String attractors of fixed points of k-bonacci-like morphisms

France Gheeraert*[1], Manon Stipulanti†[1], and Giuseppe Romana‡[2]

[1] Department of Mathematics, University of Liège, Belgium
{france.gheeraert,m.stipulanti}@uliege.be
[2] Dipartimento di Matematica e Informatica, Università di Palermo, Italy
giuseppe.romana01@unipa.it

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 obtained as fixed points of k-bonacci-like morphisms. In fact, our description involves the numeration systems classically derived from the considered morphisms.

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

2020 Mathematics Subject Classification: Primary: 68R15. Secondary: 05A05, 11A67, 68P05, 68Q45.

1 Introduction

Introduced in the data compression field by Kempa and Prezza [18], the concept of *string attractor* can be conceptualized as follows: it is a set of positions within a finite word that enables to catch all distinct factors. String attractors also have applications in combinatorial pattern matching [24], but the problem of finding a smallest string attractor is NP-hard [18]. However, it appears that combinatorial properties of words yield new strategies to find a string attractor of minimum size. Consequently, string attractors have made their entry in combinatorics on words as a systematic topic of research. Indeed, they have been studied for prefixes of the ubiquitous Thue–Morse word [19, 28] and of the period-doubling word [28], while Sturmian words can be characterized through the structure of their smallest string attractors [22, 25].

The story of the current work began during the international conference DLT 2022, where the three authors had the chance to meet for the first time and where they talked about the concept of string attractors. Romana's expertise lead us to consider them for prefixes of generalized Fibonacci words to larger alphabets (on k letters, the corresponding word is called the *k-bonacci word*), as a natural extension of Sturmian words. It turned out that the string attractors that we obtained rely on the well-known k-bonacci numbers [16]. Simultaneously, Dvořáková studied

*France Gheeraert is a Research Fellow of the FNRS.
†Supported by the FNRS Research grant 1.B.397.20F.
‡Giuseppe Romana is partly supported by MIUR project PRIN 2017 ADASCOML – 2017K7XPAN.

string attractors of factors of episturmian words [12], which covers the case of all k-bonacci words. However, her description is less explicit.

Moreover, the fact that minimal string attractors of prefixes of the k-bonacci word can be described using the k-bonacci numbers tipped us on the probable link between string attractors and numeration systems, and lead us to believe that this bond can be adapted to other morphic sequences. More specifically, we have 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, as a first step towards answering 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 $0 \leq i \leq k-2$ and $\mu_c(k-1) = 0^{c_{k-1}}$. When it exists, we then look at the fixed point of this morphism. This family was not randomly chosen. First, it generalizes the k-bonacci morphisms but the fixed points are not necessarily episturmian. In addition, some of these morphisms have already been studied in relation to numeration systems, in [13] for example. Indeed, if c is some β-representation of 1 for a simple Parry number β, using the terminology of [5], we can canonically associate a numeration system that is greedy and, in this case, corresponds to the sequence $(|\mu_c^n(0)|)_{n\in\mathbb{N}}$ of lengths of iterations of μ_c on 0 [3]. Under some conditions on the parameters, we show that the prefixes of the fixed point admit string attractors of size at most $k+1$ described using the associated numeration system.

This paper is organized as follows. In Section 2, we recall some background on combinatorics on words. We also introduce the infinite words that we will study and give some of their basic properties. Section 3 introduces numeration systems and explains how to associate one with a morphic word. After that, we give conditions on the parameters $c_0, \ldots, c_{k-1}$ for the numeration system to have "desirable" properties. Finally, in Section 4, we look at string attractors and prove the main result of this paper, namely the description of string attractors of each prefix of the studied infinite words using the associated numeration system. We present concluding remarks and future work in Section 5.

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. We let 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 $\ell \in \mathbb{N}$ and x a prefix of w such that $v = w^\ell x$. We will then write $v = w^{|v|/|w|}$. 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, $u^\omega = uuu\cdots$.

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 upmost importance. See [21] for a presentation. A celebrated result in combinatorics on words is that London words form a so-called *complete factorization of the free monoid*.

Theorem 1 (Chen-Fox-Lyndon [7])**.** *For every non-empty word $w \in A^*$, there exists a unique factorization $(\ell_1, \cdots, \ell_n)$ of w into Lyndon words over A such that $\ell_1 \geq \ell_2 \geq \cdots \ell_n$.*

Several variations of Lyndon words have been considered lately: generalized Lyndon [26], anti-Lyndon [15], inverse Lyndon [4], and Nyldon [6]. In this text, we will use the second.

Definition 2. 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 3. 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 $f\colon A^* \to B^*$, 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 [27, Example 2.11].

