New string attractor-based complexities for infinite words

Julien Cassaigne[a], France Gheeraert[b,1], Antonio Restivo[c], Giuseppe Romana[c,2], Marinella Sciortino[c,2], Manon Stipulanti[b,3]

[a] *CNRS, Aix-Marseille Univ, I2M, France*
[b] *Department of Mathematics, University of Liège, Belgium*
[c] *Department of Mathematics and Computer Science, University of Palermo, Italy*

Abstract

A *string attractor* is a set of positions in a word such that each distinct factor has an occurrence crossing a position from the set. This definition comes from the field of data compression, where the size γ^* of a smallest string attractor represents a lower bound for the output size of a wide family of string compressors exploiting repetitions in words, including BWT-based and LZ-based compressors. On finite words, the combinatorial properties of string attractors have been studied in 2021 by Mantaci et al. Later, Schaeffer and Shallit introduced the *string attractor profile function*, a complexity function which evaluates for each $n > 0$ the size γ^* of the length-n prefix of a one-sided infinite word.

A natural development of the research on the topic is to link string attractors with other classical notions of repetitiveness in combinatorics on words. Our contribution in this sense is threefold. First, we explore the relation between the string attractor profile function and other well-known combinatorial complexity functions in the context of infinite words, such as the factor complexity and the property of recurrence. Moreover, we study its asymptotic growth in the case of purely morphic words and obtain a complete description in the binary case. Second, we introduce two new string attractor-based complexity functions, in which the structure and the distribution of positions in a string attractor are taken into account, and we study their combinatorial properties. We also show that these measures provide a finer classification of some infinite families of words, namely the Sturmian and quasi-Sturmian words. Third, we explicitly give the three complexities for some specific morphic words called k-bonacci words.

A preliminary version of some results presented in this paper can be found in [Restivo, Romana, Sciortino, *String Attractors and Infinite Words*, LATIN 2022].

Keywords: String attractor, factor complexity, recurrence function, repetitiveness measure, Sturmian word, k-bonacci word
2020 MSC: 68R15, 05A05, 68Q45

Email addresses: julien.cassaigne@math.cnrs.fr (Julien Cassaigne), france.gheeraert@uliege.be (France Gheeraert), antonio.restivo@unipa.it (Antonio Restivo), giuseppe.romana01@unipa.it (Giuseppe Romana), marinella.sciortino@unipa.it (Marinella Sciortino), m.stipulanti@uliege.be (Manon Stipulanti)

[1] F. Gheeraert is a Research Fellow of the Fonds de la Recherche Scientifique – FNRS.
[2] M. Sciortino and G. Romana are partly supported by MUR project PRIN 2022 PINC – 2022YRB97K.
[3] M. Stipulanti is supported by the FNRS Research grant 1.C.104.24F.

1. Introduction

Repetitiveness is a central notion in the field of Combinatorics on Words, which has been approached from various perspectives. For instance, the *factor complexity function* is probably the most extensively studied repetitiveness measure [7]. For an infinite word $\mathbf{x}$, its factor complexity function $p_{\mathbf{x}}$ counts, for each $n \geq 0$, the number of distinct factors of length n. Intuitively, the lower the factor complexity, the more repetitive the infinite word. Indeed, a famous theorem by Morse and Hedlund characterizes the words with (eventually) constant factor complexity as being *eventually periodic*, i.e. obtained by repeating the same factor, starting after a certain finite prefix. Within the sphere of infinite aperiodic words, some of the most studied words are the *Sturmian words*, which are the infinite aperiodic words with the lowest factor complexity function, i.e. their factor complexity is $n+1$ for every n. *Quasi-Sturmian words* represent the simplest generalization of Sturmian words in terms of factor complexity, they are infinite words having factor complexity $n + d$, with $d \geq 1$, for every large enough n.

The analysis of repetitiveness in words can also be conducted using the recurrence function. It is another powerful measure that, in a complementary way, unveils the repetitive structure of infinite words. This notion was initially defined by Morse and Hedlund [32] but has found widespread recognition in the literature. See [6] for a survey. An infinite word $\mathbf{x}$ is *recurrent* if every factor of $\mathbf{x}$ occurs infinitely often. The recurrence function $R_{\mathbf{x}}$ for an infinite word $\mathbf{x}$ gives, for each $n \geq 0$, if it exists, the size of the smallest window containing all the length-n factors of $\mathbf{x}$, no matter where this window is located in $\mathbf{x}$. Intuitively, it is closely related to the maximum gap between two consecutive occurrences of any length-n factor. Essentially, it provides an idea of how quickly factors repeat within an infinite word and how distributed the repetitive elements are in the word. If $R_{\mathbf{x}}(n)$ is defined for all n, then the word is called *uniformly recurrent*, and if $R_{\mathbf{x}}$ is linear, then $\mathbf{x}$ is called *linearly recurrent*.

In application contexts, repetitiveness has recently become a fundamental concept that is gaining increasing relevance [34]. Due to the abundance of highly repetitive data and the need to manage them efficiently, being able to effectively evaluate and measure the repetitiveness of data is fundamental to optimize processes and resources. For instance, in the realm of indexing massive text collections, defining data structures that enable querying data using space proportional to the size of compressed data becomes crucial [35]. In such a scenario, finding good measures capable of capturing the level of repetitiveness in a text is strongly related to having effective parameters to evaluate the performance of such compressed data structures, both in terms of space and time. For this reason, the most commonly used measures in this field stem from compression schemes, such as the number of phrases in the LZ77 parsing and the number of equal letter runs produced by the Burrows-Wheeler Transform [36].

With the aim of unifying existing compressor-based measures, Kempa and Prezza proposed in [23] a repetitiveness measure related to combinatorial properties of the text instead of being associated with a specific compressor. A *string attractor* Γ for a text w is a set of positions in w such that each factor of w has an occurrence crossing some position in Γ. Intuitively, the more repetitive the text, the lower the number of positions needed in a string attractor. The measure $\gamma^*(w)$ is then the minimal size of a string attractor for w. On the one hand, it has been proven that γ^* is a lower bound for all other usual compressor-based repetitiveness measures. On the other hand, finding the smallest attractor size γ^* for a given text w is an NP-complete problem.

Recently much interest has been aroused by the combinatorial properties of string attractors. Firstly, in [30] the sensitivity of the measure γ^* with respect to the combinatorial operations on finite words has been studied. In particular, it has been shown that γ^* is not monotone, in the

sense that the measure γ^* of a word can be smaller than that of its prefixes. Also, the measure γ^* has been studied for families of finite prefixes of well-known infinite words such as Thue-Morse word [26, 41], Episturmian words [17], k-bonacci-like words [20] and Rote sequences [18], as well as for finite factors of the Thue-Morse word [11]. Moreover, a variation of γ^* in which cyclic factors are considered has been used to characterize the necklaces of standard Sturmian words [30], well-known infinite families of finite words used as bricks to construct particular Sturmian words, called *characteristic Sturmian words*.

A groundbreaking research connecting the notion of string attractors with previously mentioned classical combinatorial notions of repetitiveness for infinite words has been presented in [41]. In particular, the *string attractor profile function* $s_{\mathbf{x}}$ of an infinite word $\mathbf{x}$ is introduced. It measures, for each $n \geq 1$, the smallest size of a string attractor for the length-n prefix of $\mathbf{x}$. The authors study the behavior of $s_{\mathbf{x}}$ when $\mathbf{x}$ is linearly recurrent, and when $\mathbf{x}$ is *automatic*, i.e., $\mathbf{x}$ can be defined through a finite automaton [1].

In this paper, in addition to the size of a string attractor, we also take into account the distribution of the positions within the string attractor. This leads to the definition of two new measures: for a finite word w, the *span* of w is the minimal span (or width) of a string attractor of w and the *leftmost measure* of w is the smallest rightmost position of a string attractor of w. Starting from these two notions the new complexity measures $\mathsf{lm}_{\mathbf{x}}$ and $\mathsf{span}_{\mathbf{x}}$ can be defined for an infinite word $\mathbf{x}$. In particular, the *span complexity function* $\mathsf{span}_{\mathbf{x}}(n)$ and the *leftmost complexity function* $\mathsf{lm}_{\mathbf{x}}(n)$ give the value of the span and the leftmost measure applied to the length-n prefix of $\mathbf{x}$.

The main goal of this paper is to explore the relation between the three string attractor-based complexities $s_{\mathbf{x}}$, $\mathsf{lm}_{\mathbf{x}}$, and $\mathsf{span}_{\mathbf{x}}$ for an infinite word $\mathbf{x}$, and other combinatorial notions of repetitiveness. The main results shown in this paper highlight that such complexity functions are able to capture some aspects of repetitiveness that are not necessarily detected by the other known functions. A preliminary version of some of such results can be found in the conference paper [40].

Firstly, we investigate in depth the connection between the function $s_{\mathbf{x}}$ and the well-known notions of repetitiveness. We prove that the values taken by $s_{\mathbf{x}}$ for infinitely many lengths of prefixes give an upper bound on the factor complexity. Moreover, it is possible to prove that a necessary condition for aperiodic words to have bounded $s_{\mathbf{x}}$ is that the word $\mathbf{x}$ is ω-power-free and its factor complexity $p_{\mathbf{x}}$ is linear. Recall that an infinite word is ω-power-free if it does not contain arbitrarily long consecutive repetitions of any factor. Here, we prove that such a condition is not sufficient, thus answering negatively to a question raised in [40].

Secondly, we prove that, analogously to the factor complexity, the leftmost complexity characterizes eventually periodic infinite words. Moreover, we provide a new characterization of Sturmian words in terms of both the span and the leftmost complexity functions. In particular, we prove that an infinite word $\mathbf{x}$ is Sturmian if and only if $\mathsf{lm}_{\mathbf{x}}$ is unbounded and $\mathsf{span}_{\mathbf{x}} = 1$ infinitely often. Unlike the factor complexity, the span and leftmost complexities uniquely determine a characteristic Sturmian word, up to exchanging the two letters of the alphabet. Analogously to Sturmian words, a characterization in terms of both the span and the leftmost complexity functions is provided for quasi-Sturmian words.

Finally, the behaviour of the three complexity functions is studied for the k-bonacci words, which can be considered a generalization of the well-known Fibonacci word to a larger alphabet of size $k > 2$. A new technique to build a string attractor of smallest size for the finite k-bonacci words is presented. The recursive procedure to construct the positions of these string attractors can be extended to more general families of words obtained by applying morphisms, which

represent a classical mechanism to generate repetitive words.

The paper is organized as follows. In Section 2 we give all the preliminary definitions. Section 3 is focused on the string attractor profile function and its relation with the factor complexity and the recurrence property. Moreover, we study the behavior of the string attractor profile function for infinite fixed points of morphisms. In Section 4, we introduce the span and leftmost measures of a finite word. We relate these notions to the number of distinct factors of the word, to its smallest string attractor, and to images under a morphism. Afterwards, we study these measures for prefixes of an infinite word, leading to the definition of the span and the leftmost complexities. We explain some links between these complexities and the combinatorial notions of recurrence, periodicity, and factor complexity in Section 5. In Section 6, we first describe the three complexities for characteristic Sturmian words. Using this description, we obtain the previously mentioned new characterizations of Sturmian and quasi-Sturmian words. Finally, in Section 7, we move the focus to the k-bonacci words and we show that each prefix of the k-bonacci word admits a string attractor of size at most k. We also give the explicit computation of the leftmost and the span complexities. We end the paper with remarks and future works in Section 8.

