totally ordered alphabet. We let \leq_- denote the inverse order on A , i.e., $b <_- a$ if and only if $a < b$ for all $a, b \in A$. We also let \leq_- denote the inverse lexicographic order which is the lexicographic order induced by \leq_- . A word is anti-Lyndon if it is Lyndon with respect to the inverse lexicographic order.*

Otherwise stated, a word is anti-Lyndon if it is primitive and lexicographically maximal among its conjugates.

Example 2. *Let $A = \{0, 1\}$ with $0 < 1$, so $1 <_- 0$. The first few anti-Lyndon words, ordered by length, are 1, 0, 10, 110, 100, 1110, 1100, and 1000.*

2.2 Morphisms and fixed points of interest

A *morphism* is a map $f: A^* \rightarrow B^*$, where A, B are alphabets, such that $f(xy) = f(x)f(y)$ for all $x, y \in A^*$. The morphism f is *prolongable* on the letter $a \in A$ if $f(a) = ax$ for some $x \in A^*$ and $f^n(x) \neq \varepsilon$ for all $n \geq 0$. In this section, we consider a specific family of morphisms defined as follows. Note that they appear under the name *generic k -bonacci morphisms* in [32, Example 2.11].

Definition 2. *Let $k \geq 2$ be an integer and let $c_0, \dots, c_{k-1} \geq 0$ be k parameters often summarized in the shape of a word $c = c_0 \cdots c_{k-1} \geq 0^k$. The morphism $\mu_c: \{0, \dots, k-1\}^* \rightarrow \{0, \dots, k-1\}^*$ is given by $\mu_c(i) = 0^{c_i} \cdot (i+1)$ for all $i \in \{0, \dots, k-2\}$ and $\mu_c(k-1) = 0^{c_{k-1}}$. For all $n \geq 0$, we then define $u_{c,n} = \mu_c^n(0)$ and $U_{c,n} = |u_{c,n}|$.*

Table 1 Construction of the sequences $(u_n)_{n \geq 0}$ and $(U_n)_{n \geq 0}$ for $c = 102$.

n	0	1	2	3	4	5
u_n	0	01	012	01200	012000101	012000101012012
fact. of u_n	0	$u_0^1 \cdot 1$	$u_1^1 u_0^0 \cdot 2$	$u_2^1 u_1^0 u_0^2$	$u_3^1 u_2^0 u_1^2$	$u_4^1 u_3^0 u_2^2$
U_n	1	2	3	5	9	15

When the context is clear, we will usually omit the subscript c in Definition 2.

Example 3. When $c = 1^k$, we recover the k -bonacci morphism and words. For $k = 3$ and $c = 102$, the first few iterations of the corresponding morphism $\mu_c: 0 \mapsto 01, 1 \mapsto 2, 2 \mapsto 00$ are given in Table 1. Some specific factorization of the words $(u_{c,n})_{n \geq 0}$ is highlighted in Table 1.

The factorization presented in the previous example can be stated in general. It gives a recursive definition of the words $(u_{c,n})_{n \geq 0}$ and can be proven using a simple induction.

Proposition 4. For all $c = c_0 \cdots c_{k-1} \geq 0^k$, we have

$$u_n = \begin{cases} \left(\prod_{i=0}^{n-1} u_{n-i-1}^{c_i} \right) \cdot n, & \text{if } 0 \leq n \leq k-1; \\ \prod_{i=0}^{k-1} u_{n-i-1}^{c_i}, & \text{if } n \geq k. \end{cases}$$

As a consequence of Proposition 4, the sequence $(U_n)_{n \geq 0}$ respects the following recurrence relation:

$$U_n = \begin{cases} 1 + \sum_{i=0}^{n-1} c_i U_{n-i-1}, & \text{if } 0 \leq n \leq k-1; \\ \sum_{i=0}^{k-1} c_i U_{n-i-1}, & \text{if } n \geq k. \end{cases} \quad (1)$$

In the rest of the paper, we will assume the following working hypothesis (WH) on c :

$$c = c_0 \cdots c_{k-1} \geq 0^k \text{ with } c_0, c_{k-1} \geq 1. \quad (\text{WH})$$

The condition $c_{k-1} \geq 1$ ensures both that the recurrence relation is of order k and that the morphism μ_c is non-erasing, which is a classical assumption in combinatorics on words. Moreover, the condition $c_0 \geq 1$ guarantees that μ_c is prolongable. Under (WH), the morphism μ_c has an infinite fixed point starting with 0 denoted $\mathbf{u} := \lim_{n \rightarrow \infty} u_n$.

We make the following combinatorial observation.

Remark 5. Under (WH), using Proposition 4, a simple induction shows that the letter $i \in \{1, \dots, k-1\}$ can only be followed by 0 and/or $i+1$ (and only 0 in the case $i = k-1$) in \mathbf{u} and in u_n^ω .

3 Link with numeration systems

In this section, specific definitions will be recalled. For the reader unfamiliar with the theory of numeration systems, we refer to [33, Chapters 2 and 3] for an introduction and some advanced concepts.

Table 2 Illustration of the numeration system \mathcal{S}_c for $c = 102$.

n	0	1	2	3	4	5	6	7	8
$\mathbf{u}[0, n]$	ε	0	01	012	012 · 0	01200	01200 · 0	01200 · 01	01200 · 01 · 0
$\text{rep}_{\mathcal{S}_c}(n)$	ε	1	10	100	101	1000	1001	1010	1011

A *numeration system* (for natural numbers) can be defined as a triple $\mathcal{S} = (A, \text{rep}_{\mathcal{S}}, L)$, where A is an alphabet and $\text{rep}_{\mathcal{S}}: \mathbb{N} \rightarrow A^*$ is an injective function such that $L = \text{rep}_{\mathcal{S}}(\mathbb{N})$. The map $\text{rep}_{\mathcal{S}}$ is called the *representation function* and L is the *numeration language*. If $\text{rep}_{\mathcal{S}}(n) = w$ for some integer $n \geq 0$ and some word $w \in A^*$, we say that w is the *representation (in \mathcal{S})* of n and we define the *valuation (in \mathcal{S})* of w by $\text{val}_{\mathcal{S}}(w) = n$. Note that, when the context is clear, we omit the subscript \mathcal{S} in rep and val .

Any given prolongable morphism naturally gives rise to a numeration system that we will call the *associated Dumont-Thomas numeration system* [34]. These are based on particular factorizations of the prefixes of the fixed point. We only give here the definition in the particular case of the morphisms studied in this paper but the interested reader can find the general case in the original paper [34].

Proposition 6 (Dumont-Thomas [34]). *Let c satisfy (WH). For all $n \geq 1$, there exist unique integers $N, \ell_0, \dots, \ell_N \geq 0$ such that $\ell_0 \geq 1$, $\mathbf{u}[0, n] = u_N^{\ell_0} \cdots u_0^{\ell_N}$, and this factorization verifies the following: u_{N+1} is not a prefix of $\mathbf{u}[0, n]$ and, for all $0 \leq i \leq N$, $u_N^{\ell_0} \cdots u_{N-i+1}^{\ell_{i-1}} u_{N-i}^{\ell_i+1}$ is not a prefix of $\mathbf{u}[0, n]$.*

Recall that a numeration system based on a suitable sequence of integers $(U_n)_{n \geq 0}$ is called *greedy* when, at each step of the decomposition of any integer, the largest possible term of the sequence $(U_n)_{n \geq 0}$ is chosen; formally, we use the Euclidean algorithm in a greedy way. As the conditions on the factorization in Proposition 6 resemble that of greedy representations in numeration systems, we will refer to it as being *word-greedy*.

For a given c satisfying (WH), we then let \mathcal{S}_c denote the numeration system associated with the representation function $\text{rep}_{\mathcal{S}_c}: \mathbb{N} \rightarrow \mathbb{N}^*$ mapping n to $\text{rep}_{\mathcal{S}_c}(n) = \ell_0 \cdots \ell_N$, where the integers ℓ_0, \dots, ℓ_N verify the conditions of Proposition 6 for n . By convention, we set $\text{rep}_{\mathcal{S}_c}(0) = \varepsilon$.

Example 7. *Using Example 3 for $c = 102$, the representations of the first few integers are given in Table 2. The word-greedy factorization of each prefix is highlighted in the second row, leading to the representation of the corresponding integer in the third row.*

Remark 8. *If $\text{rep}_{\mathcal{S}_c}(n) = \ell_0 \cdots \ell_N$, then $n = |u_N^{\ell_0} \cdots u_0^{\ell_N}| = \sum_{i=0}^N \ell_i U_{N-i}$. In other words, $\text{val}_{\mathcal{S}_c}$ is given by the usual valuation function associated with the sequence $(U_n)_{n \geq 0}$. Such a system is sometimes called a positional numeration system. Note that this is not necessarily the case for the Dumont-Thomas numeration system associated with some other morphism.*

The Dumont-Thomas numeration systems are a particular case of abstract numeration systems introduced in [35]. A numeration system $\mathcal{S} = (A, \text{rep}, L)$ is said to be *abstract* if L is regular and $\text{rep}(n)$ is the $(n+1)$ st word of L in the genealogical order. We have the following result.

Theorem 9 (Rigo [32, Section 2.2]). *Let $\sigma: \{\alpha_0, \dots, \alpha_d\}^* \rightarrow \{\alpha_0, \dots, \alpha_d\}^*$ be a morphism prolongable on the letter α_0 . We define the automaton \mathcal{A}_σ for which $\{\alpha_0, \dots, \alpha_d\}$*

is the set of states, α_0 is the initial state, every state is final, and the (partial) transition function δ is such that, for each $\alpha \in \{\alpha_0, \dots, \alpha_d\}$ and $0 \leq i \leq |\sigma(\alpha)| - 1$, $\delta(\alpha, i)$ is the $(i + 1)$ st letter of $\sigma(\alpha)$. If $\mathcal{S} = (A, \text{rep}, L)$ is the Dumont-Thomas numeration system associated with σ , then $L = L(\mathcal{A}_\sigma) \setminus 0\mathbb{N}^*$ and $\text{rep}(n)$ is the $(n + 1)$ st word of L in the genealogical order.