Definition 4. Let $k \geq 2$ be an integer and let $c_0, \ldots, c_{k-1} \in \mathbb{N}$ be k parameters often summarized in the shape of a word $c = c_0 \cdots c_{k-1} \in \mathbb{N}^k$. The morphism $\mu_c : \{0, \ldots, k-1\}^* \to \{0, \ldots, k-1\}^*$ is given by $\mu_c(i) = 0^{c_i} \cdot (i+1)$ for all $i \in \{0, \ldots, k-2\}$ 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 $U_{c,n} = |u_{c,n}|$.

When the context is clear, we will usually omit the subscript c in Definition 4.

Example 5. 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 $\mu_c\colon 0 \mapsto 01, 1 \mapsto 2, 2 \mapsto 00$ 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 6. *For all $c = c_0 \cdots c_{k-1} \in \mathbb{N}^k$ and all $n \geq 0$, we have*

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

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

As a consequence of Proposition 6, the sequence $(U_n)_{n\in\mathbb{N}}$ respects the following recurrence relation: if $0 \leq n \leq k-1$, then $U_n = 1+\sum_{i=0}^{n-1} c_i U_{n-i-1}$, and if $n \geq k$, then $U_n = \sum_{i=0}^{k-1} c_i U_{n-i-1}$.

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

$$c = c_0 \cdots c_{k-1} \in \mathbb{N}^k \text{ with } c_0, c_{k-1} \geq 1. \tag{WH}$$

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 $\mathbf{u} := \lim_{n\to\infty} u_n$.

We make the following combinatorial observation.

Remark 7. Under (WH), using Proposition 6, a simple induction shows that the letter $1 \leq i \leq k-1$ can only be followed by 0 and/or $i+1$ (and only 0 in the case $i = k-1$) in $\mathbf{u}$.

3 Fun with numeration systems

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

A *numeration system* (for natural numbers) can be defined as a triple $\mathcal{S} = (A, \mathrm{rep}_{\mathcal{S}}, L)$, where A is an alphabet and $\mathrm{rep}_{\mathcal{S}} : \mathbb{N} \to A^*$ is an injective function such that $L = \mathrm{rep}_{\mathcal{S}}(\mathbb{N})$. The map $\mathrm{rep}_{\mathcal{S}}$ is called the *representation function* and L is the *numeration language*. If $\mathrm{rep}_{\mathcal{S}}(n) = w$ for some integer $n \in \mathbb{N}$ 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 $\mathrm{val}_{\mathcal{S}}(w) = n$. 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* [8]. 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 [8].

Proposition 8 (Dumont-Thomas [8])**.** *Let c satisfy* (WH)*. For all $n \in \mathbb{N}$, there exist unique integers $N, \ell_0, \ldots, \ell_N \in \mathbb{N}$ 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 the previous proposition 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 $\mathrm{rep}_{\mathcal{S}_c} : \mathbb{N} \to \mathbb{N}^*$ mapping n to $\mathrm{rep}_{\mathcal{S}_c}(n) = \ell_0 \cdots \ell_N$, where the integers $\ell_0, \ldots, \ell_N$ verify the conditions of Proposition 8 for n. By convention, we set $\mathrm{rep}_{\mathcal{S}_c}(0) = \varepsilon$.

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	$012\cdot 0$	01200	$01200\cdot 0$	$01200\cdot 01$	$01200\cdot 01\cdot 0$
$\mathrm{rep}_{\mathcal{S}_c}(n)$	ε	1	10	100	101	1000	1001	1010	1011

Example 9. Using Example 5 for $c = 102$, the representations of the first few integers are given 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 10. If $\mathrm{rep}_{\mathcal{S}_c}(n) = \ell_0 \cdots \ell_N$, then $n = |u_{c,N}^{\ell_0} \cdots u_{c,0}^{\ell_N}| = \sum_{i=0}^{N} \ell_i U_{c,N-i}$. In other words, $\mathrm{val}_{\mathcal{S}_c}$ is given by the usual valuation function associated with the sequence $(U_{c,n})_{n\in\mathbb{N}}$. 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 [20]. A numeration system $\mathcal{S} = (A, \mathrm{rep}, L)$ is said to be *abstract* if L is regular and $\mathrm{rep}(n)$ is the $(n+1)$st word of L in the genealogical order. We have the following result.