2. Preliminaries

Combinatorics on words. An *alphabet* is a finite set of *letters*. A *finite* (resp., *infinite*) *word* on an alphabet Σ is simply a finite (resp., infinite) sequence of letters of Σ. To distinguish them from finite words, infinite words are written in bold and we start indexing both finite and infinite words at 1, e.g., we will write $\mathbf{x} = x_1x_2\cdots$. For a finite or infinite word x, we let $|x|$ denote its *length*, i.e., the number of letters in x, and $\mathsf{alph}(x)$ denote the set of letters appearing in x. The *empty-word* ε is the only word that verifies $|\varepsilon| = 0$. We let Σ^* (resp., Σ^+) denote the set of finite (resp., non-empty finite) words over Σ. For all $n \geq 0$, we let Σ^n denote the set of length-n words over Σ.

Given a word

$$x = \begin{cases} x_1x_2\cdots x_{|x|}, & \text{if } x \text{ is finite;} \\ x_1x_2\cdots, & \text{if } x \text{ is infinite;} \end{cases}$$

an integer $1 \leq i \leq |x|$ is called a *position* within x. Given two positions $1 \leq i, j \leq |x|$, we use the notation $x[i, j] = x_ix_{i+1}\cdots x_j$; note that $x[i, j] = \varepsilon$ if $j < i$. Such a portion $x[i, j]$ for $i \leq j$ is called a *factor* of x. We let $F(x)$ denote the set of factors of x. The factor $y \in F(x)$ is *proper* if $y \neq x$. The word u is a *prefix* (resp., *suffix*) of x if $x = uv$ (resp., $x = vu$) for some word v. A factor u of x is *right special* (resp., *left special*) if there exist distinct letters $a, b \in \Sigma$ such that both ua and ub (resp., au and bu) are factors of x. The *reverse* of a finite word $x = x_1x_2\ldots x_{|x|}$ is the word read from right to left, i.e., $x^R = x_{|x|}x_{|x|-1}\cdots x_1$. If $x = x^R$, then x is a *palindrome*.

String attractor of a finite word. Roughly, a string attractor for a finite word is a set of positions within the word such that each of its factors has an occurrence "crossing" at least one element of the set. More formally, a *string attractor* of a finite word x is a set Γ of positions within x such that, for every non-empty factor $w \in F(x)$, there exist integers i, j such that $w = x[i, j]$ and $[i, j] \cap \Gamma \neq \emptyset$. We denote by $\gamma^*(x)$ the size of a smallest string attractor for x. It is easy to see that $\gamma^*(x) \geq |\mathsf{alph}(x)|$.

Example 1. Let $x = adc\underline{b}a\underline{a}d\underline{c}b\underline{a}d\underline{c}$ be a word on $\Sigma = \{a, b, c, d\}$ (the reason why some letters are underlined will become clear later on). The set $\Gamma = \{1, 4, 6, 8, 11\}$ is a string attractor for x. Note that $\Gamma' = \Gamma \setminus \{1\} = \{4, 6, 8, 11\}$ is still a string attractor for x since each factor that crosses position 1 has another occurrence that crosses a different position in Γ. The positions of Γ' are underlined above. The set Γ' is also a smallest string attractor since $|\Gamma'| = |\Sigma|$, so $\gamma^*(x) = 4$. Note that $\{3, 4, 5, 11\}$ and $\{3, 4, 6, 7, 11\}$ are also string attractors for x. It is easy to verify that the set $\Delta = \{1, 2, 3, 4\}$ is not a string attractor since, for instance, the factor aa does not intersect any position in Δ.

Factor complexity. For an infinite word $\mathbf{x}$, its *factor complexity function* $p_{\mathbf{x}}$ counts, for any integer $n \geq 0$, the distinct length-n factors of $\mathbf{x}$, i.e., $p_{\mathbf{x}}(n) = |F(\mathbf{x}) \cap \Sigma^n|$ for all $n \geq 0$.

Periodicity. Given a word x, a natural number $p \geq 1$ is called a *period* of x if $x_i = x_j$ when $i \equiv j \mod p$. An infinite word $\mathbf{x}$ is *eventually periodic* if there exist $u \in \Sigma^*$ and $v \in \Sigma^+$ such that $\mathbf{x} = uv^\omega$, i.e., $\mathbf{x}$ is the concatenation of u followed by infinite copies of a non-empty word v (denoted by v^ω). If $u = \varepsilon$, then $\mathbf{x}$ is said to be *periodic*. An infinite word is *aperiodic* if it is not eventually periodic. We recall the famous Morse-Hedlund theorem (see, for instance, [28, Theorem 1.3.13]).

Theorem 2 (Morse-Hedlund theorem)**.** *Let* $\mathbf{x}$ *be an infinite word. The following are equivalent.*

1. *The word* $\mathbf{x}$ *is eventually periodic.*
2. *We have* $p_{\mathbf{x}}(n+1) = p_{\mathbf{x}}(n)$ *for some integer* $n \geq 0$*.*
3. *The complexity function* $p_{\mathbf{x}}$ *is bounded.*

Recurrence and appearance functions. An infinite word $\mathbf{x}$ is said to be *recurrent* if every factor of $\mathbf{x}$ occurs infinitely often (in $\mathbf{x}$). The *recurrence function* $R_{\mathbf{x}} \colon n \mapsto R_{\mathbf{x}}(n)$ gives, for each n, the least integer m (or ∞ if no such m exists) such that each length-m factor of $\mathbf{x}$ contains at least an occurrence of each length-n factor of $\mathbf{x}$. An infinite word $\mathbf{x}$ is *uniformly recurrent* if $R_{\mathbf{x}}(n) < \infty$ for each $n \geq 1$. Note that $R_{\mathbf{x}}(n) - n + 1$ is the maximum gap between consecutive occurrences of the same factor, when all length-n factors are considered. If $R_{\mathbf{x}}(n)$ is linear, then $\mathbf{x}$ is *linearly recurrent*. It is easy to see that a periodic word $\mathbf{x}$ is linearly recurrent. On the other hand, if $\mathbf{x}$ is eventually periodic but not periodic, then $\mathbf{x}$ is not recurrent. Therefore, a recurrent word is either aperiodic or periodic. For an infinite word $\mathbf{x}$ and an integer n, we let $A_{\mathbf{x}}(n)$ denote the length of the shortest prefix containing all length-n factors of x. The function $n \mapsto A_{\mathbf{x}}(n)$ is called the *appearance function* of $\mathbf{x}$.

Example 3. For the binary word $\mathbf{x} = 11011100101110111\cdots$, which is the concatenation of all binary representations of the positive integers, the function $A_{\mathbf{x}}$ is easily seen to be exponential. This also follows from the fact that $p_{\mathbf{x}}$ is exponential as well, as explained in the remark below.

Remark 4. For any infinite word $\mathbf{x}$ over Σ, the fact that Σ is finite implies that $A_{\mathbf{x}}(n)$ is defined for each $n \geq 1$. One then easily sees that $p_{\mathbf{x}}(n) + n - 1 \leq A_{\mathbf{x}}(n) \leq R_{\mathbf{x}}(n)$.

Power freeness. An infinite word $\mathbf{x}$ is said to be *k-power free* for some $k > 1$ if, for every factor w of $\mathbf{x}$, w^k is not a factor of $\mathbf{x}$. If for each factor w of $\mathbf{x}$, there exists some integer $k > 1$ such that w^k is not a factor of $\mathbf{x}$, then $\mathbf{x}$ is *ω-power free*.

Morphisms. They represent a mechanism to generate infinite families of repetitive sequences, which have many mathematical properties [1, 3, 16]. Let Σ and Σ' be alphabets. A *morphism*

is a map $\varphi\colon \Sigma^* \to \Sigma'^*$ that satisfies the identity $\varphi(uv) = \varphi(u)\varphi(v)$ for all words $u, v \in \Sigma^*$. A morphism φ is *prolongable* on a letter $a \in \Sigma$ if $\varphi(a) = au$ with $u \in \Sigma^+$. If $\varphi(a) \neq \varepsilon$ for all $a \in \Sigma$, then the morphism φ is said to be *non-erasing*. Given a non-erasing morphism φ prolongable on some $a \in \Sigma$, the sequence $(\varphi^i(a))_{i\geq 0}$ of finite words gives an infinite family of prefixes of a unique infinite word $\varphi^\infty(a) = \lim_{i\to\infty} \varphi^i(a)$, which is called a *purely morphic word* or a *fixed point* of φ. A morphism φ is *primitive* if there exists $t \geq 1$ such that $b \in F(\varphi^t(a))$ for every pair of letters $a, b \in \Sigma$. If there exists k such that $|\varphi(a)| = k$ for every $a \in \Sigma$, then φ is said to be *k-uniform*.

Example 5. Let us consider the *Thue–Morse word* $\mathbf{t} = 0110100110010110\cdots$ which is the fixed point of the 2-uniform morphism $0 \mapsto 01$, $1 \mapsto 10$. It is known that the functions $p_{\mathbf{t}}(n)$, $R_{\mathbf{t}}(n)$ and $A_{\mathbf{t}}(n)$ are $\Theta(n)$. See [1] for details.

Lempel-Ziv factorization. The *Lempel-Ziv factorization* or *parsing* (*LZ77 parsing* in short) of a finite word w is its factorization $LZ(w) = v_1 v_2 \cdots v_z$ built from left to right in a greedy way as follows: each new factor (also called an *LZ-phrase*) v_i is either the leftmost occurrence of a letter in w or the longest prefix of $v_i \cdots v_z$ occurring in $v_1 \cdots v_{i-1}$. We let $z(w)$ denote the number of LZ-phrases in the LZ77 parsing of w. A measure on infinite words is naturally associated [10]: for an infinite word $\mathbf{x}$, the *LZ-complexity function* $z_{\mathbf{x}}$ maps each $n \geq 1$ to the number $z(\mathbf{x}[1, n])$ of LZ-phrases of the length-n prefix of $\mathbf{x}$.

The link between string attractors and LZ77 parsings is given in the result below. It follows from the fact that any given finite word has a string attractor of size equal to the number of its LZ-phrases.

Proposition 6 ([23])**.** *For every word $w \in \Sigma^*$, $\gamma^*(w) \leq z(w)$.*

3. String attractor profile function, factor complexity and recurrence

In this section, we explore the growth of the size of the smallest string attractor when considering larger and larger prefixes of an infinite word. Such an idea was first considered in [41].