Example 10. For $c = 102$, the automaton \mathcal{A}_{μ_c} of Theorem 9 is depicted in Figure 1 (details are left to the reader). The first few accepted words (not starting with 0) are, in genealogical order, ε , 1, 10, 100, 101, 1000, 1001, 1010, and 1011, which indeed agree with the representations of the first few integers in Example 7.

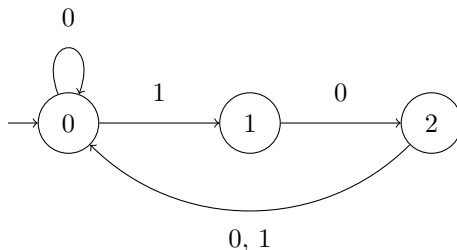


Fig. 1 The automaton \mathcal{A}_{μ_c} for $c = 102$.

As the automaton in Theorem 9 can be used to produce, for all $n \geq 0$, the letter \mathbf{u}_n when reading $\text{rep}_{\mathcal{S}_c}(n)$ by [32, Theorem 2.24], we have the following.

Corollary 11. Let c satisfy (WH). Then the sequence \mathbf{u} is \mathcal{S}_c -automatic.

Similarly to what is usually done in real base numeration systems, we will let \mathbf{d}^* denote the periodization of c , that is, $\mathbf{d}^* = (c_0 \cdots c_{k-2}(c_{k-1} - 1))^\omega$. Using Theorem 9, we deduce the next result.

Lemma 12. Under (WH), for all $n \geq 0$, we have $\text{rep}_{\mathcal{S}_c}(U_n) = 10^n$, the numbers having a representation of length $n + 1$ are those in $[U_n, U_{n+1})$, and $\text{rep}_{\mathcal{S}_c}(U_{n+1} - 1) = \mathbf{d}^*[0, n]$. In particular, $U_{n+1} - 1 = \sum_{i=0}^n \mathbf{d}_i^* U_{n-i}$.

Proof. The first claim directly follows by the definition of \mathcal{S}_c , and the second one by the genealogical order. The number $U_{n+1} - 1$ is then represented by the maximal length- $(n + 1)$ word accepted by the automaton \mathcal{A}_{μ_c} , which is the length- $(n + 1)$ prefix of \mathbf{d}^* . \square

Note that, if the numeration system \mathcal{S}_c satisfies the greedy condition, this result follows from the characterization of numeration systems in terms of dynamical systems given by Bertrand-Mathis [16, 36]. However, even though the function $\text{rep}_{\mathcal{S}_c}$ is obtained using the word-greedy factorization of prefixes of \mathbf{u} , the numeration system \mathcal{S}_c is not necessarily greedy as the following example shows.

Example 13. In Example 3 for $c = 102$, we see that $\mathbf{u}[0, 14) = 012000101 \cdot 012 \cdot 01$, so $\text{rep}_{\mathcal{S}_c}(14) = 10110$, while the greedy representation of 14 associated with the sequence $(U_n)_{n \geq 0}$ is 11000.

In fact, we have the following two characterizations.

Lemma 14. *Let c satisfy (WH). The numeration system $\mathcal{S}_c = (A, \text{rep}_{\mathcal{S}_c}, L)$ is greedy if and only if, for all $v \in L$ and for all $i \leq |v|$, the suffix of length i of v is smaller than or equal to $\mathbf{d}^*[0, i]$. Moreover, we then have*

$$L = \{v = v_1 \cdots v_n \in \mathbb{N}^* \setminus 0\mathbb{N}^* : \forall 1 \leq i \leq n, v_{n-i+1} \cdots v_n \leq \mathbf{d}^*[0, i]\}.$$

Proof. Let us denote $\mathcal{S} = (A', \text{rep}_{\mathcal{S}}, L')$ the canonical greedy numeration system associated with the sequence $(U_n)_{n \geq 0}$. In particular, by uniqueness, \mathcal{S}_c is greedy if and only if $\mathcal{S}_c = \mathcal{S}$. As \mathcal{S}_c is an abstract numeration system, $\text{rep}_{\mathcal{S}_c}$ respects the genealogical order, i.e., $n \leq m$ if and only if $\text{rep}_{\mathcal{S}_c}(n) \leq_{\text{gen}} \text{rep}_{\mathcal{S}_c}(m)$. So does $\text{rep}_{\mathcal{S}}$ by [33, Proposition 2.3.45]. Hence, $\mathcal{S}_c = \mathcal{S}$ if and only if $L = L'$. Moreover, for all $n \geq 0$, $\text{rep}_{\mathcal{S}}(U_n) = 10^n$, so L and L' contain the same number of length- n words by Lemma 12. Thus $L = L'$ if and only if $L \subseteq L'$. By [37, Lemma 5.3], we have

$$L' = \{v = v_1 \cdots v_n \in \mathbb{N}^* \setminus 0\mathbb{N}^* : \forall 1 \leq i \leq n, v_{n-i+1} \cdots v_n \leq \text{rep}_{\mathcal{S}}(U_i - 1)\},$$

so if \mathcal{S}_c is greedy, then $L = L'$ and, by Lemma 12, $\text{rep}_{\mathcal{S}}(U_i - 1) = \mathbf{d}^*[0, i]$ so we conclude. For the converse, let us proceed by contraposition and assume that \mathcal{S}_c is not greedy. Therefore, $L \not\subseteq L'$ and there exists $v \in L$ and $i \leq |v|$ such that $v_{n-i+1} \cdots v_n > \text{rep}_{\mathcal{S}}(U_i - 1)$. However, since $\mathbf{d}^*[0, i]$ is also a representation of $U_i - 1$ associated with the sequence $(U_n)_{n \geq 0}$, we have $\text{rep}_{\mathcal{S}}(U_i - 1) \geq \mathbf{d}^*[0, i]$ (see [33, Proposition 2.3.44] for example). Therefore, the length- i suffix of v is strictly greater than $\mathbf{d}^*[0, i]$, which ends the proof. \square

Theorem 15. *Let $c = c_0 \cdots c_{k-1} \geq 0^k$ with $c_0, c_{k-1} \geq 1$. The numeration system \mathcal{S}_c is greedy if and only if $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is lexicographically maximal among its conjugates.*

Proof. Using Lemma 14 and Theorem 9, \mathcal{S}_c is greedy if and only if, for all $n \geq 0$ and for all $0 \leq i \leq k - 1$, any path $\ell_0 \cdots \ell_n$ starting in state i in the automaton \mathcal{A}_{μ_c} is such that $\ell_0 \cdots \ell_n \leq \mathbf{d}^*[0, n]$. However, by definition of \mathcal{A}_{μ_c} , the lexicographically biggest path of length n starting in state i is given by the prefix of length n of $(c_i \cdots c_{k-2}(c_{k-1} - 1)c_0 \cdots c_{i-1})^\omega$. Therefore, we can conclude that \mathcal{S}_c is greedy if and only if $c_i \cdots c_{k-2}(c_{k-1} - 1)c_0 \cdots c_{i-1} \leq c_0 \cdots c_{k-2}(c_{k-1} - 1)$ for all $0 \leq i \leq k - 1$, i.e., $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates. \square

Example 16. *Let $k = 4$ and $c = 1011$. The sequence U_n satisfies the recurrence relation $U_{n+4} = U_{n+3} + U_{n+1} + U_n$ with initial conditions $U_0 = 1, U_1 = 2, U_2 = 3$, and $U_3 = 5$. A simple induction shows that $(U_n)_{n \geq 0}$ is in fact the sequence of Fibonacci numbers. As $c_0 c_1 c_2 (c_3 - 1) = 1010$ is maximal among its conjugates, the numeration system \mathcal{S}_c then corresponds to the classical Fibonacci numeration system, which can also be obtained with the parameter $c' = 11$. In this case, $c_0 c_1 c_2 (c_3 - 1) = 1010 = v^2$ with $v = 10 = c'_0 (c'_1 - 1)$, which is anti-Lyndon (see Example 2).*

The observation made in the previous example is more general, as we show below. Recall that a real number β is *Parry* if the β -expansion $d_\beta(1)$ of 1 is eventually periodic. It is *simple Parry* if $d_\beta(1)$ is finite, i.e., $d_\beta(1)$ is of the form $u0^\omega$ for some finite word u . See [33, Chapter 2] for more details.

Proposition 17. *Let c satisfy (WH) and be such that $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates.*

1. *We have $c_0 \cdots c_{k-2}(c_{k-1} - 1) = (c'_0 \cdots c'_{j-2}(c'_{j-1} - 1))^\ell$ where $c'_0 \cdots c'_{j-2}(c'_{j-1} - 1)$ is an anti-Lyndon word.*
2. *There exists a simple Parry number β such that $d_\beta(1) = c'_0 \cdots c'_{j-1} 0^\omega$.*
3. *We have $\mathcal{S}_c = \mathcal{S}_{c'}$.*
4. *If $\sigma: i \mapsto i \bmod j$ for all $0 \leq i \leq k-1$, then $\mathbf{u}_{c'} = \sigma(\mathbf{u}_c)$.*

Proof. The first item directly follows from the definition of anti-Lyndon words.

Let us turn to the proof of the second item. As $c'_0 \cdots c'_{j-2}(c'_{j-1} - 1)$ is anti-Lyndon by assumption, $c'_0 \cdots c'_{i-1} > c'_{j-i} \cdots c'_{j-2}(c'_{j-1} - 1)$ for all $i \in \{1, \dots, j-1\}$. Therefore, we obtain $c' > c'_0 \cdots c'_{i-1} \geq c'_{j-i} \cdots c'_{j-1}$. By a result of Parry [18, Corollary 4], this then implies that there exists some simple Parry number β such that $c' 0^\omega = d_\beta(1)$.