Theorem 11 (Rigo [27, Section 2.2])**.** *Let $\sigma : \{\alpha_0, \ldots, \alpha_d\}^* \to \{\alpha_0, \ldots, \alpha_d\}^*$ be a morphism prolongable on the letter α_0. We define the automaton $\mathcal{A}_\sigma$ for which $\{\alpha_0, \ldots, \alpha_d\}$ 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 $\alpha \in \{\alpha_0, \ldots, \alpha_d\}$ and $0 \le i \le |\sigma(\alpha)| - 1$, $\delta(\alpha, i)$ is the $(i+1)$st letter of $\sigma(\alpha)$. If $\mathcal{S} = (A, \mathrm{rep}, L)$ is the Dumont-Thomas numeration system associated with σ, then $L = L(\mathcal{A}_\sigma) \setminus 0\mathbb{N}^*$ and $\mathrm{rep}(n)$ is the $(n+1)$st word of L in the genealogical order.*

Example 12. For $c = 102$, the automaton $\mathcal{A}_{\mu_c}$ of Theorem 11 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 9.

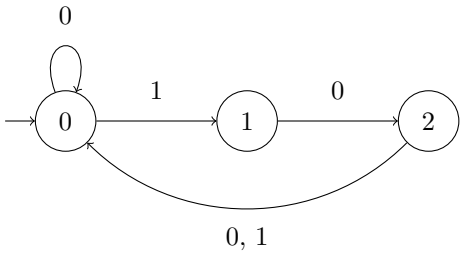

Figure 1: The automaton $\mathcal{A}_{\mu_c}$ for $c = 102$.

As the automaton in Theorem 11 can be used to produce, for all $n \ge 0$, the letter $\mathbf{u}_n$ when reading $\mathrm{rep}_{\mathcal{S}_c}(n)$ by [27, Theorem 2.24], we have the following.

Corollary 13. *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 $\mathbf{d}^\star$ denote the periodization of c, that is, $\mathbf{d}^\star = (c_0 \cdots c_{k-2}(c_{k-1} - 1))^\omega$. Using Theorem 11, we deduce the next result.

Lemma 14. *Under* (WH)*, for all $n \geq 0$, we have $\mathrm{rep}_{\mathcal{S}_c}(U_n) = 10^n$, the numbers having a representation of length $n+1$ are those in $[U_n, U_{n+1})$, and $\mathrm{rep}_{\mathcal{S}_c}(U_{n+1}-1) = \mathbf{d}^\star[0,n]$. In particular, $U_{n+1} - 1 = \sum_{i=0}^{n} \mathbf{d}_i^\star U_{n-i}$.*

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 $\mathcal{A}_{\mu_c}$, which is the length-$(n+1)$ prefix of $\mathbf{d}^\star$. □

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 [3, 5]. However, even though the function $\mathrm{rep}_{\mathcal{S}_c}$ 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 15. In Example 5 for $c = 102$, we see that $\mathbf{u}[0,14) = 012000101 \cdot 012 \cdot 01$, so $\mathrm{rep}_{\mathcal{S}_c}(14) = 10110$, while the greedy representation of 14 associated with the sequence $(U_n)_{n\in\mathbb{N}}$ is 11000.

In fact, we have the following two characterizations.

Lemma 16. *Let c satisfy* (WH)*. The numeration system $\mathcal{S}_c = (A, \mathrm{rep}_{\mathcal{S}_c}, L)$ 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 $\mathbf{d}^\star[0,i)$. Moreover, we then have*

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

Proof. Let us denote $\mathcal{S} = (A', \mathrm{rep}_{\mathcal{S}}, L')$ the canonical greedy numeration system associated with the sequence $(U_n)_{n\in\mathbb{N}}$. 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, $\mathrm{rep}_{\mathcal{S}_c}$ respects the genealogical order, i.e., $n \leq m$ if and only if $\mathrm{rep}_{\mathcal{S}_c}(n) \leq_{\mathrm{gen}} \mathrm{rep}_{\mathcal{S}_c}(m)$. So does $\mathrm{rep}_{\mathcal{S}}$ by [2, Proposition 2.3.45]. Hence, $\mathcal{S}_c = \mathcal{S}$ if and only if $L = L'$. Moreover, for all $n \geq 0$, $\mathrm{rep}_{\mathcal{S}}(U_n) = 10^n$, so L and L' contain the same number of length-n words by Lemma 14. Thus $L = L'$ if and only if $L \subseteq L'$. The statement holds since, by [17, Lemma 5.3] and by Lemma 14, we have

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

□

Theorem 17. *Let $c = c_0 \cdots c_{k-1} \in \mathbb{N}^k$ 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 16 and Theorem 11, $\mathcal{S}_c$ is greedy if and only if, for all $n \in \mathbb{N}$ and for all $0 \leq i \leq k-1$, any path $\ell_0 \cdots \ell_n$ starting in State i in the automaton $\mathcal{A}_{\mu_c}$ is such that $\ell_0 \cdots \ell_n \leq \mathbf{d}^\star[0,n]$. However, by definition of $\mathcal{A}_{\mu_c}$, the lexicographically biggest path of length n starting in state i is given by the prefix of length n of $(c_i \cdots c_{k-2}(c_{k-1}-1)c_0 \cdots c_{i-1})^\omega$. We can therefore 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. □

Observe that the condition of the previous result is equivalent to the fact that $c_0 \cdots c_{k-2}(c_{k-1}-1) = v^\ell$ for some anti-Lyndon v (in fact, v is the primitive root).