Definition 7. Let $\mathbf{x}$ be an infinite word. The *string attractor profile function* of $\mathbf{x}$ is the map $s_{\mathbf{x}}\colon n \mapsto \gamma^*(\mathbf{x}[1, n])$, i.e. $s_{\mathbf{x}}(n)$ is the size of a smallest string attractor for the length-n prefix of $\mathbf{x}$.

We will study the link between the string attractor profile function and different notions measuring the repetitiveness of factors within infinite sequences of symbols. We start by establishing a bond between the appearance, factor complexity and string attractor profile functions. In particular, it shows that upper bounds on $s_{\mathbf{x}}$ induce upper bounds on $p_{\mathbf{x}}$.

Proposition 8. *Let $\mathbf{x}$ be an infinite word. For all $n \geq 1$, we have $p_{\mathbf{x}}(n) \leq n \cdot s_{\mathbf{x}}(A_{\mathbf{x}}(n))$.*

Proof. Since alphabets are finite, so is the value $A_{\mathbf{x}}(n)$. By definition $s_{\mathbf{x}}(A_{\mathbf{x}}(n))$ is the size of a smallest string attractor Γ of the prefix of length $A_{\mathbf{x}}(n)$. Therefore, each length-n factor of $\mathbf{x}$ crosses at least one element of this string attractor. Since each element of Γ is crossed by at most n distinct length-n factors of $\mathbf{x}$, one has $p_{\mathbf{x}}(n) \leq n \cdot s_{\mathbf{x}}(A_{\mathbf{x}}(n))$. □

Using the link between string attractors and LZ77 parsings, we easily obtain an upper bound on $s_{\mathbf{x}}$ as follows.

Proposition 9. *Let* $\mathbf{x}$ *be an infinite word. Then* $s_{\mathbf{x}}(n) = O\left(\frac{n}{\log n}\right)$.

Proof. Using Proposition 6, we have $s_{\mathbf{x}}(n) \leq z_{\mathbf{x}}(n)$. To conclude, it suffices to use the following upper bound on $z_{\mathbf{x}}(n)$ from [27]: the number of LZ-phrases for a length-n word on an alphabet Σ is bounded by $\frac{n}{(1-\epsilon_n)\log_{|\Sigma|}(n)}$, where $|\Sigma|$ denotes the size of the alphabet Σ and $\epsilon_n = 2\,\frac{1+\log_{|\Sigma|}(\log_{|\Sigma|}(n|\Sigma|))}{\log_{|\Sigma|}(n)}$. □

It is possible to construct an infinite word $\mathbf{x}$ for which there exists a sequence of positive integers n_i, $i \geq 1$, such that $s_{\mathbf{x}}(n_i) = \Theta(\frac{n_i}{\log n_i})$. For instance, such a word $\mathbf{x}$ can be constructed by using a suitable sequence of de Bruijn words. However, having information on the values of the string attractor profile function over a sequence $(n_i)_{i\geq 1}$ does not allow us to precisely determine its entire behavior. Therefore, we do not know whether the bound of Proposition 9 is tight.

However, if we assume that the appearance function is linear, a better bound on the function $s_{\mathbf{x}}$ is given below.

Theorem 10 ([41])**.** *Let* $\mathbf{x}$ *be an infinite word. If* $A_{\mathbf{x}}(n) = \Theta(n)$*, then* $s_{\mathbf{x}}(n) = O(\log n)$.

In the following subsections we show several examples in which different repetitiveness aspects are considered (Subsection 3.1, we analyse which combinatorial notions of repetitiveness are related to the boundedness of the string attractor profile function (Subsection 3.2) and, finally, we study the behaviour of the string attractor profile function in case of infinite words generated by morphisms (Subsection 3.3).

3.1. Some concrete examples

In this subsection, we study the behavior of the string attractor profile function of various types of infinite words and we focus on the relation with other measures of repetitiveness.

First, we look at the string attractor profile function of a periodic word representing the simplest case of repetitiveness.

Example 11. Let us consider the word $(01)^\omega = 01010101\cdots$. The word is periodic, and therefore $p_{(01)^\omega}(n) = \Theta(1)$ and $A_{(01)^\omega}(n) = n + 1$. Since each non-empty factor v of $(01)^\omega$ has an occurrence starting either in the first or in the second position (respectively when v starts with 0 or 1), the set $\{1, 2\}$ is a string attractor for each prefix of length $n \geq 2$ of $(01)^\omega$, and therefore $s_{(01)^\omega}(n) = \Theta(1)$.

As shown later in Proposition 19, every infinite word with factor complexity $\Theta(1)$ has a bounded string attractor profile function. Since by Proposition 8 we can see that a word with factor complexity $\Theta(n^{1+\varepsilon})$ can not have a bounded string attractor profile function, in this subsection we consider different case studies with linear factor complexity. In the following example, we provide a non-recurrent infinite word having linear complexity function and unbounded string attractor profile function.

Example 12. Let us consider the characteristic sequence $\mathbf{c} = 1101000100000001\cdots$ of powers of 2, i.e., $c_i = 1$ if $i = 2^j$ for some $j \geq 0$, $c_i = 0$ otherwise. It is easy to see that $\mathbf{c}$ is aperiodic and not recurrent (e.g., the factor 11 occurrs only once). It is known that $p_{\mathbf{c}}(n)$ and $A_{\mathbf{c}}(n)$ are $\Theta(n)$ [1]. One can prove that $s_{\mathbf{c}}(n) = \Theta(\log n)$ [25, 30, 41].

Example 13 gives a recurrent (not uniformly) infinite word with linear factor complexity and unbounded string attractor profile function.

Example 13. Let $\mu\colon \{0,1\}^* \to \{0,1\}^*$ be the 3-uniform morphism defined by $\mu(0) = 010$ and $\mu(1) = 111$. The infinite word $\mathbf{w} = \mu^\omega(0) = 0101110101111111111010\cdots$ has linear factor complexity $p_{\mathbf{w}}$. Moreover, it is recurrent, but not uniformly. Finally, since all factors $01^{3^k}0$, $k \geq 1$, occur in $\mathbf{w}$ and do not overlap with each other, the string attractor profile function $s_{\mathbf{w}}$ is not bounded by a constant. Actually, as a consequence of Theorem 22 (proved in Subsection 3.3), we can conclude that $s_{\mathbf{w}}(n) = \Theta(\log n)$.

In the previous example, the fact that the string attractor profile function is unbounded follows from the existence of arbitrary large powers of 1. The example below uses the Thue-Morse word to give an ω-power free infinite word with linear factor complexity and unbounded string attractor profile function.

Example 14. Let $\psi\colon \{s, a_0, b_0, a_1, b_1\}^* \to \{s, a_0, b_0, a_1, b_1\}^*$ be the 2-uniform morphism defined by $\psi(s) = sb_0$, $\psi(a_x) = a_{\overline{x}}b_{\overline{x}}$, and $\psi(b_x) = b_{\overline{x}}a_{\overline{x}}$ for all $x \in \{0,1\}$, where $\overline{x} = 1 - x$. Since ψ is 2-uniform, it follows that the infinite word $\mathbf{v} = \psi^\omega(s) = sb_0b_1a_1b_0a_0a_0b_0b_1a_1a_1b_1\cdots$ has linear factor complexity [1]. Moreover, one can observe that if we consider the coding $\lambda\colon \{s, a_0, b_0, a_1, b_1\}^* \mapsto \{0,1\}^*$ defined by $\lambda(s) = \lambda(a_0) = \lambda(a_1) = 0$ and $\lambda(b_0) = \lambda(b_1) = 1$ and apply it on $\mathbf{v}$, we obtain the Thue-Morse word $\mathbf{t} = 01101001100101101\cdots$. Since $\mathbf{t}$ is 3-power free [1], it follows that $\mathbf{v}$ is ω-power free. Finally, since all the factors $b_0\psi^{2k-1}(b_0)b_0$, $k \geq 1$, occur only once in $\mathbf{v}$ and do not overlap with each other, the string profile function $s_{\mathbf{v}}$ is not bounded by a constant.

The next example shows an infinite word that is uniformly recurrent and with linear factor complexity. Also in this case the string attractor profile function is unbounded.

Example 15. Let us consider the two 3-uniform morphisms

$$\mu\colon \begin{cases} 0 \mapsto 010 \\ 1 \mapsto 111 \end{cases} \quad \text{and} \quad \bar{\mu}\colon \begin{cases} 0 \mapsto 000 \\ 1 \mapsto 101 \end{cases}$$

and the word $\mathbf{q} = \lim_{n\to\infty} \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \circ \bar{\mu}^n \circ \mu^n(0)$. This word is of linear factor complexity [14, Proposition 2.1] and is uniformly recurrent [13, Lemma 7]. Let us show that, for all $n \geq 1$, the prefix $u_n = \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \bar{\mu}^n \circ \mu^n(0)$ of $\mathbf{q}$ requires at least $n-1$ positions in any of its string attractors. Observe that, in $\mu^n(0)$, we have the factors $01^{3^i}0$ for all $1 \leq i \leq n-1$ which do not overlap one another. Let us show that their images under $\sigma = \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \circ \bar{\mu}^n$ do not overlap either. We first make the following observation. By definition of the morphism μ, for any words u and w, if u contains (at least) a 0, then any occurrence of $\mu(u)$ in $\mu(w)$ corresponds to an occurrence of u in w. In other words, for any u and v containing (at least) a 0 each, $\mu(u)$ and $\mu(v)$ overlap in $\mu(w)$ if and only if u and v overlap in w. Similarly, for any u and v containing (at least) a 1 each, $\bar{\mu}(u)$ and $\bar{\mu}(v)$ overlap in $\eta(w)$ if and only if u and v overlap in w. As σ is a composition of μ and $\bar{\mu}$, this shows that the factors $\sigma(01^{3^i}0)$, $1 \leq i \leq n-1$, do not overlap in u_n. We conclude that $s_{\mathbf{q}}$ is not bounded.

However, many classical infinite words in literature have a known string attractor profile function bounded by a constant. It is the case of the Thue–Morse word (Example 24), the period-doubling word (Example 25), and, as shown in this paper, the characteristic Sturmian words (Theorem 43), the k-bonacci words (Theorem 57) and the family of words defined by S. Holub in [22] (Example 16).

Example 16. Let us define an infinite word $\mathbf{u}$ introduced by S. Holub in [22]. For that, let $(n_i)_{i\geq 1}$ be an increasing sequence of positive integers with $n_1 \geq 2$. We recursively define the sequence $(u_i)_{i\geq 0}$ as $u_0 = \varepsilon$ and $u_i = u_{i-1}0(u_{i-1}1)^{n_i}u_{i-1}$. It has been proven in [22] that $\mathbf{u} = \lim_{i\to\infty} u_i$ is uniformly recurrent but not linearly recurrent. Moreover, for each $i \geq 1$, $\mathbf{u}$ can be factorized as a product of words u_i0 and u_i1, i.e., $\mathbf{u} = u_ic_1u_ic_2u_ic_3\cdots$, where $c_j \in \{0, 1\}$. More precisely it has been proved in [22] that each occurrence of u_i starts at position that is a multiple of $|u_i| + 1$. By using such a property, the word u has exactly two right special factors of length n, for each $n \geq 1$. They are precisely the length-n suffixes of $u_{i-1}0(u_{i-1}1)^{n_i}u_{i-1}$ and $(u_{i-1}1)^{n_i}u_{i-1}0u_{i-1}$ where $|u_{i-1}| + 1 \leq n \leq |u_i|$. Consequently, $p_{\mathbf{u}}(n) = 2n$ [4].