Let us turn to the proof of the third item. Write $v = c'_0 \cdots c'_{j-2}(c'_{j-1} - 1)$ (simply put, to get c' , we add 1 to the last letter of v). By the first item, $c = v^{\ell-1} c'$ is a “partial” cyclization of c' . In particular, by definition, we obtain $\mathbf{d}_c^* = \mathbf{d}_v^*$ (where the dependence of \mathbf{d}^* on the chosen parameters is emphasized via a subscript). The numeration systems \mathcal{S}_c and $\mathcal{S}_{c'}$ thus coincide by Theorem 15 and Lemma 14.

Finally, let us show the fourth item. For all $i \not\equiv j-1 \pmod j$, we have

$$\mu_{c'} \circ \sigma(i) = 0^{c'_i \bmod j} ((i \bmod j) + 1) = 0^{c_i} (i + 1 \bmod j) = \sigma \circ \mu_c(i).$$

Similarly, if $i \equiv j-1 \pmod j$, then

$$\begin{aligned} \mu_{c'} \circ \sigma(i) &= 0^{c'_{j-1}} = \begin{cases} 0^{c_i+1}, & \text{if } i \neq k-1; \\ 0^{c_k-1}, & \text{if } i = k-1; \end{cases} \\ &= \sigma \circ \mu_c(i). \end{aligned}$$

This shows that $\mu_{c'} \circ \sigma = \sigma \circ \mu_c$. By induction, we can then show that $\mu_{c'}^n(0) = \sigma(\mu_c^n(0))$ for all $n \geq 0$. Hence, $\mathbf{u}_{c'} = \sigma(\mathbf{u}_c)$. \square

Combined with Corollary 11, this result implies that, if $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates, the word \mathbf{u} is simple-Parry automatic in the sense that it is automatic for the integer numeration system classically associated with a simple Parry number.

Example 18. *We illustrate Proposition 17 by resuming Example 16. We have $c_0 c_1 c_2 (c_3 - 1) = 1010 = v^2$ with $v = 10$ and $c' = 11$. The corresponding simple Parry number is the Golden ratio φ . Moreover, if $\sigma: i \mapsto i \bmod 2$, then*

$$\sigma(\mu_c^\omega(0)) = \sigma(0120301001201 \dots) = 0100101001001 \dots$$

is the Fibonacci word.

4 String attractors of the prefixes

We now turn to the concept of string attractors in relation to the fixed points of the morphisms μ_c , $c \geq 0^k$. A *string attractor* of a finite word $y = y_1 \cdots y_n$ is a set $\Gamma \subseteq \{1, \dots, n\}$ such that every non-empty factor of y has an occurrence crossing a position in Γ , i.e., for each factor $x \in A^m$ of y , there exists $i \in \Gamma$ and j such that $i \in \{j, \dots, j + m - 1\}$ and $x = y[j, j + m)$.

Example 19. *The set $\{2, 3, 4\}$ is a string attractor of the word $0\underline{1}2\underline{0}01$. Indeed, it suffices to check that the factors 0, 1 and 01 have an occurrence crossing one of the underlined positions. No smaller string attractor exists since at least one position in the set is needed per different letter in the word.*

Warning. We would like to stress the following crucial point: in this paper, the letters of infinite words are indexed starting from 0 while the positions in a string attractor are counted starting at 1. This could be seen as confusing, but we use the same notation as the original paper on string attractors [1]. Where ambiguity may occur, we explicitly declare how finite words are indexed.

The family of words \mathbf{u}_c contains the famous k -bonacci words (when $c = 1^k$), and it is known for these words that the positions in $\{U_n : n \geq 0\}$ are sufficient to find string attractors of minimal sizes [10]. It is thus natural to wonder if it is also the case for each word \mathbf{u}_c . We first obtain the following result.

Proposition 20. *Let c satisfy (WH). If every prefix of \mathbf{u} has a string attractor made of elements of $\{U_n : n \geq 0\}$, then $\mathbf{u}[0, U_{n+1} - 1)$ is a fractional power of u_n for all $n \geq 0$.*

Proof. Assume to the contrary that there exists N such that $\mathbf{u}[0, U_{N+1} - 1)$ is not a fractional power of u_N . Therefore, let us denote $u_N x_N b_N$, where x_N is a possibly empty word and b_N is a letter, the shortest prefix of $\mathbf{u}[0, U_{N+1} - 1)$ that is not a fractional power of u_N . We show that $x_N b_N$ is not a factor of $u_N x_N$, and therefore that any string attractor of $u_N x_N b_N$ must contain a position in $[U_N + 1, U_{N+1} - 1)$, which contradicts the assumption.

More generally, we show that, for all n , if $u_n x_n b_n$ is the shortest prefix of \mathbf{u} that is not a fractional power of u_n , then either $x_n = \varepsilon$ or it is always followed in $u_n x_n$ by the letter $b \neq b_n$ such that $u_n x_n b$ is a prefix of u_n^ω . Observe that, in the first case, by Proposition 4, $c_0 = 1$ and $c_i = 0$ for all $1 \leq i \leq n$. In other words, $u_n = 012 \cdots n$ and $b_n = n + 1$. Therefore we can indeed conclude that $x_n b_n$ is not a factor of $u_n x_n$ in both cases.

We proceed by induction on n . If $n = 0$, then $u_n = 0$ and $u_n x_n b_n = 0^{c_0} 1$, so $x_n = 0^{c_0 - 1}$ is always followed by $b = 0$ in $u_n x_n$.

Assume now that the claim is true for $n - 1$ and let us prove it for n . If $x_{n-1} = \varepsilon$, then, as above, $u_n = 012 \cdots n$. Therefore, $u_{n+1} = \mu(u_n) = 012 \cdots n \mu(n)$ and x_n is either ε (if $\mu(n) = (n + 1)$) or 0. In the second case, we have $b_n \in \{0, n + 1\}$ and x_n is only followed by $1 \neq b_n$ in $u_n x_n$ so we conclude.

If $x_{n-1} \neq \varepsilon$, then $\mu(u_{n-1} x_{n-1})$ is a prefix of \mathbf{u} that is a fractional power of u_n , so $\mu(x_{n-1})$ is a prefix of x_n . Notice that x_{n-1} cannot end with $(k - 1)$. Indeed, this follows from Remark 5 since $u_{n-1} x_{n-1}$ is the longest common prefix between \mathbf{u} and u_{n-1}^ω . This implies that any non-suffix occurrence of x_n in $u_n x_n$ comes from a non-suffix

occurrence of x_{n-1} in $u_{n-1}x_{n-1}$. By the induction hypothesis, such an occurrence of x_{n-1} is always followed by $b \neq b_{n-1}$ such that $u_{n-1}x_{n-1}b$ is a prefix of u_{n-1}^ω . Therefore, if 0^ℓ is the longest common prefix between $\mu(b)$ (resp., $\mu(b0)$ if $b = k-1$) and $\mu(b_{n-1})$ (resp., $\mu(b_{n-1}0)$ if $b_{n-1} = k-1$), then $x_n = \mu(x_{n-1})0^\ell$ and x_n is always followed in $u_n x_n$ by $c \neq b_n$ such that $\mu(u_{n-1}x_{n-1})0^\ell c = u_n x_n c$ is a prefix of u_n^ω . \square

Based on this result, we see that fractional powers of the word u_n will play a key role in determining string attractors. We thus introduce some notations.

Definition 3. Let c satisfy (WH). For all $n \geq 0$, we let q_n denote the longest prefix of \mathbf{u} that is a fractional power of u_n , i.e., the longest common prefix between \mathbf{u} and $(u_n)^\omega$. For all $n \geq 0$, we also let $Q_n = |q_n|$.

Working with fractional powers also has another advantage from the string attractor point of view. Indeed, there is no trivial link in general between the string attractors of the finite words w and wa , where a is a letter. However, we have the following result which can be derived from the proofs of [7, Propositions 12 and 15].

Proposition 21. Let z be a non-empty word and let x and y be fractional powers of z with $|z| \leq |x| \leq |y|$. If Γ is a string attractor of x , then $\Gamma \cup \{z\}$ is a string attractor of y .

Motivated by this result, to describe string attractors of each prefix, it is now sufficient to be able to describe, for all $n \geq 1$, a string attractor of a prefix of length m_n for some $m_n \in [U_n - 1, Q_{n-1}]$. This argument is the key in the proof of our main theorem.

For $n \geq 0$, we denote

$$\Gamma_n = \begin{cases} \{U_0, \dots, U_n\}, & \text{if } 0 \leq n \leq k-1; \\ \{U_{n-k+1}, \dots, U_n\}, & \text{if } n \geq k. \end{cases}$$

We also define

$$P_n = \begin{cases} U_n, & \text{if } 0 \leq n \leq k-1; \\ U_n + U_{n-k+1} - U_{n-k} - 1, & \text{if } n \geq k. \end{cases} \quad (2)$$

The next lemma directly follows from the definition of P_n and from Proposition 4.

Lemma 22. Let c satisfy (WH). Then $P_n \leq U_{n+1} - 1$ for all $n \geq 0$.

To simplify the statement of the following theorem, we set $\Gamma_{-1} = \emptyset$.

Theorem 23. Let $c = c_0 \cdots c_{k-1} \geq 0^k$ with $c_0, c_{k-1} \geq 1$ and such that $\mathbf{u}[0, U_{n+1} - 1]$ is a fractional power of u_n , and let $n \geq 0$.

1. If $m \in [U_n, Q_n]$, then $\Gamma_{n-1} \cup \{U_n\}$ is a string attractor of $\mathbf{u}[0, m)$.
2. If $m \in [P_n, Q_n]$, then Γ_n is a string attractor of $\mathbf{u}[0, m)$.

Proof. Let us simultaneously prove the two claims by induction on n . If $n = 0$, then $1 \leq m \leq c_0$, so $\mathbf{u}[0, m) = 0^m$ and the conclusion directly follows for both claims. Assume now that the claims are satisfied for $n-1$ and let us prove them for n . By Lemma 22, $P_{n-1} \leq U_n - 1$, and by hypothesis on c , $U_n - 1 \leq Q_{n-1}$. Therefore, by the induction hypothesis, Γ_{n-1} is a string attractor of $\mathbf{u}[0, U_n - 1)$. This implies

that $\Gamma_{n-1} \cup \{U_n\}$ is a string attractor of u_n so, by Proposition 21, of $\mathbf{u}[0, m)$ for all $m \in [U_n, Q_n]$. This ends the proof of the first claim.