Example 18. Let $k = 4$ and $c = 1011$. In this case, $c_0c_1c_2(c_3-1) = 1010 = v^2$ with $v = 10$, which is anti-Lyndon (see Example 3). 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 $(U_n)_{n\in\mathbb{N}}$ is in fact the sequence of Fibonacci numbers. Therefore the numeration system $\mathcal{S}_c$ corresponds to the classical Fibonacci numeration system, which can also be obtained with the parameter $c = 11$.

The observation made in the previous example is more general.

Remark 19. Let c satisfy (WH). If $c_0 \cdots c_{k-2}(c_{k-1}-1) = v^\ell$ with v anti-Lyndon, we define the word $v' := v_1 \cdots v_{|v|-1}(v_{|v|}+1)$ (simply put, we add 1 to the last letter of v). Then $c = v^{\ell-1}v'$ is a "partial" cyclization of v'. In particular, since $\mathbf{d}^\star_c = \mathbf{d}^\star_{v'}$ (where the dependence of $\mathbf{d}^\star$ on the chosen parameters is emphasized via a subscript), the numeration systems $\mathcal{S}_c$ and $\mathcal{S}_{v'}$ coincide by Lemma 14.

For the reader familiar with the general theory of numerations, v' satisfies $v'_i \cdots v'_{|v|} < v'$ for all indices $i \in \{2, \ldots, |v|\}$. This implies that v' is the β-expansion $d_\beta(1)$ of 1 for a simple Parry number β [23]. Therefore, c is also a representation of 1 in base β.

Example 20. We illustrate the previous remark by resuming Example 18. We have $v = 10$ and $v' = 11$. The corresponding simple Parry number is the Golden ratio φ. Observe that indeed $c = vv' = 1011$ is a representation of 1 in base φ.

4 Link to string attractors

Using the results and concepts of the previous sections, we now turn to the concept of string attractors in relation to the fixed points of the morphisms μ_c, $c \in \mathbb{N}^k$. A *string attractor* of a finite word $y = y_1 \cdots y_n$ is a set $\Gamma \subseteq \{1, \ldots, n\}$ such that every factor of y has an occurrence crossing a position in Γ, i.e., for all factor $x \in A^m$ of y, there exists $i \in \Gamma$ and j such that $i \in \{j, \ldots, j+m-1\}$ and $x = y_j \cdots y_i \cdots y_{j+m-1}$.

Example 21. The set $\{2, 3, 4\}$ is a string attractor of the word $0\,\underline{1}\,\underline{2}\,\underline{0}\,0\,1$. 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 [18]. Where ambiguity may occur, we explicitly declare how finite words are indexed.

As we will look at prefixes of infinite words, it is natural to wonder if there is a link between the string attractors of the finite words w and wa, where a is a letter. In general, there is no trivial link although we have the following result which can be derived from the proofs of [22, Propositions 12 and 15].

Proposition 22. *Let z be a non-empty word and let $x = z^r$, $y = z^s$ be fractional powers of z with $1 \leq r \leq s$. If Γ is a string attractor of x, then $\Gamma \cup \{|z|\}$ is a string attractor of y.*

Since the considered infinite words are the limits of the sequence $(u_n)_{n\in\mathbb{N}}$, we are interested in the prefixes which are fractional powers of some u_n.

Definition 23. 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 $Q_n = |q_n|$.

4.1 Fractional power prefixes and anti-Lyndon words

In this subsection, we will prove that the words defined above have a particular structure related to (anti-)Lyndon words in Proposition 30. To do so, we introduce some notations. For all $n \geq 0$, the pair $\{i_n, j_n\}$ will designate the two (distinct) letters following q_n in $\mathbf{u}$ and in $(u_n)^\omega$. Without loss of generality, we will always assume that $i_n < j_n$.

Example 24. Set $c = 102$. Recall from Examples 5 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 $(i_n)_{n\in\mathbb{N}}$ and $(j_n)_{n\in\mathbb{N}}$, 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 $c_0 = \max\{c_0, \ldots, c_{j_n - 1}\}$ 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 $\mu(q_n) = u_n^{\ell_0} \cdots u_1^{\ell_n}$ by the induction hypothesis.