Furthermore, it is possible to prove that, for $i \geq 1$, the set

$$\Gamma^{(i)} = \left\{ |u_{i-1}| + 1, \sum_{k=0}^{i-1}(|u_k| + 1), 2|u_{i-1}| + 2 \right\}$$

is a string attractor for u_i. In fact, given the recursive construction of $\mathbf{u}$, for each factor v of u_i we can find $0 \leq j \leq i - 1$ such that $|u_j| < |v| \leq |u_{j+1}|$, and each of these factors must fall in one of the following mutually exclusive cases:

1. $v = s_j(1u_j)^{q_1}0(u_j1)^{q_2}p_j$, for some $q_1, q_2 \geq 0$ such that $q_1 + q_2 \leq n_{j+1}$, and for some prefix p_j and suffix s_j of u_j;
2. $v = s_j(1u_j)^{h_1}0u_j0(u_j1)^{h_2}p_j$, for some $j < i - 1$ and $h_1, h_2 \geq 0$ such that $h_1 + h_2 < n_{j+1}$, and for some prefix p_j and suffix s_j of u_j;
3. $v = s_j(1u_j)^k1p_j$, for some $0 \leq k < n_i$ (resp. $0 \leq k \leq n_j$) if $j = i - 1$ (resp. if $j < i - 1$), and for some prefix p_j and suffix s_j of u_j.

One can observe that for all $j < i - 1$, the factors v from case 1. have an occurrence crossing the position $\sum_{k=0}^{i-1}(|u_k| + 1) \in \Gamma^{(i)}$, while if $j = i - 1$ the only occurrence of v in u_i crosses the position $|u_{i-1}| + 1 \in \Gamma^{(i)}$. The factors v that fall in case 2. on the other hand overlap the first position in $\Gamma^{(i)}$ in correspondence to the 0 in v at position $|s_j| + h_1(1 + |u_j|) + 1$, that is $|u_j| + 1 \in \Gamma^{(i)}$ once again. Finally, one occurrence of each factor falling in case 3. can be found overlapping the last position in $\Gamma^{(i)}$, where the last 1 right before p_j ends up exactly at position $2|u_i| + 2 \in \Gamma^{(i)}$.

We deduce a string attractor for the length-n prefix of $\mathbf{u}$ as follows: if i is such that $|u_i| < n < |u_{i+1}|$, we can merge the set $\Gamma^{(i)}$ with the positions $\leq n$ in $\Gamma^{(i+1)}$ to obtain a string attractor for the length-n prefix. Such a string attractor can have up to 6 positions, and it follows that $s_{\mathbf{u}}(n) = \Theta(1)$.

3.2. The bounded case

Supported by the previous subsection, it is relevant to detect which combinatorial properties of infinite words are related to the boundedness of the string attractor profile function. Observe that we already know the following result.

Theorem 17 ([41])**.** *For any linearly recurrent infinite word* $\mathbf{x}$*, we have* $s_{\mathbf{x}}(n) = \Theta(1)$.

The previous theorem is not a characterization. Indeed, Example 16 exhibits uniformly (and not linearly) recurrent words $\mathbf{x}$ for which $s_{\mathbf{x}}$ is bounded. In this section, we gather results towards a characterization.

First, we analyze how the boundedness of $s_{\mathbf{x}}$ structures the infinite word $\mathbf{x}$ and we show that if an infinite word ha s its string attractor profile function bounded by some constant value, then it has at most linear factor complexity. More precisely, we have the following result.

Theorem 18. *Let* $\mathbf{x}$ *be an infinite word. If* $s_{\mathbf{x}} = \Theta(1)$*, then either* $\mathbf{x}$ *is eventually periodic, or* $\mathbf{x}$ *is* ω*-power free and* $p_{\mathbf{x}} = \Theta(n)$.

Proof. First, Proposition 8 implies that, if k is such that $s_{\mathbf{x}}(n) < k$ for each $n \geq 1$, then $p_{\mathbf{x}}(n) \leq n \cdot k$ for each $n \geq 1$. Therefore, the factor complexity is (at most) linear. Towards a contradiction, let us assume now that $\mathbf{x}$ is aperiodic and not ω-power free. Then there exists a factor w of $\mathbf{x}$ such that, for every $q \geq 1$, w^q is factor of $\mathbf{x}$. Moreover, the assumption on $\mathbf{x}$ implies that $\mathbf{x} \neq uw^\omega$ for any $u \in \Sigma^*$. It follows that there exists an increasing sequence $(q_j)_{j\geq 1}$ of integers such that, for each j, there exist a proper suffix s_j and a proper prefix p_j of w, and two letters a_j and b_j such that $a_j s_j$ is not a suffix of w, $p_j b_j$ is not a prefix of w, and $a_j s_j w^{q_j} p_j b_j$ is a factor of $\mathbf{x}$. As any position can cover at most two such factors, $s_{\mathbf{x}}$ is unbounded. This is a contradiction. □

The following proposition shows that, in the case of eventually periodic words, the string attractor profile function is bounded by a constant.

Proposition 19. *For any eventually periodic infinite word* $\mathbf{x}$*, we have* $s_{\mathbf{x}}(n) = \Theta(1)$.

Proof. Let $u \in \Sigma^*$ and $v \in \Sigma^+$ such that $\mathbf{x} = uv^\omega$. For all $n \geq 1$, $\{1, \ldots, \min\{n, |uv|\}\}$ is a string attractor for the length-n prefix. Therefore, $s_{\mathbf{x}}(n) \leq |uv|$ for all n. □

However, the converse of Theorem 18 does not hold. Indeed, in Example 14 a ω-power free word with linear factor complexity and unbounded string attractor profile function is given.

Note that even the stronger hypothesis of uniform recurrence together with linear factorial complexity does not guarantee that the string profile function is bounded, as shown in Example 15.

Observe that Examples 15 and 14 negatively answer to the questions posed in [40] in which it was asked whether linear factor complexity along with either uniformly recurrence or ω-power freeness property are sufficient to guarantee a bounded string attractor profile function for a given infinite word.

The problem of finding a complete characterization of the infinite words having a bounded string attractor profile function is still open.

We conclude this subsection with Table 1 showing a synoptic overview of the factor complexity, repetitiveness properties, and string attractor profile function for both the infinite words described in this section and those that will be considered in the rest of the paper. Note that apart from the periodic word $(01)^\omega$, all words considered in the table have linear factor complexity. By Theorem 18, linear factor complexity is a necessary but not sufficient condition for an aperiodic infinite word to have a bounded string attractor profile function. Four infinite words with unbounded string attractor profile function are shown while, for the other words, we have $s_{\mathbf{x}}(n) = \Theta(1)$. In these cases, the exact values of $s_{\mathbf{x}}(n)$, for n large enough, are reported in the table. This points out both that different repetitiveness aspects may be hiding behind a constant string attractor profile function but also that infinite words having deeply different combinatorial structures and properties may have point-wise equal values of the function $s_{\mathbf{x}}(n)$. This fact motivates the use of the notion of string attractor to define new complexity measures with the goal of capturing such combinatorial properties, as we describe in the next sections.

3.3. The case of purely morphic words

Some data compression measures have been explored when applied to fixed points of morphisms, or more specifically, to iterated images of a morphism. It is the case of the number of

Infinite word $\mathbf{x}$	$p_{\mathbf{x}}(n)$	Recurrence	ω-power free	$s_{\mathbf{x}}(n)$
$(01)^\omega$ (Ex. 11)	$\Theta(1)$	linearly recurrent	No	2
$\mathbf{c}$ (Ex. 12)	$\Theta(n)$	not recurrent	No	$\Theta(\log n)$
$\mathbf{w}$ (Ex. 13)	$\Theta(n)$	recurrent	No	$\Theta(\log n)$
$\mathbf{v}$ (Ex. 14)	$\Theta(n)$	not recurrent	Yes	$\Theta(\log n)$
$\mathbf{q}$ (Ex. 15)	$\Theta(n)$	uniformly recurrent	Yes	$\Theta(\log n)$
$\mathbf{u}$ (Ex. 16)	$\Theta(n)$	uniformly recurrent	Yes	3
$\mathbf{s}$ (Sec. 6)	$\Theta(n)$	uniformly recurrent	Yes	2
$\mathbf{t}$ (Ex. 24)	$\Theta(n)$	linearly recurrent	Yes	4
$\mathbf{pd}$ (Ex. 25)	$\Theta(n)$	linearly recurrent	Yes	2
$\mathbf{b}^{(k)}$ (Sec. 7)	$\Theta(n)$	linearly recurrent	Yes	k

Table 1: The table shows the factor complexity $p_{\mathbf{x}}(n)$, recurrence properties, ω-power freeness and the string attractor profile function $s_{\mathbf{x}}(n)$ for large enough n, for all the infinite words $\mathbf{x}$ considered in Sections 3, 6 and 7, namely: the periodic word $(01)^\omega$, the characteristic sequence $\mathbf{c}$ of powers of 2; the purely morphic word $\mathbf{w}$ generated by the morphism μ defined by $\mu(0) = 010$ and $\mu(1) = 111$; a purely morphic word $\mathbf{v}$ generated by a 2-uniform morphism; a uniformly recurrent word $\mathbf{q}$ defined using μ and its counterpart obtained by echanging 0 and 1; an infinite word $\mathbf{u}$ introduced by Holub in [22]; any characteristic Sturmian word $\mathbf{s}$; the Thue-Morse word $\mathbf{t}$; the period doubling word $\mathbf{pd}$; the k-bonacci word $\mathbf{b}^{(k)}$ defined over an alphabet of size k.

BWT equal-letter runs [19] and of the LZ-complexity function [10]. Therefore, it is natural to wonder if similar results can be obtained for the string attractor profile function.

First we present an upper bound on the string attractor profile function of purely morphic words.

Theorem 20. *Let $\mathbf{x} = \varphi^\infty(a)$ be the fixed point of a morphism φ prolongable on $a \in \Sigma$. Then $s_{\mathbf{x}}(n) = O(i)$, where i is such that $|\varphi^i(a)| \leq n < |\varphi^{i+1}(a)|$. In particular, if there exists $\rho > 1$ such that $|\varphi^i(a)| = \Omega(\rho^i)$, then $s_{\mathbf{x}}(n) = O(\log n)$.*

To prove Theorem 20, we use Proposition 6 and the following result about the number of LZ-phrases in the LZ77 parsing in purely morphic words.