Let us now prove the second claim. Observe that, using Proposition 21, it suffices to prove that Γ_n is a string attractor of $\mathbf{u}[0, P_n)$. If $0 \leq n \leq k-1$, then $\Gamma_n = \Gamma_{n-1} \cup \{U_n\}$ so we can directly conclude using the first claim. Thus assume that $n \geq k$. Then by the first claim, $\Gamma_n \cup \{U_{n-k}\} = \Gamma_{n-1} \cup \{U_n\}$ is a string attractor of $\mathbf{u}[0, P_n)$. Therefore, it remains to show that the position U_{n-k} is not needed in the string attractor. In other words, we prove that the factors of $\mathbf{u}[0, P_n)$ having an occurrence crossing position U_{n-k} (and no other position of $\Gamma_n \cup \{U_{n-k}\}$) have another occurrence crossing a position in Γ_n . More precisely, we show that they have an occurrence crossing position U_n . To help the reader with the proof, we illustrate the situation in Figure 2.

As the smallest position in Γ_n is U_{n-k+1} , we need to consider the factor occurrences crossing position U_{n-k} in $\mathbf{u}[0, U_{n-k+1} - 1)$. So, if we write $\mathbf{u}[0, P_n) = u_n w$, it is sufficient to show that u_{n-k} is a suffix of u_n and that $w' := \mathbf{u}[U_{n-k}, U_{n-k+1} - 1)$ is a prefix of w . Observe that

$$|w| = P_n - U_n = U_{n-k+1} - U_{n-k} - 1 \quad (3)$$

by definition of P_n , so $|w'| = |w|$ and we actually show that $w' = w$.

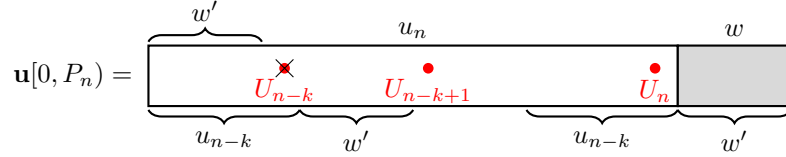


Fig. 2 Representation of the proof of the second claim of Theorem 23. As we warned the reader before, elements in a string attractor are indexed starting at 1 (in red), while indices of letters in \mathbf{u} start at 0.

The fact that u_{n-k} is a suffix of u_n is a direct consequence of Proposition 4 as $c_{k-1} \geq 1$ by assumption. Since $\mathbf{u}[0, U_{n-k+1} - 1)$ is a fractional power of u_{n-k} by assumption, w' is a prefix of u_n . By Lemma 22 and by assumption, we also have $P_n \leq U_{n+1} - 1 \leq Q_n$, so $\mathbf{u}[0, P_n)$ is a fractional power of u_n . This implies that $w = w'$. \square

5 Fractional power prefixes and anti-Lyndon words

In this section, we study the words q_n and their lengths Q_n . As we will show in Proposition 31, these words have a particular structure related to (anti-)Lyndon words. To prove this, we introduce some more notations. For all $n \geq 0$, the pair $\{i_n, j_n\}$ designates the two (distinct) letters following q_n in \mathbf{u} and in $(u_n)^\omega$. Without loss of generality, we always assume that $i_n < j_n$.

Example 24. Set $c = 102$. Recall from Example 3 that the first few words in $(u_n)_{n \geq 0}$ are 0, 01, 012, 01200, 012000101, 012000101012012. It is then easy to see that the first

few words in $(q_n)_{n \geq 0}$ are 0, 01, 0120, 0120001, 0120001010120. So we conclude that the first few pairs in $(\{i_n, j_n\})_{n \geq 0}$ are $\{0, 1\}$, $\{0, 2\}$, $\{0, 1\}$, $\{0, 2\}$, $\{0, 1\}$.

The following lemma gives a recursive construction for the sequences $(i_n)_{n \geq 0}$ and $(j_n)_{n \geq 0}$, as well as a first structure for the words q_n .

Lemma 25. *Let c satisfy (WH). For all $n \geq 0$, we have $q_n = u_n^{\ell_0} u_{n-1}^{\ell_1} \cdots u_0^{\ell_n}$ where the sequences $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ are recursively constructed as follows: $\ell_0 = c_0$, $i_0 = 0$, $j_0 = 1$, and for all $n \geq 0$, if $j_n \leq k - 2$, we have*

$$\{\ell_{n+1}, i_{n+1}, j_{n+1}\} = \begin{cases} \{c_{j_n}, 0, j_n + 1\}, & \text{if } c_{i_n} > c_{j_n}; \\ \{c_{j_n}, i_n + 1, j_n + 1\}, & \text{if } c_{i_n} = c_{j_n}; \\ \{c_{i_n}, 0, i_n + 1\}, & \text{if } c_{i_n} < c_{j_n}; \end{cases}$$

and if $j_n = k - 1$, we have $\{\ell_{n+1}, i_{n+1}, j_{n+1}\} = \{c_{i_n}, 0, i_n + 1\}$.

Proof. We prove the claimed structure for the sequences $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ and also that $c_0 = \max\{c_0, \dots, c_{j_n-1}\}$ for all $n \geq 0$ by induction.

For the base case $n = 0$, as $u_0 = 0$ and $u_1 = 0^{c_0}1$ is a prefix of \mathbf{u} , we directly have $\ell_0 = c_0$, $i_0 = 0$, $j_0 = 1$ and $c_0 = \max\{c_0\}$.

Let us now move to the induction step: assume that both claims are satisfied for n and let us prove them for $n + 1$. For the first claim, by definition, $\mu(q_n)$ is a prefix of both $\mu(\mathbf{u}) = \mathbf{u}$ and $\mu(u_n)^\omega = (u_{n+1})^\omega$. Moreover, it is followed in one of them by $\mu(i_n) = 0^{c_{i_n}} \cdot (i_n + 1)$ and in the other by $\mu(j_n)$. The image of j_n under μ takes two forms.

If $j_n \leq k - 2$, then $\mu(j_n) = 0^{c_{j_n}} \cdot (j_n + 1)$. Thus, as $i_n + 1 \neq j_n + 1$, we have $q_{n+1} = \mu(q_n)0^{\ell_{n+1}}$ where $0^{\ell_{n+1}}$ is the longest common prefix between $\mu(i_n)$ and $\mu(j_n)$. We then have

$$\{\ell_{n+1}, i_{n+1}, j_{n+1}\} = \begin{cases} \{c_{j_n}, 0, j_n + 1\}, & \text{if } c_{i_n} > c_{j_n}; \\ \{c_{j_n}, i_n + 1, j_n + 1\}, & \text{if } c_{i_n} = c_{j_n}; \\ \{c_{i_n}, 0, i_n + 1\}, & \text{if } c_{i_n} < c_{j_n}. \end{cases}$$

The conclusion of the first claim follows from the fact that $\mu(q_n) = u_{n+1}^{\ell_0} \cdots u_1^{\ell_n}$ by the induction hypothesis.

If $j_n = k - 1$, then by Remark 5, q_{n+1} is not only followed by $\mu(k - 1)$ but by $\mu(k - 1)\mu(0) = 0^{c_{k-1}+c_0} \cdot 1$. By the second claim, we have

$$c_{i_n} \leq \max\{c_0, \dots, c_{k-2}\} = c_0 < c_{k-1} + c_0$$

as $c_{k-1} \geq 1$ by assumption. We conclude that $\{\ell_{n+1}, i_{n+1}, j_{n+1}\} = \{c_{i_n}, 0, i_n + 1\}$.

The second claim is also satisfied as $\max\{c_0, \dots, c_{j_{n+1}-1}\} \leq \max\{c_0, \dots, c_{j_n-1}\}$. Indeed, in all cases, either $j_{n+1} \leq j_n$, or $j_{n+1} = j_n + 1$ and $c_{j_n} \leq \max\{c_0, \dots, c_{j_n-1}\}$. \square

Example 26. *Let us take $c = 210221$ for which $k = 6$. The first few elements of the sequences $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ are given in Table 3. We already observe that they*

Table 3 Illustration of the construction of the sequences $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ in the case where $c = 210221$.

n	Comparison	ℓ_n	$\{i_n, j_n\}$
0	/	$c_0 = 2$	$\{0, 1\}$
1	$c_0 > c_1$	$c_1 = 1$	$\{0, 2\}$
2	$c_0 > c_2$	$c_2 = 0$	$\{0, 3\}$
3	$c_0 = c_3$	$c_3 = 2$	$\{1, 4\}$
4	$c_1 < c_4$	$c_1 = 1$	$\{0, 2\}$
5	$c_0 > c_2$	$c_2 = 0$	$\{0, 3\}$
6	$c_0 = c_3$	$c_3 = 2$	$\{1, 4\}$

are (eventually) periodic. Indeed, $\{i_1, j_1\} = \{0, 2\} = \{i_4, j_4\}$ and, as $\{i_N, j_N\}$ entirely determines the rest of the sequences, $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ are eventually periodic of period length 3 starting from index 1 (and even from index 0 for $(\ell_n)_{n \geq 0}$).

From the recursive definition given in Lemma 25, we derive the following result.

Lemma 27. *Let c satisfy (WH). For all $n \geq 0$, the word $c_0 \cdots c_{i_n-1}$ is a border of the word $c_0 \cdots c_{j_n-1}$, i.e., $c_0 \cdots c_{i_n-1} = c_{j_n-i_n} \cdots c_{j_n-1}$.*

Proof. Once again, we prove the result by induction on $n \geq 0$. Notice that, if $i_n = 0$, then the word $c_{j_n-i_n} \cdots c_{j_n-1}$ is empty, hence the conclusion. This is in particular the case for $n = 0$. Assume now that the claim holds for n and let us prove it for $n + 1$. By Lemma 25, we have $i_{n+1} = 0$ unless $c_{i_n} = c_{j_n}$. In this case, $i_{n+1} = i_n + 1$ and $j_{n+1} = j_n + 1$ so, as $c_0 \cdots c_{i_n-1} = c_{j_n-i_n} \cdots c_{j_n-1}$ by the induction hypothesis, we directly have $c_0 \cdots c_{i_{n+1}-1} = c_{j_{n+1}-i_{n+1}} \cdots c_{j_{n+1}-1}$. \square

We now show the link with (anti-)Lyndon words. Before doing so, we recall some famous properties of Lyndon words that will be useful. The first result is part of the folklore, but a proof can be found, for instance, in [38].