If $j_n = k-1$ then by Remark 7, 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

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

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 $\max\{c_0, \ldots, c_{j_{n+1}-1}\} \leq \max\{c_0, \ldots, c_{j_n - 1}\}$. Indeed, in all cases, either $j_{n+1} \leq j_n$, or $j_{n+1} = j_n + 1$ and $c_{j_n} \leq \max\{c_0, \ldots, c_{j_n - 1}\}$. □

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 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.

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	$c_0 = c_3$	$c_3 = 2$	$\{1, 4\}$

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 $c_0 \cdots c_{i_{n+1}-1} = c_{j_{n+1}-i_{n+1}} \cdots c_{j_{n+1}-1}$. □

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 [11].

Proposition 28. *Lyndon words are unbordered, i.e. if w is a 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 Theorem 1. See, for instance, [21, Theorem 5.1.5].

Proposition 29. *Let $w \in A^*$ be a non-empty word and let $(\ell_1, \cdots, \ell_n)$ 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 [10]. 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 1 (Duval [10])**.** Let $(A, \leq)$ be an ordered set and let $w = w_0 \dots w_n$ 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 $w_{i-j} \cdots w_n$.

Using the notation of the paragraph preceding Algorithm 1, 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 [9, 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 30. *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 $\ell_n = c_{n \bmod p}$, $j_n \equiv (n+1) \mod p$, 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 $N + 1 = p + i_N \equiv i_N \mod p$. Indeed, by Lemma 25 and by definition of N, we have $\ell_{N+1} = c_{i_N} = c_{N+1 \bmod p}$ and

$$j_{N+1} = i_N + 1 \equiv N + 2 \mod p. \tag{1}$$

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 $n + 1$ mod 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 $j_{n+1} \equiv n + 2 \mod p$.

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,

$$i_n \equiv i_N \mod p. \tag{2}$$

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

$$\ell_{n+1} = \ell_{N+1} \quad \text{and} \quad j_{n+1} = i_n + 1 \leq N+1. \tag{3}$$

By the induction hypothesis for n, we have

$$N+1 = j_n \equiv n+1 \mod p. \tag{4}$$

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 (3), the second by the induction hypothesis for $N+1$, and the last by Congruence (4), and

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

where the first equality follows from (3), the second congruence from (2), the third by (1), the fourth by the induction hypothesis for $N+1$, and the last by Congruence (4). This ends the proof. □

Example 31. Let us pursue Example 31 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 $c_0c_1 = 10$ is 10 itself so $\mathbf{a} = (10)^\omega$. We can easily check that the first few q_n's indeed satisfy Proposition 30.

4.2 String attractors of the prefixes

Motivated by Proposition 22, 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 the main theorem. However, we first have to ensure that this interval is well defined. For that, we will need the following lemma.

Lemma 32. *Let c satisfy* (WH). *Then $c_0 \cdots c_{k-2} \geq \mathbf{a}[0, k-2]$.*

Proof. Assume the contrary and let w be the longest anti-Lyndon prefix of $c_0 \cdots c_{k-2}$. If $|w| \leq i \leq k-2$ is the smallest index such that $c_0 \cdots c_i < \mathbf{a}[0, i]$, then $c_0 \cdots c_i = w^\ell va$ with v a proper prefix of w, a a letter, and $va < w$. So [9, 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 definition (maximality) of w. □

In fact, the condition obtained for the greediness of the numeration system is related to the relation between $U_n - 1$ and Q_{n-1}. This is detailed in the next two results.

Proposition 33. *Let c satisfy* (WH). *If $c_0 \cdots c_{k-2}(c_{k-1}-1)$ is lexicographically maximal among its conjugates, then $\mathbf{d}^\star[0, n] \leq \mathbf{a}[0, n]$ for all $n \geq 0$.*

Proof. Let w denote the longest anti-Lyndon prefix of $c_0 \cdots c_{k-2}$. 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 $c_0 \cdots c_{k-2}(c_{k-1}-1) \leq \mathbf{a}[0, k-1]$. By Lemma 32, we get $c_0 \cdots c_{k-2} = \mathbf{a}[0, k-2]$ and $c_{k-1} - 1 \leq \mathbf{a}_{k-1}$.