Proposition 21 ([10])**.** *Let $\mathbf{x} = \varphi^\infty(a)$ be the fixed point of a non-erasing morphism φ prolongable on $a \in \Sigma$. Then*

$$z(\varphi^i(a)) = \begin{cases} \Theta(1), & \text{if } \mathbf{x} \text{ is eventually periodic;} \\ \Theta(i), & \text{otherwise.} \end{cases}$$

Proof of Theorem 20. For all $i \geq 0$, define $n_i = |\varphi^i(a)|$. By Proposition 21, there exist two constant $c_1, c_2 \geq 1$ such that for all $n \in [n_i, n_{i+1})$, we have $c_1 \cdot i \leq z_{\mathbf{x}}(n_i) \leq z_{\mathbf{x}}(n) \leq z_{\mathbf{x}}(n_{i+1}) \leq$

$c_2 \cdot i + c_2$. Note that the second and third inequalities follow by the monotonicity of the measure z (i.e., $z(u) \leq z(uv)$ for all $u, v \in \Sigma^*$). This implies that $z_{\mathbf{x}}(n) = \Theta(i)$, and by Proposition 6 it follows that $s_{\mathbf{x}} = O(i)$. In particular, if $|\varphi^i(a)| = \Omega(\rho^i)$ for some $\rho > 1$, then one has $n \in \Omega(\rho^i)$ or, conversely, $i = O(\log n)$ so the conclusion $s_{\mathbf{x}}(n) = O(i) = O(\log n)$ follows. □

In the following theorem, we provide a finer result in the case of binary purely morphic word.

Theorem 22. *Let $\mu\colon \{a, b\}^* \to \{a, b\}^*$ be a morphism prolongable on a and $\mathbf{x} = \mu^\infty(a)$. Then either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$, and it is decidable whether the former or the latter occurs.*

Proof. Based on the morphism μ, we can decide in which of the following (mutually exclusive) cases we are.

1. The word $\mathbf{x}$ is eventually periodic [38, Theorem 4].
2. The word $\mathbf{x}$ is aperiodic and there exist a non-erasing morphism $\tau\colon \Sigma^* \to \{a, b\}^*$ and a primitive morphism $\varphi\colon \Sigma^* \to \Sigma^*$ such that $\mathbf{x} = \mu^\infty(a) = \tau(\varphi^\infty(a))$ (whenever μ is primitive, as well as some decidable cases where $\mu(b) = b$ by [37, Theorem 4.1] and its proof).
3. The word $\mathbf{x}$ is aperiodic and contains arbitrarily large powers of b's (whenever $\mu = b^k$, $k \geq 2$, as well as some decidable cases where $\mu(b) = b$ by [37, Theorem 4.1]).

Let us now show that, in each case, we have either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$. For the first case, we have $s_{\mathbf{x}}(n) = \Theta(1)$ as a direct consequence of Proposition 19. In the second case, as φ is primitive, $\varphi^\infty(a)$ is linearly recurrent (see [15, Proposition 25]). This implies that $\mathbf{x}$ is also linearly recurrent and thus that $s_{\mathbf{x}}(n) = \Theta(1)$ by Theorem 17.

We now turn to the third case. Observe that, by Theorem 18, we cannot have $s_{\mathbf{x}}(n) = \Theta(1)$, so we show that $s_{\mathbf{x}}(n) = \Theta(\log n)$. By [19, Proposition 20 and Corollary 27], the number of distinct maximal runs of b's grows logarithmically with respect to the length of the prefixes of $\mathbf{x}$. As a position in a string attractor can cover at most two different runs of b's, this implies that $s_{\mathbf{x}}(n) = \Omega(\log n)$. On the other hand, observe that by aperiodicity $\mu(a)$ contains at least two occurrences of a. Therefore, $|\mu^n(a)| = \Omega(2^n)$ and, by Theorem 20, we conclude that $s_{\mathbf{x}}(n) = O(\log n)$ so $s_{\mathbf{x}}(n) = \Theta(\log n)$. □

The same result has been obtained for another class of words as reported below. In short, an infinite word $\mathbf{x}$ is *k-automatic*, with $k \geq 2$, if and only if there exist a coding $\tau\colon \Sigma \to \Sigma$ and a k-uniform morphism μ_k such that $\mathbf{x} = \tau(\mu_k^\infty(a))$, for some $a \in \Sigma$ [1]. An infinite word is called *automatic* if it is k-automatic for some $k \geq 2$.

Theorem 23 ([41])**.** *Let $\mathbf{x}$ be an automatic infinite word. Then, either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$, and it is decidable whether the former or the latter occurs.*

Examples 12 and 13 show two automatic sequences for which the string attractor profile function is $\Theta(\log n)$.

For some particular automatic words obtained as fixed points of morphisms, string attractors may be found by using their specific combinatorial structure and properties, as shown in the next example.

Example 24. Let us consider the Thue–Morse word $\mathbf{t} = 0110100110010110\cdots$. It is a purely morphic word, as described in Example 5. It has been proven in [41] (cf. also [26, 11]) that $s_{\mathbf{t}}(n) = 4$ for all $n \geq 25$.

Moreover, the authors of [41] show that, if the string attractor profile function is bounded, it is possible to build an automaton which returns the positions of a smallest string attractor for each prefix. However, the construction of such an automaton is done case by case using the theorem-proving software `Walnut` [33]. Such a technique has been used in [41] to find a string attractor of smallest size for the automatic infinite word considered in the next example.

Example 25. Consider the *period-doubling word* $\mathbf{pd} = 101110101011\cdots$, which is the fixed point of the morphism $1 \mapsto 10$, $0 \mapsto 11$. It has been proven in [41, Theorem 3] that $s_{\mathbf{pd}}(n) = 2$ for all $n \geq 1$. In particular, it has been shown [41, Theorem 4] that for the prefix of $\mathbf{pd}$ of length $n \geq 6$, a string attractor of smallest size is

$$\Gamma(\mathbf{pd}[1,n]) = \begin{cases} \{3\cdot 2^{i-3}, 3\cdot 2^{i-2}\}, & \text{if } 2^i \leq n < 3\cdot 2^i; \\ \{2^i, 2^{i+1}\}, & \text{if } 3\cdot 2^i \leq n < 2^{i+1}. \end{cases}$$

4. Two new string attractor-based measures

In this section, we introduce two new notions related to the string attractors of a word. Indeed, knowing the minimal size of a string attractor is often not sufficient to understand the structure of a word or choose interesting string attractors. Therefore, it can be useful to also take into account the distribution of the positions in the string attractors. This is what our new measures capture and, as we will show later on, they will allow us to distinguish families of words.

The first measure is the span of a word, which gives the minimum distance between the rightmost and the leftmost positions of any string attractor.

Definition 26. Let w be a finite word and let $\mathcal{G}$ be the set of all string attractors for w. The *(string attractor) span* of w is the value $\mathsf{span}(w) = \min_{\Gamma\in\mathcal{G}}(\max\Gamma - \min\Gamma)$. We will also abusively say that the quantity $(\max\Gamma - \min\Gamma)$ is the span of the string attractor Γ.

Example 27. Let us consider the word $w = \overline{abc}\underline{\overline{c}abc}$ on the alphabet $\Sigma = \{a, b, c\}$. One can see that the sets $\Gamma_1 = \{4, 5, 6\}$ (underlined positions) and $\Gamma_2 = \{1, 2, 4\}$ (overlined positions) are two suitable string attractors for w. Both are of minimal size as $|\Gamma_1| = |\Gamma_2| = |\Sigma|$ but they have different spans. Moreover, since all of the positions of Γ_1 are consecutive, it is of minimal span and therefore $\mathsf{span}(w) = 6 - 4 = 2$.

The span can be used to derive an upper-bound on the number of distinct factors, as shown below.

Proposition 28. *For any finite word w over Σ, we have $|F(w) \cap \Sigma^n| \leq n + \mathsf{span}(w)$ for all $1 \leq n \leq |w|$.*

Proof. Let Γ be a string attractor of minimal span and write $\delta = \min\Gamma$ and $\delta' = \max\Gamma$. Then, the interval $\Gamma' = [\delta, \delta']$ contains Γ and is a string attractor for w. Since every factor has an occurrence crossing a position in Γ', it is possible to find all length-n factors of w by considering a window of length n sliding from position $\max\{\delta - n + 1, 1\}$ to position $\min\{\delta', |w| - n + 1\}$. One can see that this interval is of size at most $\delta' - (\delta - n + 1) + 1 = n + \mathsf{span}(w)$. This ends the proof. □

In addition, we may compare string attractors of a given word according to their rightmost positions. More specifically, we will want the string attractor having the smallest such position. This gives the notion defined below.

Definition 29. Let w be a finite word and let $\mathcal{G}$ be the set of all string attractors for w. The *leftmost string attractor* for w is a string attractor $\Gamma \in \mathcal{G}$ such that, for all $\Gamma' \in \mathcal{G}$, we have $\max \Gamma \leq \max \Gamma'$. The *(string attractor) leftmost measure* of w is then $\mathsf{lm}(w) = \max \Gamma$, where Γ is a leftmost string attractor.

Example 30. We resume Example 27. First, we have $4 = \max \Gamma_2 < \max \Gamma_1 = 6$. Second, the set $\Delta = \{1, 2, 3\}$ is not a string attractor for w. Therefore $\mathsf{lm}(w) = 4$.

Examples 27 and 30 show that for the finite word $w = abccabc$, these two measures can be realized by distinct string attractors. In fact, in this case, it is not possible to find a leftmost string attractor having minimal span since $\{2, 3, 4\}$ is not a string attractor.

Similarly to what we did for the span, we can use the leftmost measure to obtain an upper-bound on the number of distinct factors.

Proposition 31. *For any finite word w over Σ, we have $|F(w) \cap \Sigma^n| \leq \mathsf{lm}(w)$ for all $1 \leq n \leq |w|$.*

Proof. The proof follows the same lines as that of Proposition 28 by considering a leftmost string attractor Γ, and $\Gamma' = [1, \max \Gamma]$ instead. □

In Examples 27 and 30, we can see that $\gamma^*(w) - 1 \leq \mathsf{span}(w) \leq \mathsf{lm}(w) - 1$. This is a general result as shown below.

Proposition 32. *Let w be an finite word. Then, $\gamma^*(w) - 1 \leq \mathsf{span}(w) \leq \mathsf{lm}(w) - 1$.*

Proof. Let Γ be a string attractor of w with minimal span. It contains at most $\max \Gamma - \min \Gamma + 1 = \mathsf{span}(w) + 1$ elements, therefore $\gamma^*(w) \leq \mathsf{span}(w) + 1$.

Let Γ' be a leftmost string attractor of w. Its span is at most $\max \Gamma' - 1 = \mathsf{lm}(w) - 1$, therefore $\mathsf{span}(w) \leq \mathsf{lm}(w) - 1$. □

The following proposition shows how the size of the smallest string attractor, the span and the leftmost measure of a word yield bounds on the corresponding measures for its image under a morphism.

Proposition 33. *Let $\varphi \colon \Sigma^* \to \Sigma'^*$ be a morphism. There exists a constant $C \geq 1$ which depends only on φ such that, for every $w \in \Sigma^+$, the following hold:*

1. $\gamma^*(\varphi(w)) \leq 2\gamma^*(w) + C$;
2. $\mathsf{span}(\varphi(w)) \leq C \cdot \mathsf{span}(w)$;
3. $\mathsf{lm}(\varphi(w)) \leq C \cdot \mathsf{lm}(w)$.