Proposition 28. *Lyndon words are unbordered, i.e., if w is both a prefix and a suffix of a Lyndon word v , then $w = \varepsilon$ or $w = v$.*

The next result is shown within the proof of the Chen-Fox-Lyndon Theorem (Theorem 1). See, for instance, [26, Theorem 5.1.5].

Proposition 29. *Let $w \in A^*$ be a non-empty word and let (ℓ_1, \dots, ℓ_n) be its Lyndon factorization as in Theorem 1. Then ℓ_1 is the longest Lyndon prefix of w .*

Duval provided an algorithm computing the Lyndon factorization of a word in linear time [39]. It is based on a decomposition of the word into three parts xyz : we already computed the Lyndon factorization of x and we are now looking at $w = yz$, where y is a fractional power of a Lyndon word v and z is the part that we still need to explore. We keep track of the position of the first letter of z with an index j , and of the period of y (i.e., the length of v) using an index i such that $j - i = |v|$.

Algorithm 30 (Duval [39]). *Let (A, \leq) be an ordered set and let $w = w_0 \cdots w_n$ be a length- n word over A . We denote w_{n+1} a new symbol smaller than all the letters of w . Set $i = 0$ and $j = 1$. While $i \leq n$, compare w_i and w_j and do the following:*

- if $w_i < w_j$, then set $j = j + 1$ and $i = 0$;
- if $w_i = w_j$, then set $j = j + 1$ and $i = i + 1$;

- if $w_i > w_j$, then output $w_0 \cdots w_{j-i-1}$ as the next element in the Lyndon factorization and restart the algorithm with the word $w_{j-i} \cdots w_n$.

Using the notation of the paragraph preceding Algorithm 30, we explain the three cases present in the algorithm. We want to compute the next Lyndon word in the Lyndon factorization of a word, knowing that of some of its prefixes. By definition of i and j , we compare the letter w_j in z with the letter w_i , spaced by $|v|$ letters.

- If $w_i < w_j$, then yw_j is a Lyndon word by [40, Lemme 2], so we update y to yw_j and v to y .
- If $w_i = w_j$, then yw_j is still a fractional power of v , so we simply update y to yw_j without changing the length of v (that is, we do not modify $j - i$).
- If $w_i > w_j$, then yw_j cannot be a prefix of a Lyndon word, so the longest Lyndon prefix of w is v .

We are now ready to prove the structure of the words q_n and its link with anti-Lyndon words.

Proposition 31. *Let c satisfy (WH). Define \mathbf{a} as the infinite concatenation of the longest anti-Lyndon prefix of the word $c_0 \cdots c_{k-2}$. Then for all $n \geq 0$, $q_n = u_n^{\mathbf{a}_0} u_{n-1}^{\mathbf{a}_1} \cdots u_0^{\mathbf{a}_n}$. In particular, $Q_n = \sum_{i=0}^n \mathbf{a}_i U_{n-i}$.*

Proof. By Lemma 25, the beginning of the construction of the sequences $(\ell_n)_{n \geq 0}$, $(i_n)_{n \geq 0}$, $(j_n)_{n \geq 0}$ corresponds exactly to the first application of Duval's algorithm to the word $c_0 \cdots c_{k-2}$ with the order \leq_- . More specifically, letting N denote the first index n for which $c_{i_n} < c_{j_n}$ or $j_n = k - 1$ and setting $p = j_N - i_N$, then Duval's algorithm for \leq_- implies that the word $\ell_0 \cdots \ell_{p-1}$ is the first element in the Lyndon factorization of $c_0 \cdots c_{k-2}$ for the order \leq_- . Therefore, $\ell_0 \cdots \ell_{p-1} = c_0 \cdots c_{p-1}$ is the longest anti-Lyndon prefix of $c_0 \cdots c_{k-2}$ by Proposition 29. Let us denote it v . As in the statement, let $\mathbf{a} = vvv \cdots$.

Observe that, by definition of N and by Lemma 25, for all $1 \leq n \leq N$, we have $j_n = n + 1$ as it is incremented at each step, and $\ell_n = c_{j_{n-1}} = c_n$. In particular, $p = j_N - i_N = N + 1 - i_N$.

We now prove that $\ell_n = \mathbf{a}_n$ for all $n \geq 0$. By definition of \mathbf{a} , the equality holds for $0 \leq n < p$, so it is enough to look at all $n \geq p$. We show by induction on $n \geq p$ that $\ell_n = c_{n \bmod p}$, $j_n \equiv (n + 1) \bmod p$, and $j_n \leq N + 1$.

For $p \leq n \leq N$, we already have $\ell_n = c_n$, $j_n = n + 1$, and $j_n \leq N + 1$ by the observation made above. Moreover, Duval's algorithm implies that $c_0 \cdots c_N$ is periodic of period length p , so $\ell_n = c_n = c_{n \bmod p}$. This is also true for $n = N + 1$ as $N + 1 = p + i_N \equiv i_N \bmod p$. Indeed, by Lemma 25 and by definition of N , we have $\ell_{N+1} = c_{i_N} = c_{N+1 \bmod p}$ and

$$j_{N+1} = i_N + 1 \equiv N + 2 \bmod p. \quad (4)$$

Assume now that the claim is true for indices up to $n \geq N + 1$ and let us prove it for $n + 1$. By the induction hypothesis, we have $j_n \leq N + 1$, so we distinguish two cases.

Case 1. If $j_n \leq N$, then $j_n \leq k - 2$ (as $j_N = N + 1 \leq k - 1$). By Lemma 27, comparing c_{i_n} and c_{j_n} is equivalent to comparing $c_0 \cdots c_{i_n}$ and $c_{j_n - i_n} \cdots c_{j_n}$. As mentioned earlier in the proof, $c_0 \cdots c_N$ is a fractional power of v , so $c_0 \cdots c_{i_n}$ is a prefix

of a power of v while $c_{j_n-i_n} \cdots c_{j_n}$ is a prefix of a power of a conjugate of v . As v is Lyndon for \leq_- , its powers are smaller than the powers of its conjugates for \leq_- , thus $c_0 \cdots c_{i_n} \leq_- c_{j_n-i_n} \cdots c_{j_n}$ and $c_{i_n} \leq_- c_{j_n}$, i.e., $c_{i_n} \geq c_{j_n}$. Using Lemma 25, we conclude that $\ell_{n+1} = c_{j_n} = c_{n+1 \bmod p}$ as $j_n \leq N$ is congruent to $n+1 \bmod p$ by the induction hypothesis and $c_0 \cdots c_N$ has period length p . We also have $j_{n+1} = j_n + 1$ thus $j_{n+1} \leq N+1$ and $j_{n+1} \equiv n+2 \bmod p$.

Case 2. If $j_n = N+1$, then using Lemma 27, we know that $c_0 \cdots c_N = c_0 \cdots c_{j_n-1}$ has a border of length i_n so $c_0 \cdots c_N$ has period length $N+1-i_n$. Since it also has period length p and $c_0 \cdots c_{p-1}$ is anti-Lyndon thus unbordered by Proposition 28, we must have that $N+1-i_n$ is a multiple of $p = N+1-i_N$. In other words,

$$i_n \equiv i_N \pmod{p}. \quad (5)$$

In particular, by periodicity, $c_{i_n} = c_{i_N}$. Moreover, $j_n = N+1 = j_N$ so $\{c_{i_n}, c_{j_n}\} = \{c_{i_N}, c_{j_N}\}$. Therefore, by Lemma 25 and by definition of N , we have

$$\ell_{n+1} = \ell_{N+1} \quad \text{and} \quad j_{n+1} = i_n + 1 \leq N+1. \quad (6)$$

By the induction hypothesis for n , we have

$$N+1 = j_n \equiv n+1 \pmod{p}. \quad (7)$$

We conclude that

$$\ell_{n+1} = \ell_{N+1} = c_{(N+1) \bmod p} = c_{(n+1) \bmod p},$$

where the first equality follows by (6), the second by the induction hypothesis for $N+1$, and the last by Congruence (7), and

$$j_{n+1} = i_n + 1 \equiv i_N + 1 \equiv N+2 \equiv n+2 \pmod{p},$$

where the first equality follows from (6), the first congruence from (5), the second by (4), and the last by Congruence (7). This ends the proof. \square

Example 32. Let us pursue Example 24 for which $c = 102$. The first few words in $(q_n)_{n \geq 0}$ are 0, 01, 0120, 0120001, 0120001010120. The longest anti-Lyndon prefix of $c_0 c_1 = 10$ is 10 itself so $\mathbf{a} = (10)^\omega$. We can easily check that the first few q_n 's indeed satisfy Proposition 31.

Now that we have a good understanding of the fractional powers of the words u_n , we look for an equivalent description of the condition in Proposition 20 and Theorem 23. This is the purpose of Proposition 34, but we first need the following technical lemma.

Lemma 33. Let c satisfy (WH) and let w denote the longest anti-Lyndon prefix of $c_0 \cdots c_{k-2}$.

1. Then $c_0 \cdots c_{k-2} \geq \mathbf{a}[0, k-2]$.
2. Moreover, $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates if and only if the following three assertions hold:

- (a) We have $c_0 \cdots c_{k-2} = \mathbf{a}[0, k-2]$.
- (b) We have $c_{k-1} - 1 \leq \mathbf{a}_{k-1}$.
- (c) If $c_{k-1} - 1 = \mathbf{a}_{k-1}$, then $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is an integer power of w .