We now prove that $\mathbf{d}^\star[0,n] \leq \mathbf{a}[0,n]$ for all $n \geq 0$. If $c_{k-1}-1 < \mathbf{a}_{k-1}$, then the conclusion is direct. 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 so there exist $\ell \geq 1$ and u a proper prefix of w such that $c_0\cdots c_{k-2}(c_{k-1}-1) = w^\ell u$. Let us write $w = uv$. If $u \neq \varepsilon$, we then have

$$c_0\cdots c_{k-2}(c_{k-1}-1) = w^\ell u = u(vu)^\ell < uw^\ell$$

as w is anti-Lyndon thus strictly greater than its conjugates. This contradicts the assumption that $c_0\cdots c_{k-2}(c_{k-1}-1)$ is maximal among its conjugates. Therefore, $u = \varepsilon$ and $c_0\cdots c_{k-2}(c_{k-1}-1)$ is a (natural) power of w. We conclude that $\mathbf{a} = \mathbf{d}^\star$, which ends the proof of the first item. □

Proposition 34. *Let c satisfy* (WH). *If $c_0\cdots c_{k-2}(c_{k-1}-1)$ is lexicographically maximal among its conjugates, then $U_{n+1}-1 \leq Q_n$ for all $n \geq 0$.*

Proof. Let us show the claim by contraposition. So assume that there exists an integer n such that $U_{n+1}-1 > Q_n$. Thus $q_n = u_n^{\mathbf{a}_0}\cdots u_0^{\mathbf{a}_n}$ is a proper prefix of $\mathbf{u}[0,U_{n+1}-1)$. By Lemma 14, $\mathrm{rep}_{\mathcal{S}_c}(U_{n+1}-1) = \mathbf{d}^\star[0,n]$, so $\mathbf{d}_0^\star$ is the largest exponent e such that u_n^e is a prefix of $\mathbf{u}[0,U_{n+1}-1)$. This implies that $\mathbf{d}_0^\star \geq \mathbf{a}_0$. Moreover, if $\mathbf{a}_0 = \mathbf{d}_0^\star$, the same argument implies that $\mathbf{d}_1^\star$ is the largest exponent e such that $u_n^{\mathbf{d}_0^\star}u_{n-1}^e$ is a prefix of $\mathbf{u}[0,U_{n+1}-1)$. In both cases, we have $\mathbf{d}_0^\star\mathbf{d}_1^\star \geq \mathbf{a}_0\mathbf{a}_1$. We may iterate the reasoning to obtain $\mathbf{d}^\star[0,n] \geq \mathbf{a}[0,n]$. As q_n is a proper prefix of $\mathbf{u}[0,U_{n+1}-1)$, the inequality cannot be an equality. This contradicts Proposition 33, which ends the proof. □

We will now prove that, under the conditions of the previous result, we can describe string attractors of every prefix of $\mathbf{u}$ using the elements of $(U_n)_{n\in\mathbb{N}}$. For $n \in \mathbb{N}$, we denote

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

We also define

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

The next lemma directly follows from Proposition 34 and the definition of P_n.

Lemma 35. *Let c satisfy* (WH). *If $c_0\cdots c_{k-2}(c_{k-1}-1)$ is maximal among its conjugates, then $P_n \leq U_{n+1}-1 \leq Q_n$ for all $n \in \mathbb{N}$.*

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

Theorem 36. *Let $c = c_0\cdots c_{k-1} \in \mathbb{N}^k$ with $c_0, c_{k-1} \geq 1$ and $c_0\cdots c_{k-2}(c_{k-1}-1)$ maximal among its conjugates. Fix an integer $n \geq 0$. If $m \in [U_n, Q_n]$, then $\Gamma_{n-1}\cup\{U_n\}$ is a string attractor of $\mathbf{u}[0,m)$. Furthermore, 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 35 and 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 22 and by definition of Q_n (Definition 23), 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 22, 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 $\mathbf{u}[0, P_n)$ that have 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

$$|w| = P_n - U_n = U_{n-k+1} - U_{n-k} - 1 \tag{5}$$

by definition of P_n, so $|w'| = |w|$. We will actually show that $w' = w$.

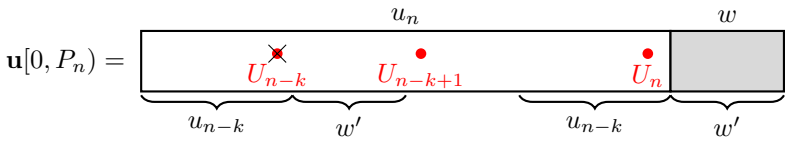

Figure 2: Representation of the proof of the second claim of Theorem 36. As we warned the reader before, elements in a string attractor are indexed starting at 1 (in red), while indices of letters in $\mathbf{u}$ start at 0.