Proof. Starting from a given string attractor Γ for w, we show how one can build a valid string attractor for $\varphi(w)$ in two steps.

Step 1. First, we consider the factors of the images of letters, i.e., the elements of $F_\varphi = \bigcup_{a \in \Sigma} F(\varphi(a))$. Recall that for every symbol $a \in \Sigma$ there is at least one position $j \in \Gamma$ such that $w_j = a$; let us denote j_a such a position. Then, for every $a \in \Sigma$ we can choose any minimum string attractor Γ_a of $\varphi(a)$ and overlay it on the occurrence of $\varphi(w_{j_a})$ to cover the factors of $\varphi(a)$. In other words, every element of F_φ has an occurrence in w crossing at least a position in

$$\mathcal{T}_\varphi = \bigcup_{a \in \Sigma} \{|\varphi(w[1, j_a - 1])| + \delta : \delta \in \Gamma_a\}.$$

Step 2. Let us now consider the other factors of $\varphi(w)$, i.e., the elements of $F(\varphi(w))$ which are not in F_φ. To cover these factors, we define two sets of positions. Let $\mathcal{T}_f = \{|\varphi(w[1, j-1])| + 1 : j \in \Gamma\}$ be the set of positions corresponding to the first letter of $\varphi(w_j)$, where j is a position in Γ. Analogously, we define the set $\mathcal{T}_\ell = \{|\varphi(w[1, j])| : j \in \Gamma\}$ as the set of positions corresponding to the last letter of some $\varphi(w_j)$ with $j \in \Gamma$.

Let $u \in F(\varphi(w)) \setminus F_\varphi$ and let v be a factor of w of minimal length such that u is a factor of $\varphi(v)$. Observe that, by definition of F_φ, v is of length at least 2. As v is a factor of w, it has an occurrence crossing some position $j \in \Gamma$. By minimality of v, we know that u has an occurrence crossing either the first position of $\varphi(w_j)$, or the last position of $\varphi(w_j)$ (or both). Therefore, u crosses a position in $\mathcal{T}_f$ or $\mathcal{T}_\ell$.

As a consequence of the previous two steps, $\Gamma' = \mathcal{T}_\varphi \cup \mathcal{T}_f \cup \mathcal{T}_\ell$ is a string attractor for $\varphi(w)$. Recall that this construction can be done starting from any string attractor Γ of w, giving different corresponding string attractors Γ'. To obtain the three claimed inequalities, we will consider different string attractors Γ of w. Now let us denote $\ell = \max_{a\in\Sigma} |\varphi(a)|$, i.e., ℓ is the longest image of a letter.

1. If Γ is such that $|\Gamma| = \gamma^*(w)$, then

$$\gamma^*(\varphi(w)) \leq |\Gamma'| \leq |\mathcal{T}^f| + |\mathcal{T}^l| + |\mathcal{T}^\varphi| \leq 2\gamma^*(w) + \sum_{a\in\Sigma} \gamma^*(\varphi(a)).$$

2. If Γ is such that $\delta = \min\Gamma$, $\delta' = \max\Gamma$ and $\delta' - \delta = \mathsf{span}(w)$, then by construction we have $\min\Gamma' = |\varphi(w[1, \delta - 1])| + 1 \in \mathcal{T}_f$ and $\max\Gamma' = |\varphi(w[1, \delta'])| \in \mathcal{T}_\ell$, and therefore

$$\mathsf{span}(\varphi(w)) \leq |\varphi(w[1, \delta'])| - (|\varphi(w[1, \delta - 1])| + 1) = |\varphi(w[\delta, \delta'])| - 1 \leq \ell \cdot (\mathsf{span}(w) + 1).$$

3. If Γ is such that $\max\Gamma = \mathsf{lm}(w)$, then

$$\mathsf{lm}(\varphi(w)) \leq \max\Gamma' = |\varphi(w[1, \max\Gamma])| \leq \ell \cdot \mathsf{lm}(w).$$

To end the proof, we can choose the constant $C = \ell(|\Sigma| + 1)$ (which is independent of w), and the conclusion will follow for all three cases. □

5. Span and leftmost complexities

Based on the two new measures introduced in the previous section, we can define related complexity functions for infinite words, respectively called the *span complexity* and the *leftmost complexity*, which allow us to obtain a finer classification of infinite words. Indeed, Examples 35 and 44 highlight two infinite words, the period-doubling word and the Fibonacci word, which are not distinguishable if we consider their respective string attractor profile function as they are eventually equal to 2. However, the situation is very different if we look at how the positions within a string attractor are arranged.

Definition 34. Let $\mathbf{x}$ be an infinite word. The *span* and *leftmost complexities* of $\mathbf{x}$ are respectively defined by $\mathsf{span}_\mathbf{x}(n) = \mathsf{span}(\mathbf{x}[1, n])$ and $\mathsf{lm}_\mathbf{x}(n) = \mathsf{lm}(\mathbf{x}[1, n])$ for all $n \geq 1$.

The span complexity for the period doubling word is described below.

Example 35. Consider the period-doubling word $\mathbf{pd} = 101110101011\cdots$ described in Example 25 in which we recalled that $s_\mathbf{pd}(n) = 2$ for all $n \geq 2$. It has been proven in [41, Theorem 10] that

$$\mathsf{span}_\mathbf{pd}(n) = \begin{cases} 1, & \text{if } 2 \leq n \leq 5; \\ 2^i, & \text{if } 3 \cdot 2^i \leq n < 3 \cdot 2^{i+1} \text{ for some } i \geq 1. \end{cases}$$

For Holub's words, we can use Example 16 to obtain the span and the leftmost complexities for particular prefixes, as shown below.

Example 36. Consider the word $\mathbf{u}$ from Example 16 in which we proved that, for all $i \geq 0$, the set $\Gamma^{(i)} = \left\{|u_i| + 1, \sum_{k=0}^{i}(|u_k| + 1), 2|u_i| + 2\right\}$ is a string attractor of the length-$|u_{i+1}|$ prefix of $\mathbf{u}$. This directly implies that $\mathsf{span}_{\mathbf{u}}(|u_{i+1}|) \leq \max \Gamma^{(i)} - \min \Gamma^{(i)} = |u_i|+1$ and that $\mathsf{lm}_{\mathbf{u}}(|u_{i+1}|) \leq 2|u_i|+2$. Recall moreover that consecutive occurrences of u_i in $\mathbf{u}$ are separated by at least $|u_i| + 1$ letters. In particular, as $u_{i+1} = u_i 0 (u_i 1)^{n_i} u_{i-1}$ with $n_i \geq 2$, the factor $u_i 0$ only occurs as a prefix in u_{i+1}, and $1u_i 1$ does not occur before position $2|u_i| + 2$. It follows that $\mathsf{span}_{\mathbf{u}}(|u_{i+1}|) = |u_i| + 1$ and that $\mathsf{lm}_{\mathbf{u}}(|u_{i+1}|) = 2|u_i| + 2$.

The next result directly follows from Proposition 32 and establishes the relationship between the profile function, the span and leftmost complexities.

Proposition 37. *For any infinite word $\mathbf{x}$, we have $s_{\mathbf{x}}(n) - 1 \leq \mathsf{span}_{\mathbf{x}}(n) \leq \mathsf{lm}_{\mathbf{x}}(n) - 1$ for all $n \geq 1$.*

As we did for the string attractor profile function, we will now focus on the case where these new complexities are "bounded". More specifically, we will characterize the infinite words such that these complexities are bounded infinitely many times.

We first look at the leftmost complexity. We will use the following intermediate result, which can be deduced from the proofs of [30, Propositions 12 and 15].

Proposition 38. *Let w be a non-empty word and let $u = w^r$, $v = w^s$ be fractional powers of w with $1 \leq r \leq s$. If Γ is a string attractor of u, then $\Gamma \cup \{|w|\}$ is a string attractor of v.*

Proposition 39. *For any infinite word $\mathbf{x}$, the following are equivalent:*

1. *There exists a constant $C \geq 1$ such that $\mathsf{lm}_{\mathbf{x}}(n) \leq C$ for infinitely many n.*
2. *The word $\mathbf{x}$ is eventually periodic.*
3. *The leftmost complexity $\mathsf{lm}_{\mathbf{x}}$ is bounded.*

Proof. The implication (1) $\implies$ (2) follows from Proposition 31. Indeed, for all $m \geq 1$, there exists an integer n such that $\mathsf{lm}_{\mathbf{x}}(n) \leq C$ and $\mathbf{x}[1, n]$ contains all length-m factors. Therefore, $p_{\mathbf{x}}(m) \leq C$. Using Theorem 2, this implies that x is eventually periodic.

The implication (2) $\implies$ (3) follows from Proposition 38. Indeed, if $\mathbf{x} = uv^\omega$, then for all $n \geq 1$, $\{1, 2, \ldots, \min\{n, |uv|\}\}$ is a string attractor for the word $\mathbf{x}[1, n]$. Therefore, $\mathsf{lm}_{\mathbf{x}}(n) \leq |uv|$ for all $n \geq 1$.

The implication (3) $\implies$ (1) is direct. □

This result gives a new characterization of eventually periodic words. Observe that the proof uses the well-known characterization by Morse and Hedlund (Theorem 2). Note that, in the following, we will mostly use the contraposition of Proposition 39.

We now look at a similar description for the span.

Proposition 40. *Let $\mathbf{x}$ be an infinite word. If there exists a constant $C \geq 1$ such that $\mathsf{span}_{\mathbf{x}}(n) \leq C$ for infinitely many n, then $\mathbf{x}$ is eventually periodic, or it is recurrent and there exists $d \leq C$ such that $p_{\mathbf{x}}(n) = n + d$ for all large enough n.*

Proof. Let us suppose that $\mathbf{x}$ is aperiodic. We first show that $\mathbf{x}$ is recurrent. Towards a contradiction, we assume that $\mathbf{x}$ is not recurrent. Therefore, there exists a factor that only occurs once in $\mathbf{x}$. Say that this occurrence ends at position k. This implies that, for all $n \geq k$, any string attractor of $\mathbf{x}[1, n]$ contains a position smaller than or equal to k. As $\mathsf{span}_{\mathbf{x}}(n) \leq C$ for infinitely many n, then $\mathsf{lm}_{\mathbf{x}}(n) \leq k + C$ for infinitely many n, which contradicts Proposition 39.

We now show that $\mathbf{x}$ has the claimed factor complexity. For all $m \geq 1$, there exists an integer n such that $\mathsf{span}_{\mathbf{x}}(n) \leq C$ and $\mathbf{x}[1, n]$ contains all length-m factors. By Proposition 28, we have $p_{\mathbf{x}}(m) \leq m + C$. Using Theorem 2 and as $\mathbf{x}$ is aperiodic, we conclude that $p_{\mathbf{x}}(m) = m + d$ for all large enough m and for some $d \leq C$. □

Note that a converse-like characterization will be given in Theorem 49.