Proof. We show the first claim. Assume by contradiction that there exists a minimal index $i \in \{|w|, \dots, k-2\}$ such that $c_0 \cdots c_i < \mathbf{a}[0, i]$. Then $c_0 \cdots c_i = w^\ell va$ with a proper prefix v of w and a letter a such that $va < w$. So [40, Lemme 2] implies that $c_0 \cdots c_i$ is an anti-Lyndon prefix of $c_0 \cdots c_{k-2}$. As $i \geq |w|$, this contradicts the maximality of w .

We now turn to the second claim. Assume that $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates. We first show that $c_0 \cdots c_{k-2}(c_{k-1} - 1) \leq \mathbf{a}[0, k-1]$. If it is not the case, there exist $\ell \geq 1$, a proper prefix u of w , a letter a and a word v such that $c_0 \cdots c_{k-2}(c_{k-1} - 1) = w^\ell uav$ and $ua > w$. Then $uavw^\ell > c_0 \cdots c_{k-2}(c_{k-1} - 1)$, so $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is not maximal among its conjugates. This is a contradiction. Therefore, we have $c_0 \cdots c_{k-2}(c_{k-1} - 1) \leq \mathbf{a}[0, k-1]$. Using the first claim, we get $c_0 \cdots c_{k-2} = \mathbf{a}[0, k-2]$ and $c_{k-1} - 1 \leq \mathbf{a}_{k-1}$, which gives Items 2a and 2b. Now if $c_{k-1} - 1 = \mathbf{a}_{k-1}$, then $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is a fractional power of w . It is moreover maximal among its conjugates by assumption so, since w is anti-Lyndon, $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is in fact an integer power of w . This proves Item 2c.

Let us now suppose that Assumptions 2a, 2b, and 2c hold. Let us consider a conjugate $c_i \cdots c_{k-2}(c_{k-1} - 1)c_0 \cdots c_{i-1}$ for some $1 \leq i \leq k-1$. We have

$$c_i \cdots c_{k-2}(c_{k-1} - 1) = \mathbf{a}[i, k-2](c_{k-1} - 1) \leq \mathbf{a}[i, k-1] \leq \mathbf{a}[0, k-1-i] = c_0 \cdots c_{k-1-i},$$

where the first and last equalities come from Assumption 2a, the first inequality comes from Assumption 2b, and the second inequality comes from the fact that $\mathbf{a} = w^\omega$ and w is anti-Lyndon. Therefore, we have two cases. If $c_i \cdots c_{k-2}(c_{k-1} - 1) < c_0 \cdots c_{k-1-i}$, then

$$c_i \cdots c_{k-2}(c_{k-1} - 1)c_0 \cdots c_{i-1} < c_0 \cdots c_{k-2}(c_{k-1} - 1),$$

which is enough in this case. Otherwise, we get $c_{k-1} - 1 = \mathbf{a}_{k-1}$, so $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is an integer power of w by Assumption 2c. Hence it is maximal among its conjugates since w is anti-Lyndon. \square

Proposition 34. *Let c satisfy (WH). The following assertions are equivalent.*

1. The word $\mathbf{u}[0, U_{n+1} - 1)$ is a fractional power of u_n , i.e., $U_{n+1} - 1 \leq Q_n$, for all $n \geq 0$.
2. We have $\mathbf{d}^*[0, n] \leq \mathbf{a}[0, n]$ for all $n \geq 0$.
3. The word $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates.
4. The numeration system \mathcal{S}_c is greedy.

Proof. We prove that 1 implies 2 by contraposition. Let $n \geq 0$ be the smallest integer such that $\mathbf{d}^*[0, n] > \mathbf{a}[0, n]$. Note that $n \neq 0$ since $\mathbf{d}_0^* = c_0 = \mathbf{a}_0$. By minimality, we have $\mathbf{d}^*[0, n-1] \leq \mathbf{a}[0, n-1]$, so $\mathbf{d}^*[0, n-1] = \mathbf{a}[0, n-1]$ and $\mathbf{d}_n^* > \mathbf{a}_n$. Therefore, by the last parts of both Lemma 12 and Proposition 31, we have $U_{n+1} - 1 > Q_n$.

We show that 2 implies 1 by contraposition. Assume that there exists an integer n such that $U_{n+1} - 1 > Q_n$ and let us show that $\mathbf{d}^*[0, n] > \mathbf{a}[0, n]$. By Proposition 31,

$q_n = u_n^{\mathbf{a}_0} \cdots u_0^{\mathbf{a}_n}$ is a proper prefix of $\mathbf{u}[0, U_{n+1} - 1)$. By Lemma 12, $\text{rep}_{\mathcal{S}_c}(U_{n+1} - 1) = \mathbf{d}^*[0, n]$, so \mathbf{d}_0^* is the largest exponent e such that u_n^e is a prefix of $\mathbf{u}[0, U_{n+1} - 1)$. This implies that $\mathbf{d}_0^* \geq \mathbf{a}_0$. Moreover, if $\mathbf{a}_0 = \mathbf{d}_0^*$, the same argument implies that \mathbf{d}_1^* is the largest exponent e such that $u_n^{\mathbf{d}_0^*} u_{n-1}^e$ is a prefix of $\mathbf{u}[0, U_{n+1} - 1)$. In both cases, we have $\mathbf{d}_0^* \mathbf{d}_1^* \geq \mathbf{a}_0 \mathbf{a}_1$. We may iterate the reasoning to obtain $\mathbf{d}^*[0, n] \geq \mathbf{a}[0, n]$. As q_n is a proper prefix of $\mathbf{u}[0, U_{n+1} - 1)$, the inequality cannot be an equality so we conclude.

We prove that 2 implies 3. Assume that $\mathbf{d}^*[0, n] \leq \mathbf{a}[0, n]$ for all $n \geq 0$. Using the first part of Lemma 33, this directly implies Items 2a and 2b of Lemma 33. Let us show Item 2c to conclude that $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates by Lemma 33. Assume that $c_{k-1} - 1 = \mathbf{a}_{k-1}$. Therefore, $c_0 \cdots c_{k-2}(c_{k-1} - 1) = w^\ell v$ for some proper prefix v of w and $\ell \geq 1$. Let u be such that $w = vu$. We then have

$$w^\ell v u = \mathbf{d}^*[0, (\ell + 1)|w| + |v|] \leq \mathbf{a}[0, (\ell + 1)|w| + |v|] = w^\ell v u v.$$

Since w is anti-Lyndon, the only possibility is to have $v = \varepsilon$ and $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is an integer power of w . This proves Item 2c.

We show that 3 implies 2. By Item 2b of Lemma 33, we have $c_{k-1} - 1 \leq \mathbf{a}_{k-1}$. If the previous inequality is strict, then the conclusion is direct by Item 2a of Lemma 33. Otherwise, by Item 2c of Lemma 33, $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is an integer power of w and we conclude that $\mathbf{a} = \mathbf{d}^*$, which is enough.

Finally, the Assertions 3 and 4 are equivalent by Theorem 15. This ends the proof. \square

Remark 35. Examining the proof of 2 implies 3 of Proposition 34, we observe that it is enough to know that $\mathbf{d}^*[0, k - 1 + |w|] \leq \mathbf{a}[0, k - 1 + |w|]$, so Assertions 1 and 2 may be respectively replaced as follows.

- 1'. The word $\mathbf{u}[0, U_{n+1} - 1)$ is a fractional power of u_n , i.e., $U_{n+1} - 1 \leq Q_n$, for all $0 \leq n \leq k - 1 + |w|$.
- 2'. We have $\mathbf{d}^*[0, n] \leq \mathbf{a}[0, n]$ for all $0 \leq n \leq k - 1 + |w|$.

From Proposition 20 and Theorem 23, we directly obtain the following corollary.

Corollary 36. Let c satisfy (WH). Every prefix of \mathbf{u}_c has a string attractor made of elements of $\{U_n : n \geq 0\}$ if and only if one of the four assertions of Proposition 34 is satisfied.

6 Optimality of the string attractors

So far we were interested in obtaining a precise description (related to a specific numeration system) of a string attractor for each prefix of the infinite word of interest. In this section, we rather focus on the size of attractors and therefore recall the following concept from [4, 8]. Given an infinite word \mathbf{x} and any integer $n \geq 1$, we let $s_{\mathbf{x}}(n)$ denote the size of a smallest string attractor for the length- n prefix of \mathbf{x} . The function $s_{\mathbf{x}} : n \mapsto s_{\mathbf{x}}(n)$ is called the *string attractor profile function* of \mathbf{x} . As a consequence of Theorem 23, we obtain the following.

Corollary 37. Let c satisfy (WH) and assume that one of the assertions of Proposition 34 holds.

1. For all $n \geq 0$ and all $m \in [P_n, Q_n]$, $s_{\mathbf{u}}(m) = \text{Card}(\Gamma_n)$.
2. For all large enough m , $k \leq s_{\mathbf{u}}(m) \leq k + 1$.

Proof. Using a simple induction, one can check that the positions in Γ_n all correspond to different letters in \mathbf{u} . As any string attractor must contain at least one position per letter, we deduce the first claim from Theorem 23. The second part similarly follows from Theorem 23 since, for all large enough n , Γ_n is of size k , and the intervals $[U_n, Q_n]$ cover $\mathbb{N} \setminus \{0\}$ by Proposition 34. \square

For some parameters c , the intervals $[P_n, Q_n]$ also cover $\mathbb{N} \setminus \{0\}$, implying that a string attractor of minimal size can always be given by some Γ_n . This is for example the case of $c = 211$. We characterize these parameters c in the following result. Recall the definition of the sequence $(P_n)_{n \geq 0}$ given in Equation (2):

$$P_n = \begin{cases} U_n, & \text{if } 0 \leq n \leq k-1; \\ U_n + U_{n-k+1} - U_{n-k} - 1, & \text{if } n \geq k. \end{cases}$$

Proposition 38. *Let c satisfy (WH) and let w denote the longest anti-Lyndon prefix of $c_0 \cdots c_{k-2}$. Moreover assume that one of the assertions of Proposition 34 holds. The inequality $P_n - 1 \leq Q_{n-1}$ holds for all n if and only if the following conditions are satisfied:*

1. we have $c_{k-1} = 1$;
2. the word $c_0 \cdots c_{k-2}$ is an integer power of w .