The fact that u_{n-k} is a suffix of u_n is a direct consequence of Proposition 6 as $c_{k-1} \geq 1$ by assumption. To prove that $w' = w$, we first make the following observation: Proposition 6 again implies that u_n is followed by $u_n^{c_0-1} u_{n-1}^{c_1} \cdots u_{n-k+1}^{c_{k-1}}$ in $\mathbf{u}$. Since u_{n-k+1} is a prefix of all the words $u_{n-k+1}, \ldots, u_{n-1}$, the word u_n is in particular followed by u_{n-k+1} in $\mathbf{u}$. As $|w| \leq U_{n-k+1}$ by Equation (5), this implies that w is a prefix of u_{n-k+1}, so also of $\mathbf{u}$. To conclude with the claim, it is then enough to show that w' is also a prefix of $\mathbf{u}$. To prove this, we will use the numeration system $\mathcal{S}_c$ and consider two cases.

First, assume that $n - 2k + 1 \geq 0$. By definition of w' and by Proposition 6, w' is a prefix of $v := u_{n-k}^{c_0-1} \cdots u_{n-2k+1}^{c_{k-1}}$. Define the word $x = (c_0 - 1)c_1 \cdots c_{k-1} 0^{n-2k+1}$. If it begins with 0's, we consider instead the word obtained by removing the leading 0's. Note that x corresponds to a factorization of v into the words $u_{n-k}, \ldots, u_0$. As $c_0 \cdots c_{k-2}(c_{k-1} - 1)$ is maximal among its conjugates by assumption, x is in the numeration language by Lemma 16. By definition of $\mathcal{S}_c$, x is the Dumont-Thomas factorization of v, implying that v is a prefix of $\mathbf{u}$.

Second, if $n-2k+1 < 0$, then we conclude in a similar way by considering $v = u_{n-k}^{c_0-1} \cdots u_0^{c_{n-k}}$ and $x = (c_0 - 1)c_1 \cdots c_{n-k}$ instead. □

As a consequence, for some values of c ($c = 211$ for example), we can say that every prefix of $\mathbf{u}$ has a string attractor of size at most k. Moreover, these string attractors are optimal as every position in Γ_n covers a different letter (this can be proved using a simple induction).

Observe that the bounds given in the previous theorem are 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 [16].

5 Final comments

We end this paper by discussing the scope of use of our main result. For a given $c \in \mathbb{N}^k$ satisfying specific properties, Theorem 36 states that we can easily describe a string attractor of size at most $k+1$ for any prefix of the fixed point $\mathbf{u}$ of μ_c defined in Section 2.2 and that, in some cases, we can even lower its size to k. On the one hand, this result is not necessarily optimal. For

example, if $c = 12$, the corresponding fixed point is referred to as the *period-doubling* word [1]. Our result particularly implies that $\{2, 4, 8\}$ is a string attractor of its length-8 prefix 01000101 and we can check that every position is needed. However, in [28], the authors proved that we can find a string attractor of size 2, namely, $\{3, 6\}$. On the other hand, for some $c \in \mathbb{N}^k$, 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 $c = 3203$ [14]. As a consequence, the approach from [28] does not apply; in particular, we study words outside the framework needed to use the software *Walnut* [29].

Finally, we wish to point out that this paper is a first exploration into the possible link between string attractors of prefixes of morphic words and general numeration systems. As stated in the Question presented Section 1, we believe that this connection can be extended to other morphisms, which is a path that we will continue exploring in the future.

Acknowledgements

We warmly thank M. Rigo and S. Kreczman for useful discussions on numeration systems, especially for indicating [8] and [17] respectively.

References

[1] J.-P. Allouche and J. Shallit. *Automatic sequences*. Cambridge University Press, Cambridge, 2003. Theory, applications, generalizations. doi:10.1017/CBO9780511546563.

[2] Valérie Berthé and Michel Rigo, editors. *Combinatorics, automata and number theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010. doi:10.1017/CBO9780511777653.

[3] A. Bertrand-Mathis. Comment écrire les nombres entiers dans une base qui n'est pas entière. *Acta Math. Hungar.*, 54(3-4):237–241, 1989. doi:10.1007/BF01952053.

[4] P. Bonizzoni, C. De Felice, R. Zaccagnino, and R. Zizza. Inverse Lyndon words and inverse Lyndon factorizations of words. *Adv. in Appl. Math.*, 101:281–319, 2018. doi:10.1016/j.aam.2018.08.005.

[5] É. Charlier, C. Cisternino, and M. Stipulanti. A full characterization of Bertrand numeration systems. In *Developments in language theory*, volume 13257 of *Lecture Notes in Comput. Sci.*, pages 102–114. Springer, Cham, 2022. doi:10.1007/978-3-031-05578-2_8.