On the other hand, some infinite words have maximal span complexity, as stated in the following result.

Proposition 41. *For any linearly recurrent word* $\mathbf{x}$*, if* $p_{\mathbf{x}}(n) = n + \Omega(n)$*, then* $\mathsf{span}_{\mathbf{x}}(n) = \Theta(n)$.

Proof. Since $\mathbf{x}$ is linearly recurrent, by Remark 4, there exists an integer A such that, for all m, the length-(Am) prefix of $\mathbf{x}$ contains all length-m factors of $\mathbf{x}$. For all n, if m is such that $n \in [Am + 1, A(m + 1)]$, Proposition 28 implies that $\mathsf{span}_{\mathbf{x}}(n) \geq p_{\mathbf{x}}(m) - m$. By assumption on the factor complexity function, we have $p_{\mathbf{x}}(m) \geq Cm$ for a constant $C > 1$. Therefore $\mathsf{span}_{\mathbf{x}}(n) \geq (C - 1)m \geq (C - 1)\,(n/A - 1)$. This shows that $\mathsf{span}_{\mathbf{x}}(n) = \Omega(n)$. But since we trivially have $\mathsf{span}_{\mathbf{x}}(n) = O(n)$, the conclusion follows. □

6. The case of Sturmian words

Sturmian words are famous combinatorial objects having a large number of mathematical properties and characterizations. They also have a geometric description as approximations of straight lines [28, Chapter 2]. Among aperiodic binary infinite words, they are those with minimal factor complexity, i.e., an aperiodic infinite word $\mathbf{x}$ is a *Sturmian word* if $p_{\mathbf{x}}(n) = n + 1$, for all $n \geq 0$. Moreover, Sturmian words are uniformly recurrent.

In this section, we analyze properties of the three new string-attractor related complexities for Sturmian words and two related families of infinite words. On the one hand, we consider the subfamily of *characteristic Sturmian words*, defined as follows: a Sturmian word $\mathbf{s}$ is *characteristic* if both $0\mathbf{s}$ and $1\mathbf{s}$ are Sturmian words. On the other hand, we investigate the superfamily of *quasi-Sturmian* words, which can be considered the simplest generalizations of Sturmian words in terms of factor complexity. Indeed, they are defined as follows [5]: a word $\mathbf{x}$ is *quasi-Sturmian* if there exist integers d and n_0 such that $p_{\mathbf{x}}(n) = n + d$, for each $n \geq n_0$. The infinite words having factor complexity $n + d$ have been also studied in [21] where they are called "words with minimal block growth".

6.1. On the string attractor-based complexities for characteristic Sturmian words

We focus here on the family of characteristic Sturmian words, for which we can explicitly give the string attractor profile function, the span complexity and the leftmost complexity by giving string attractors realizing them.

It is based on the construction of characteristic Sturmian words via particular finite words called *standard Sturmian words*. The latter have many interesting combinatorial properties and appear as extremal cases for several algorithms and data structures [9, 8, 24, 31, 42]. The standard Sturmian words are defined recursively as follows [39].

Definition 42. A *directive sequence* is an infinite sequence of integers $(q_i)_{i\geq 0}$ such that $q_0 \geq 0$ and $q_i \geq 1$ for all $i \geq 1$. The corresponding sequence of *standard Sturmian words* $(x_i)_{i\geq 0}$ is defined by $x_0 = b$, $x_1 = a$, and $x_{i+1} = x_i^{q_{i-1}} x_{i-1}$ for all $i \geq 1$.

The limits $\mathbf{s} = \lim_{i\to\infty} x_i$ of such sequences of standard Sturmian words are exactly the characteristic Sturmian words [28, Proposition 2.2.24]. Note that $\mathbf{s}$ starts with the letter a if and only if $q_0 \geq 1$, and $\mathbf{s}$ starts with the letter b otherwise. We let $E\colon \{a,b\}^* \to \{a,b\}^*$ be the exchange morphism, i.e., $E(a) = b$ and $E(b) = a$. A well-known property of characteristic Sturmian words is the following: $\mathbf{s}$ starts with a letter a and has $(q_i)_{i\geq 0}$ as directive sequence if and only if $E(\mathbf{s})$ starts with a letter b and has $(q'_i)_{i\geq 0}$ as directive sequence with $q'_0 = 0$ and $q'_{i+1} = q_i$ for all $i \geq 0$ [29, Section 2]. Therefore, in what follows, we only consider the case where $q_0 \geq 1$.

The following result shows that each prefix of a characteristic Sturmian word has a smallest string attractor of span 1, i.e., consisting of two consecutive positions.

Theorem 43. *Consider a directive sequence $(q_i)_{i\geq 0}$ with $q_0 \geq 1$, the corresponding sequence $(x_i)_{i\geq 0}$ of standard Sturmian words and the associated characteristic Sturmian word $\mathbf{s} = \lim_{i\to\infty} x_i$ as in Definition 42. Then we have*

$$s_{\mathbf{s}}(n) = \begin{cases} 1, & \text{if } n < |x_2|; \\ 2, & \text{if } n \geq |x_2|; \end{cases} \qquad \mathit{span}_{\mathbf{s}}(n) = \begin{cases} 0, & \text{if } n < |x_2|; \\ 1, & \text{if } n \geq |x_2|; \end{cases}$$

and

$$\mathit{lm}_{\mathbf{s}}(n) = \begin{cases} 1, & \text{if } n < |x_2|; \\ |x_k|, & \text{if } |x_k| + |x_{k-1}| - 1 \leq n \leq |x_{k+1}| + |x_k| - 2 \text{ for some } k \geq 2. \end{cases}$$

More specifically, for all $n \geq 1$, a string attractor for $\mathbf{s}[1,n]$ is given by

$$\Gamma_n = \begin{cases} \{1\}, & \text{if } n < |x_2|; \\ \{|x_k| - 1, |x_k|\}, & \text{if } |x_k| + |x_{k-1}| - 1 \leq n \leq |x_{k+1}| + |x_k| - 2 \text{ for some } k \geq 2. \end{cases}$$

Proof. We start the proof by showing the last part of the statement, i.e., we show that, for all $n \geq 1$, the given Γ_n is a string attractor for $\mathbf{s}[1,n]$. Observe first that, if $n < |x_2|$, then $\mathbf{s}[1,n] = a^n$, so $\{1\}$ is directly a string attractor. For the case $n \geq |x_2|$, we will need the following notations. For all $k \geq 2$, using [29, Theorem 3], we factorize the standard Sturmian word x_k into $x_k = C_k u_k$ where C_k is a palindrome and $u_k = ab$ if k is even and $u_k = ba$ if k is odd. We also recall the following observation from [28, Theorem 2.2.11]: for all $k \geq 2$, since

$$\mathbf{s}[1, |x_{k+1}| + |x_k| - 2] = x_{k+1} C_k = C_{k+1} u_{k+1} C_k,$$

the previous word is periodic of period $|C_k| + 2 = |x_k|$.

Assume now that $n \geq |x_2|$, and let $k \geq 2$ be such that $|x_k| + |x_{k-1}| - 1 \leq n \leq |x_{k+1}| + |x_k| - 2$ (such a k exists since $|x_2| + |x_1| - 1 = |x_2|$). Since $\mathbf{s}[1, |x_{k+1}| + |x_k| - 2]$ is periodic of period $|x_k|$, then it is a fractional power of x_k. Therefore, using Proposition 38, it is enough to show that $\Gamma_n = \{|x_k| - 1, |x_k|\}$ is a string attractor of the length-$(|x_k| + |x_{k-1}| - 1)$ prefix of $\mathbf{s}$, that we will denote p_k.

If $k = 2$ or $k = 3$, the conclusion is direct as $p_2 = x_2 = a^{q_0} b$ and $p_3 = (a^{q_0} b)^{q_1} a a^{q_0}$. If $k \geq 4$, we use the fact that a similar result was proved for the standard Sturmian words in [30,

n	1	2	3	4	5	6	7	8
$\mathbf{f}[1,n]$	$\underline{a}$	$\underline{ab}$	$\underline{ab}a$	$a\underline{ba}a$	$a\underline{baa}b$	$a\underline{baa}ba$	$aba\underline{ab}ab$	$aba\underline{ab}aba$
Γ_n	$\{1\}$	$\{1,2\}$	$\{1,2\}$	$\{2,3\}$	$\{2,3\}$	$\{2,3\}$	$\{4,5\}$	$\{4,5\}$

Table 2: For $n \in [1,8]$, the length-n prefix of the Fibonacci word $\mathbf{f} = abaababaabaab\cdots$ and its leftmost string attractor Γ_n.

Theorem 22]. Namely, Γ_n is a string attractor for x_{k+1}. To show that Γ_n is also a string attractor for p_k, we will show that $w := \mathbf{s}[|x_k|, |p_k|]$ does not occur elsewhere in p_k. Indeed, this will imply that for each factor of p_k its occurrence that was covered by Γ_n in x_{k+1} is an occurrence in p_k (also covered by Γ_n).

Observe that, as $k \geq 4$, $w = cC_{k-1}c$ where c is the last letter of u_k and the first letter of u_{k-1}. Note that w is not a suffix of $\mathbf{s}[1, |p_k| - 1] = x_kC_{k-1}$ as x_k ends with $u_k = dc$, $d \neq c$. Therefore, if w is a factor of x_kC_{k-1}, it is followed by c since x_kC_{k-1} is periodic of period $|x_{k-1}| = |w|$. In particular, $C_{k-1}cc$ and $C_{k-1}cd = x_{k-1}$ are factors of x_kC_{k-1}. This implies that $C_{k-1}c$ is right special and, by [28, Proposition 2.1.23], cC_{k-1} is a prefix of x_k. As $C_{k-1}c$ is also a prefix of x_k, this implies that C_{k-1} is periodic of period 1, a contradiction as $k \geq 4$. This ends the proof that w is not a factor of x_kC_{k-1} and, with it, the proof that Γ_n is a string attractor of $\mathbf{s}[1,n]$.

Moreover, we directly have that Γ_n is of minimal size and of minimal span among the string attractors of $\mathbf{s}[1,n]$. It is also a leftmost string attractor as each string attractor of $\mathbf{s}[1,n]$ will contain a position greater than or equal to $|x_k|$ to cover w. This proves the three claimed complexities. □

Example 44. Consider the infinite Fibonacci word $\mathbf{f} = abaababaabaababaa\cdots$, which is a characteristic Sturmian word with directive sequence $(1)_{i\geq 0}$. In Table 2, for $1 \leq n \leq 8$, we exhibit the length-n prefixes of $\mathbf{f}$ and their respective leftmost string attractor Γ_n. The underlined positions in $\mathbf{f}[1,n]$ correspond to those in Γ_n, while the first few lengths $|x_k|$, $k \in [1,5]$ are given by $\{1,2,3,5,8\}$.

Note that different size-2 string attractors are obtained in Subsection 7.2.