Proof. Observe that, for all $n \leq k-1$, Equation (1) gives

$$P_n - 1 = U_n - 1 = c_0 U_{n-1} + \cdots + c_n U_0.$$

On the other hand, Lemma 33 implies that $c_0 \cdots c_{k-2} = \mathbf{a}[0, k-2]$, so

$$Q_{n-1} = \mathbf{a}_0 U_{n-1} + \cdots + \mathbf{a}_n U_0 = P_n - 1$$

by Proposition 31. This shows that the inequality of the statement always holds for all $n \leq k-1$. Let us now show that it is also satisfied for $n \geq k$ if and only if the two conditions of the statement are satisfied.

For $n \geq k$, we have

$$\begin{aligned} P_n - 1 &= U_n + U_{n-k+1} - U_{n-k} - 2 \\ &= c_0 U_{n-1} + \cdots + c_{k-2} U_{n-k+1} + (c_{k-1} - 1) U_{n-k} + U_{n-k+1} - 2, \end{aligned}$$

where the second equality follows from Equation (1), and

$$Q_{n-1} = \mathbf{a}_0 U_{n-1} + \cdots + \mathbf{a}_{k-2} U_{n-k+1} + \mathbf{a}_{k-1} U_{n-k} + \cdots + \mathbf{a}_{n-1} U_0.$$

Therefore, by Lemma 33,

$$P_n - 1 \leq Q_{n-1} \iff (c_{k-1} - 1) U_{n-k} + U_{n-k+1} - 2 \leq \mathbf{a}_{k-1} U_{n-k} + \cdots + \mathbf{a}_{n-1} U_0. \quad (8)$$

Now, if the two conditions of the statement are fulfilled, i.e., if $c_{k-1} = 1$ and $c_0 \cdots c_{k-2}$ is an integer power of w , then $\mathbf{a}_{k-1+\ell} = \mathbf{a}_\ell$ for all $\ell \geq 0$, so

$$P_n - 1 \leq Q_{n-1} \iff U_{n-k+1} - 2 \leq Q_{n-k}.$$

Since $U_{n-k+1} - 2 \leq P_{n-k+1} - 1$, we can then conclude by induction that the inequality of the statement is satisfied for all $n \geq 0$.

Let us now prove the converse by a detailed case-analysis, and assume first by contradiction that $c_{k-1} \neq 1$. Since $c_{k-1} \geq 1$, this in fact means that $c_{k-1} \geq 2$. For $n = k$, the right-hand side of (8) becomes

$$(c_{k-1} - 1)U_0 + U_1 - 2 \leq \mathbf{a}_{k-1}U_0,$$

which gives, since $U_0 = 1$ and $U_1 = c_0 + 1$,

$$c_0 \leq c_{k-1} - 2 + c_0 \leq \mathbf{a}_{k-1}.$$

Since \mathbf{a} is the infinite concatenation of an anti-Lyndon word, we have $\mathbf{a}_{k-1} \leq \mathbf{a}_0 = c_0$. Therefore, we deduce that $\mathbf{a}_{k-1} = c_0$ and $c_{k-1} = 2$. Now, if $n = k + 1$, the right-hand side of (8) becomes

$$(c_{k-1} - 1)U_1 + U_2 - 2 \leq \mathbf{a}_{k-1}U_1 + \mathbf{a}_kU_0 \iff c_0 + 1 + (U_2 - c_0U_1) - 2 \leq \mathbf{a}_k.$$

This leads us to consider two cases depending on k . If $k = 2$, we get $U_2 - c_0U_1 = c_1 = 2$ so $c_0 + 1 + 2 - 2 \leq \mathbf{a}_2$, but as $\mathbf{a}_2 \leq \mathbf{a}_0 = c_0$, this is impossible. If $k \geq 3$, then $U_2 - c_0U_1 = c_1 + 1$, however since $\mathbf{a}_{k-1}\mathbf{a}_k \leq \mathbf{a}_0\mathbf{a}_1 = c_0c_1$ and $\mathbf{a}_{k-1} = c_0$ as established above, we must have $c_0 + 1 + c_1 + 1 - 2 \leq \mathbf{a}_k \leq c_1$ which is also impossible.

This shows that, if $P_n - 1 \leq Q_{n-1}$ for all $n \geq 0$, then $c_{k-1} = 1$. Assume now by contradiction that $c_0 \cdots c_{k-2}$ is not an integer power of w . Therefore, by Lemma 33, there exist $\ell \geq 1$ and a proper non-empty prefix x of w such that $c_0 \cdots c_{k-2} = w^\ell x$. Let us denote $w = xy$ for some (non-empty) word y .

For $n = k + |w|$, the right-hand side of (8) becomes

$$\begin{aligned} U_{|w|+1} - 2 &\leq \mathbf{a}_{k-1}U_{|w|} + \cdots + \mathbf{a}_{k-1+|w|}U_0 \\ \iff c_0U_{|w|} + \cdots + (c_{|w|} - 1)U_0 &\leq \mathbf{a}_{k-1}U_{|w|} + \cdots + \mathbf{a}_{k-1+|w|}U_0, \end{aligned} \quad (9)$$

where we used Equation (1) to develop $U_{|w|+1}$ (recall that $|w| + 1 \leq \ell|w| + |x| = k - 1$). Notice that $\mathbf{a}[k - 1, +\infty) = yw^\omega$, so $\mathbf{a}[k - 1, k - 1 + |w|] = yxy_0$. On the other hand, $c_0 \cdots c_{|w|-1}(c_{|w|} - 1) = w(w_0 - 1)$ (recall that $w_0 \geq 1$). As w is anti-Lyndon, we have $w > yx$ so $w(w_0 - 1) > yxy_0$. To obtain a contradiction, let us go back to numeration systems.

For all $i \leq |w| + 1 \leq k + 1$, the length- i suffix v of yxy_0 (resp., $w(w_0 - 1)$) is (resp., is smaller than) a factor of \mathbf{a} , so, since w is anti-Lyndon, $v \leq \mathbf{a}[0, i) = \mathbf{d}^*[0, i)$ by Lemma 33. Therefore, by Lemma 14 and Proposition 34, yxy_0 (resp., $w(w_0 - 1)$) is in the numeration language of \mathcal{S}_c . By Theorem 9, \mathcal{S}_c respects the genealogical order,

and by Remark 8, the valuation is given by the sequence $(U_n)_{n \geq 0}$. Therefore, the word inequality $w(w_0 - 1) > yxy_0$ implies the (integer) inequality

$$c_0 U_{|w|} + \cdots + c_{|w|-1} + (c_{|w|} - 1)U_0 > \mathbf{a}_{k-1} U_{|w|} + \cdots + \mathbf{a}_{k-1+|w|} U_0,$$

which contradicts Inequality (9). This ends the proof that, if $P_n - 1 \leq Q_{n-1}$ for all n , then $c_{k-1} = 1$ and $c_0 \cdots c_{k-2}$ is an integer power of w . \square

We immediately obtain the next result.

Corollary 39. *Let c satisfy (WH). If $c = w^\ell 1$ for some $\ell \in \mathbb{N} \setminus \{0\}$ and some anti-Lyndon word w , then*

$$s_{\mathbf{u}}(m) = \begin{cases} i + 1, & \text{if } m \in [U_i, U_{i+1}) \text{ with } 0 \leq i \leq k - 2; \\ k, & \text{if } m \geq U_{k-1}. \end{cases}$$

Remark 40. *We note that the conditions of Proposition 38 are precisely those of [41, Theorem 1.1.], which characterizes the words \mathbf{u}_c for which there exists a simple Parry number β such that $d_\beta(1) = c0^\omega$ and the factor complexity function of \mathbf{u}_c is affine.*

However, Proposition 38 does not ban the existence of other parameters c for which a string attractor of minimal size can always be given by some Γ_n . This follows from the next remark.

Remark 41. *The proof of Proposition 20 shows that the factor $\mathbf{u}[U_n, Q_n]$ does not appear before. Therefore Γ_n cannot be a string attractor of $\mathbf{u}[0, m)$ if $m \geq Q_n + 1$; in other words the upper bound Q_n is tight in Theorem 23. However, the lower bound P_n is not necessarily tight. For example, if $c = 23$, then $\Gamma_2 = \{3, 9\}$ is a string attractor of the length-9 prefix $\mathbf{u}[0, 9) = 001001000$, while $P_2 = 10$. This is also the case for the k -bonacci morphisms ($c = 1^k$) where better bounds are provided in [10].*

On the other hand, there exist parameters c satisfying the conditions of Proposition 34 but for which Γ_n is sometimes not sufficient. The simplest such example is the *period-doubling word* corresponding to $c = 12$. Indeed, the length-8 prefix is given by 01000101 and the first few elements of the sequence $(U_n)_{n \geq 0}$ are 1, 2, 4, 8. One then easily checks that none of $\Gamma_1 = \{1, 2\}$, $\Gamma_2 = \{2, 4\}$, and $\Gamma_3 = \{4, 8\}$ is a string attractor. However, this word admits the size-2 string attractors $\{3, 6\}$ and $\{4, 6\}$ for example. In fact, Schaeffer and Shallit [4] proved that, if \mathbf{u} is the period-doubling word,

$$s_{\mathbf{u}}(n) = \begin{cases} 1, & \text{if } n = 1; \\ 2, & \text{if } n \geq 2. \end{cases}$$

Based on this observation and our experiments, we state the following conjecture.

Conjecture 42. *Let c satisfy (WH) and assume that one of the assertions of Proposition 34 holds. Then,*

$$s_{\mathbf{u}}(m) = \begin{cases} i + 1, & \text{if } m \in [U_i, U_{i+1}) \text{ with } 0 \leq i \leq k - 2; \\ k, & \text{if } m \geq U_{k-1}. \end{cases}$$

Observe that, using the first part of Corollary 37, this equality is known to be true for all $m \leq U_k - 1$ and for infinitely many values of m .