[6] É. Charlier, M. Philibert, and M; Stipulanti. Nyldon words. *J. Combin. Theory Ser. A*, 167:60–90, 2019. doi:10.1016/j.jcta.2019.04.002.

[7] K.-T. Chen, R. H. Fox, and R. C. Lyndon. Free differential calculus. IV. The quotient groups of the lower central series. *Ann. of Math. (2)*, 68:81–95, 1958. doi:10.2307/1970044.

[8] J.-M. Dumont and A. Thomas. Systèmes de numération et fonctions fractales relatifs aux substitutions. *Theoret. Comput. Sci.*, 65(2):153–169, 1989. doi:10.1016/0304-3975(89)90041-8.

[9] J.-P. Duval. Mots de Lyndon et périodicité. *RAIRO Inform. Théor.*, 14(2):181–191, 1980. doi:10.1051/ita/1980140201811.

[10] J.-P. Duval. Factorizing words over an ordered alphabet. *J. Algorithms*, 4(4):363–381, 1983. doi:10.1016/0196-6774(83)90017-2.

[11] J.-P. Duval, T. Harju, and D. Nowotka. Unbordered factors and lyndon words. *Discrete Mathematics*, 308(11):2261–2264, 2008. doi:https://doi.org/10.1016/j.disc.2006.09.054.

[12] L. Dvořáková. String attractors of episturmian sequence, 2022. preprint available at arXiv:2211.01660.

[13] S. Fabre. Substitutions et β-systèmes de numération. *Theoret. Comput. Sci.*, 137(2):219–236, 1995. doi:10.1016/0304-3975(95)91132-A.

[14] C. Frougny. On the sequentiality of the successor function. *Inform. and Comput.*, 139(1):17–38, 1997. doi:10.1006/inco.1997.2650.

[15] D. A. Gewurz and F. Merola. Numeration and enumeration. *European J. Combin.*, 33(7):1547–1556, 2012. doi:10.1016/j.ejc.2012.03.017.

[16] F. Gheeraert, A. Restivo, G. Romana, M. Sciortino, and M. Stipulanti. String attractors and infinite words, 2023. Work in progress; preliminary version available at arXiv:2206.00376.

[17] M. Hollander. Greedy numeration systems and regularity. *Theory Comput. Syst.*, 31(2):111–133, 1998. doi:10.1007/s002240000082.

[18] Dominik Kempa and Nicola Prezza. At the roots of dictionary compression: string attractors. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 827–840. ACM, New York, 2018. doi:10.1145/3188745.3188814.

[19] Kanaru Kutsukake, Takuya Matsumoto, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. On repetitiveness measures of Thue-Morse words. In *String processing and information retrieval*, volume 12303 of *Lecture Notes in Comput. Sci.*, pages 213–220. Springer, Cham, 2020. doi:10.1007/978-3-030-59212-7_15.

[20] P. B. A. Lecomte and M. Rigo. Numeration systems on a regular language. *Theory Comput. Syst.*, 34(1):27–44, 2001. doi:10.1007/s002240010014.

[21] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997. Corrected reprint of the 1983 original,. doi:10.1017/CBO9780511566097.

[22] S. Mantaci, A. Restivo, G. Romana, G. Rosone, and M. Sciortino. A combinatorial view on string attractors. *Theoret. Comput. Sci.*, 850:236–248, 2021. doi:10.1016/j.tcs.2020.11.006.

[23] W. Parry. On the β-expansions of real numbers. *Acta Math. Acad. Sci. Hungar.*, 11:401–416, 1960. doi:10.1007/BF02020954.

[24] Christiansen A. R., Ettienne M. B., Kociumaka T., Navarro G., and Prezza N. Optimal-time dictionary-compressed indexes. *ACM Trans. Algorithms*, 17(1):8:1–8:39, 2021.

[25] A. Restivo, G. Romana, and M. Sciortino. String attractors and infinite words. In *LATIN 2022: Theoretical informatics*, volume 13568 of *Lecture Notes in Comput. Sci.*, pages 426–442. Springer, Cham, 2022. doi:10.1007/978-3-031-20624-5_26.

[26] C. Reutenauer. Mots de Lyndon généralisés. *Sém. Lothar. Combin.*, 54:Art. B54h, 16, 2005/07.

[27] M. Rigo. *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.

[28] L. Schaeffer and J. Shallit. String attractor for automatic sequences, 2022. preprint available at arXiv:2012.06840.

[29] Jeffrey Shallit. *The logical approach to automatic sequences. Exploring combinatorics on words with Walnut*, volume 482 of *Lond. Math. Soc. Lect. Note Ser.* Cambridge: Cambridge University Press, 2023. doi:10.1017/9781108775267.