Acknowledgments. We warmly thank É. Charlier, S. Kreczman, and M. Rigo for useful discussions on numeration systems, especially the last two indicated [37] and [34] respectively.

Giuseppe Romana is partly supported by MUR PRIN project no. 2022YRB97K - 'PINC' (Pangenome INformatiCs: From Theory to Applications). Manon Stipulanti is an FNRS Research Associate supported by the Research grant 1.C.104.24F.

References

- [1] Kempa, D., Prezza, N.: At the roots of dictionary compression: string attractors. In: STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 827–840. ACM, New York (2018). <https://doi.org/10.1145/3188745.3188814>
- [2] Christiansen, A.R., Ettienne, M.B., Kociumaka, T., Navarro, G., Prezza, N.: Optimal-time dictionary-compressed indexes. *ACM Trans. Algorithms* **17**(1), 8–1839 (2021)
- [3] Kociumaka, T., Navarro, G., Prezza, N.: Toward a definitive compressibility measure for repetitive sequences. *IEEE Trans. Inform. Theory* **69**(4), 2074–2092 (2023) <https://doi.org/10.1109/tit.2022.3224382>
- [4] Schaeffer, L., Shallit, J.: String attractor for automatic sequences. Preprint available at <https://arxiv.org/abs/2012.06840> (2022)
- [5] Kutsukake, K., Matsumoto, T., Nakashima, Y., Inenaga, S., Bannai, H., Takeda, M.: On repetitiveness measures of Thue-Morse words. In: String Processing and Information Retrieval. *Lecture Notes in Comput. Sci.*, vol. 12303, pp. 213–220. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59212-7_15
- [6] Dolce, F.: String attractors for factors of the Thue-Morse word. In: Combinatorics on Words. *Lecture Notes in Comput. Sci.*, vol. 13899, pp. 117–129. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-33180-0_9
- [7] Mantaci, S., Restivo, A., Romana, G., Rosone, G., Sciortino, M.: A combinatorial view on string attractors. *Theoret. Comput. Sci.* **850**, 236–248 (2021) <https://doi.org/10.1016/j.tcs.2020.11.006>
- [8] Restivo, A., Romana, G., Sciortino, M.: String attractors and infinite words. In: LATIN 2022: Theoretical Informatics. *Lecture Notes in Comput. Sci.*, vol. 13568, pp. 426–442. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20624-5_26

- [9] Dvořáková, Ľ.: String attractors of episturmian sequences. *Theoret. Comput. Sci.* **986**, 114341 (2024) <https://doi.org/10.1016/j.tcs.2023.114341>
- [10] Cassaigne, J., Gheeraert, F., Restivo, A., Romana, G., Sciortino, M., Stipulanti, M.: New string attractor-based complexities on infinite words. Submitted work. Preprint available at <https://hdl.handle.net/2268/309672> (2024)
- [11] Dvořáková, Ľ., Hendrychová, V.: String attractors of Rote sequences. Preprint available at <https://arxiv.org/abs/2308.00850> (2023)
- [12] Béaur, P., Gheeraert, F., Menibus, B.: String attractors and bi-infinite words. Preprint available at <https://arxiv.org/abs/2403.13449> (2024)
- [13] Coven, E.M., Hedlund, G.A.: Sequences with minimal block growth. *Math. Systems Theory* **7**, 138–153 (1973) <https://doi.org/10.1007/BF01762232>
- [14] Rigo, M., Salimov, P.: Another generalization of abelian equivalence: binomial complexity of infinite words. *Theoret. Comput. Sci.* **601**, 47–57 (2015) <https://doi.org/10.1016/j.tcs.2015.07.025>
- [15] Fabre, S.: Substitutions et β -systèmes de numération. *Theoret. Comput. Sci.* **137**(2), 219–236 (1995) [https://doi.org/10.1016/0304-3975\(95\)91132-A](https://doi.org/10.1016/0304-3975(95)91132-A)
- [16] Bertrand-Mathis, A.: Comment écrire les nombres entiers dans une base qui n'est pas entière. *Acta Math. Hungar.* **54**(3-4), 237–241 (1989) <https://doi.org/10.1007/BF01952053>
- [17] Thurston, W.P.: Groups, tilings, and finite state automata summer 1989 AMS colloquium lectures (version 1.5, July 20, 1989). In: *Collected Works of William P. Thurston with Commentary. Vol. II. 3-manifolds, Complexity and Geometric Group theory*, pp. 553–602. Amer. Math. Soc., Providence, RI (2022). 1989 preprint
- [18] Parry, W.: On the β -expansions of real numbers. *Acta Math. Acad. Sci. Hungar.* **11**, 401–416 (1960) <https://doi.org/10.1007/BF02020954>
- [19] Burdík, v., Frougny, C., Gazeau, J.P., Krejcar, R.: Beta-integers as natural counting systems for quasicrystals. *J. Phys. A* **31**(30), 6449–6472 (1998) <https://doi.org/10.1088/0305-4470/31/30/011>
- [20] Frougny, C., Masáková, Z., Pelantová, E.: Complexity of infinite words associated with beta-expansions. *Theor. Inform. Appl.* **38**(2), 163–185 (2004) <https://doi.org/10.1051/ita:2004009>
- [21] Ambrož, P., Masáková, Z., Pelantová, E., Frougny, C.: Palindromic complexity of infinite words associated with simple Parry numbers. vol. 56, pp. 2131–2160 (2006). *Numération, pavages, substitutions*. http://aif.cedram.org/item?id=AIF_2006__56-7_2131_0

- [22] Turek, O.: Abelian properties of Parry words. *Theoret. Comput. Sci.* **566**, 26–38 (2015) <https://doi.org/10.1016/j.tcs.2014.11.024>
- [23] Frougny, C.: On the sequentiality of the successor function. *Inform. and Comput.* **139**(1), 17–38 (1997) <https://doi.org/10.1006/inco.1997.2650>
- [24] Shallit, J.: *The Logical Approach to Automatic Sequences: Exploring Combinatorics on Words with Walnut*. London Mathematical Society Lecture Note Series, vol. 482, p. 358. Cambridge University Press, Cambridge (2023). <https://doi.org/10.1017/9781108775267>
- [25] Gheeraert, F., Romana, G., Stipulanti, M.: String attractors of fixed points of k -Bonacci-like morphisms. In: *Combinatorics on Words. Lecture Notes in Comput. Sci.*, vol. 13899, pp. 192–205. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-33180-0_15
- [26] Lothaire, M.: *Combinatorics on Words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1997). <https://doi.org/10.1017/CBO9780511566097>. Corrected reprint of the 1983 original,
- [27] Chen, K.-T., Fox, R.H., Lyndon, R.C.: Free differential calculus. IV. The quotient groups of the lower central series. *Ann. of Math. (2)* **68**, 81–95 (1958) <https://doi.org/10.2307/1970044>
- [28] Reutenauer, C.: Mots de Lyndon généralisés. *Sém. Lothar. Combin.* **54**, 54–16 (2005/07)
- [29] Gewurz, D.A., Merola, F.: Numeration and enumeration. *European J. Combin.* **33**(7), 1547–1556 (2012) <https://doi.org/10.1016/j.ejc.2012.03.017>
- [30] Bonizzoni, P., De Felice, C., Zaccagnino, R., Zizza, R.: Inverse Lyndon words and inverse Lyndon factorizations of words. *Adv. in Appl. Math.* **101**, 281–319 (2018) <https://doi.org/10.1016/j.aam.2018.08.005>
- [31] Charlier, E., Philibert, M., Stipulanti, M.: Nyldon words. *J. Combin. Theory Ser. A* **167**, 60–90 (2019) <https://doi.org/10.1016/j.jcta.2019.04.002>
- [32] Rigo, M.: *Formal Languages, Automata and Numeration Systems. 2. Applications to Recognizability and Decidability*. Networks and Telecommunications Series. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ (2014)
- [33] Berthé, V., Rigo, M. (eds.): *Combinatorics, Automata and Number Theory*. Encyclopedia of Mathematics and its Applications, vol. 135. Cambridge University Press, Cambridge (2010). <https://doi.org/10.1017/CBO9780511777653>
- [34] Dumont, J.-M., Thomas, A.: Systèmes de numération et fonctions fractales relatifs aux substitutions. *Theoret. Comput. Sci.* **65**(2), 153–169 (1989) <https://doi.org/>

10.1016/0304-3975(89)90041-8

- [35] Lecomte, P.B.A., Rigo, M.: Numeration systems on a regular language. *Theory Comput. Syst.* **34**(1), 27–44 (2001) <https://doi.org/10.1007/s002240010014>
- [36] Charlier, E., Cisternino, C., Stipulanti, M.: A full characterization of Bertrand numeration systems. In: *Developments in Language Theory. Lecture Notes in Comput. Sci.*, vol. 13257, pp. 102–114. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05578-2_8
- [37] Hollander, M.: Greedy numeration systems and regularity. *Theory Comput. Syst.* **31**(2), 111–133 (1998) <https://doi.org/10.1007/s002240000082>
- [38] Duval, J.-P., Harju, T., Nowotka, D.: Unbordered factors and lyndon words. *Discrete Mathematics* **308**(11), 2261–2264 (2008) <https://doi.org/10.1016/j.disc.2006.09.054>
- [39] Duval, J.-P.: Factorizing words over an ordered alphabet. *J. Algorithms* **4**(4), 363–381 (1983) [https://doi.org/10.1016/0196-6774\(83\)90017-2](https://doi.org/10.1016/0196-6774(83)90017-2)
- [40] Duval, J.-P.: Mots de Lyndon et périodicité. *RAIRO Inform. Théor.* **14**(2), 181–191 (1980) <https://doi.org/10.1051/ita/1980140201811>
- [41] Bernat, J., Masáková, Z., Pelantová, E.: On a class of infinite words with affine factor complexity. *Theoret. Comput. Sci.* **389**(1-2), 12–25 (2007) <https://doi.org/10.1016/j.tcs.2007.07.001>