While infinitely many characteristic Sturmian words have the same string attractor profile function (resp., the same span complexity), the leftmost complexity uniquely determines the characteristic Sturmian word (up to exchanging the letters a and b, captured by the exchange morphism E). This is the object of the result below.

Proposition 45. *Let* $\mathbf{s}$ *and* $\mathbf{s}'$ *be two characteristic Sturmian words such that* $\mathsf{lm}_{\mathbf{s}} = \mathsf{lm}_{\mathbf{s}'}$. *Then, either* $\mathbf{s} = \mathbf{s}'$ *or* $\mathbf{s} = E(\mathbf{s}')$.

Proof. As in Definition 42, let $(q_i)_{i\geq 0}$ and $(p_i)_{i\geq 0}$ be two directive sequences and let $(x_i)_{i\geq 0}$ and $(y_i)_{i\geq 0}$ be the corresponding sequences of standard Sturmian words that are prefixes of $\mathbf{s}$ and $\mathbf{s}'$ respectively. Now consider the associated characteristic Sturmian words $\mathbf{s}$ and $\mathbf{s}'$. Due to the observation made after Definition 42, we may assume that, up to exchanging a and b, both $\mathbf{s}$ and $\mathbf{s}'$ start with the letter a (i.e., $q_0, p_0 \geq 1$). The assumption that $\mathsf{lm}_{\mathbf{s}} = \mathsf{lm}_{\mathbf{s}'}$ together with Theorem 43 now implies that the sequences $(|x_i|)_{i\geq 0}$ and $(|y_i|)_{i\geq 0}$ are equal. A simple induction shows that $q_i = p_i$ for all i, therefore $\mathbf{s} = \mathbf{s}'$. □

Observe that Theorem 43 is only true for characteristic Sturmian words since some prefixes of non-characteristic Sturmian words do not admit any string attractor of span 1, as shown in the following example.

Example 46. Let $\mathbf{s} = aaaaaabaaaaaabaaaaaaab\cdots$ be a characteristic Sturmian word whose directive sequence begins with $q_0 = 6$ and $q_1 = 2$ and let $\mathbf{x}$ be the non-characteristic Sturmian word such that $\mathbf{s} = aaaa \cdot \mathbf{x}$, hence $\mathbf{x} = aabaaaaaabaaaaaaab\cdots$. We consider the prefix $\mathbf{x}[1, 14] = aabaaaaaabaaaa$. Since b occurs only at positions 3 and 10 and the factor $aaaaaa$ only in $\mathbf{x}[4, 9]$, the candidates as string attractor with two consecutive positions are $\Gamma_1 = \{3, 4\}$ and $\Gamma_2 = \{9, 10\}$. However, one can check that the factors $aaab$ and $baaaaa$ do not cross any position in Γ_1 and Γ_2 respectively, showing that $\mathsf{span}_{\mathbf{x}}(14) \geq 2$. Nonetheless, $\mathbf{x}'[1, 14]$ admits a string attractor of size 2 (but with a larger span), i.e., $\Gamma = \{4, 10\}$.

6.2. Characterization of Sturmian and quasi-Sturmian words

We now turn to the families of Sturmian and quasi-Sturmian words. For each, we provide a new characterization in terms of both the span and leftmost complexities.

We start off with Sturmian words.

Theorem 47. *An infinite word* $\mathbf{x}$ *is Sturmian if and only if* $\mathsf{lm}_{\mathbf{x}}$ *is unbounded and* $\mathsf{span}_{\mathbf{x}}(n) = 1$ *for infinitely many* $n \geq 1$.

Proof. For the first implication, let $\mathbf{x}$ be a Sturmian word. Since $\mathbf{x}$ is aperiodic, Proposition 39 shows that $\mathsf{lm}_{\mathbf{x}}$ satisfies the statement. We now establish the claimed property on $\mathsf{span}_{\mathbf{x}}$. As $\mathbf{x}$ is aperiodic and recurrent, it has infinitely many right special prefixes. Moreover, for every such prefix v, there is a characteristic Sturmian word $\mathbf{s}$ having v^R as a prefix [28, Proposition 2.1.23]. Therefore, $\mathsf{span}(v) = \mathsf{span}(v^R) = 1$ for all long enough v by Theorem 43 and the proof of [30, Proposition 11].

For the other implication, consider an infinite word $\mathbf{x}$ satisfying the assumptions. First, it is aperiodic by Proposition 39. Moreover, by assumption, for all $m \geq 1$, there exists an integer n such that $\mathbf{x}[1, n]$ contains all length-m factors and $\mathsf{span}_{\mathbf{x}}(n) = 1$. Therefore, $p_{\mathbf{x}}(m) \leq m + 1$ by Proposition 28. The fact that $\mathbf{x}$ is Sturmian follows from Theorem 2. □

We now turn to quasi-Sturmian words. As announced, we prove a sort of converse of Proposition 40. We will make use of the following characterization of quasi-Sturmian words [5].

Theorem 48 ([5])**.** *An infinite word* $\mathbf{x}$ *over the alphabet* Σ *is quasi-Sturmian if and only if it can be written as* $\mathbf{x} = w\varphi(\mathbf{s})$*, where* w *is a finite word,* $\mathbf{s}$ *is a Sturmian word on the alphabet* $\{a, b\}$*, and* φ *is a morphism from* $\{a, b\}^*$ *to* Σ^* *such that* $\varphi(ab) \neq \varphi(ba)$.

Theorem 49. *An infinite word* $\mathbf{x}$ *is quasi-Sturmian if and only if* $\mathsf{lm}_{\mathbf{x}}$ *is unbounded and there exist a suffix* $\mathbf{y}$ *of* $\mathbf{x}$ *and a constant* $C' \geq 1$ *such that* $\mathsf{span}_{\mathbf{y}}(n) \leq C'$ *for infinitely many* $n \geq 1$.

Proof. For the first implication, as quasi-Sturmian words are aperiodic by Theorem 2, $\mathsf{lm}_{\mathbf{x}}$ is unbounded by Proposition 39. In addition, by Theorem 48, there exists a finite word w, a Sturmian word $\mathbf{s}$, and a morphism φ such that $\mathbf{x} = w\varphi(\mathbf{s})$. Consider the suffix $\mathbf{y} = \varphi(\mathbf{s})$. By Theorem 47, there are infinitely many integers n such that $\mathsf{span}_{\mathbf{s}}(n) = 1$, and by Proposition 33, there exists a constant $C' \geq 1$ such that, for all $N = |\varphi(\mathbf{s}[1, n])|$,

$$\mathsf{span}_{\mathbf{y}}(N) = \mathsf{span}(\varphi(\mathbf{s}[1, n])) \leq C' \cdot \mathsf{span}(\mathbf{s}[1, n]) = C'.$$

For the other implication, by Propositions 39 and 40, $p_{\mathbf{y}}(n) = n + d$ with $d \leq C'$ for all large enough n. Since $\mathbf{x} = w\mathbf{y}$ for some finite word w, we have $p_{\mathbf{x}}(n) \leq p_{\mathbf{y}}(n) + |w| = n + d + |w|$ for all large enough n. We conclude by Theorem 2 that x is quasi-Sturmian. □

7. String attractors and complexities for k-bonacci words

In this section, we study string attractors of prefixes of some purely morphic words over an alphabet of size $k \geq 2$, namely the so-called *k-bonacci words*. The case $k = 2$ corresponds to the famous Fibonacci word[4], which is a Sturmian word and for which string-attractor related concepts have already been studied. For $k = 3$, each prefix of the Tribonacci word admits a string attractor of size at most 3 as shown in [41].

More generally, as k-bonacci words are *episturmian*, L. Dvořáková showed that each prefix admits a string attractor of size at most k [17, Theorem 10]. However, this result is not constructive in the sense that the string attractors are not explicitly given. Our contribution is to provide a constructive description of a string attractor of size at most k for each prefix. Our new approach differs from the techniques used to obtain string attractors for the Thue–Morse word, the period-doubling word, and standard Sturmian words, and may be extended to other purely morphic words. Moreover, we precisely describe our string attractors in terms of the corresponding k-bonacci numbers, which opens the door to considerations related to numeration systems. In fact, a first attempt towards these considerations was done in [20] using a similar construction.

Furthermore we then study the leftmost and the span complexities of the k-bonacci words.

7.1. Useful definitions and intermediate results

Let us consider an integer $k \geq 2$ and the morphism $\mu_k \colon \{0, \dots, k-1\}^* \to \{0, \dots, k-1\}^*$ defined by $\mu_k(i) = 0(i+1)$ for all $i \in \{0, 1, \dots, k-2\}$ and $\mu_k(k-1) = 0$. The *infinite k-bonacci word* $\mathbf{b}^{(k)}$ is defined as the fixed-point $\mathbf{b}^{(k)} = \mu_k^\infty(0)$. The cases $k = 2$ and $k = 3$ correspond to the Fibonacci and Tribonacci words respectively.

Furthermore, for all $n \geq 0$, we let $b_n^{(k)} = \mu_k^n(0)$ denote the *nth finite k-bonacci word*. We also set $b_n^{(k)} = \varepsilon$ for all $-k \leq n < 0$. For any $n \geq 0$, we let $B_n^{(k)} = |b_n^{(k)}|$ denote the length of the nth finite k-bonacci word. The sequence $(B_n^{(k)})_{n \geq 0}$ will be referred to as the sequence of *k-bonacci numbers*. When the context is clear, we will drop the superscript (k) in all of these notations.

Example 50. For $k = 3$, we write the first few non empty finite Tribonacci words in Table 3.

n	0	1	2	3	4	5
$b_n^{(3)}$	0	0 1	01 0 2	0102 01 0	0102010 0102 01	0102010010201 0102010 0102

Table 3: The first few finite Tribonacci words $(b_n^{(3)})_{0 \leq n \leq 5}$ (some particular decomposition is highlighted for a latter purpose, see Proposition 51).

[4]Note that in Example 44, the Fibonacci word is defined on the alphabet $\{a, b\}$ to match the general definition of Sturmian words. In this section, for the sake of simplicity, we define it on $\{0, 1\}$ instead.

Another way of seeing the sequence $(b_n^{(k)})_{n\geq -k}$ is the following, which can be proven by an easy induction. See Table 3 for an example with $k = 3$.

Proposition 51. *We have*

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

We now define two sequences of integers $(L_n^{(k)})_{n\geq 0}$ and $(U_n^{(k)})_{n\geq 0}$ linked to k-bonacci numbers that will help us partition $\mathbb{N}$.

Definition 52. For all $n \geq 0$, we set

$$L_n^{(k)} = \begin{cases} B_n^{(k)}, & \text{if } n \leq k; \\ B_n^{(k)} + B_{n-k-1}^{(k)} - 1, & \text{otherwise;} \end{cases}$$

and

$$U_n^{(k)} = \sum_{i=0}^{n} B_i^{(k)}.$$

Example 53. When $k = 3$, we obtain $(L_n^{(3)})_{n\geq 0} = 1, 2, 4, 7, 13, 25, 47, 87, \ldots$ and $(U_n^{(3)})_{n\geq 0} = 1, 3, 7, 14, 27, 51, 95, 176, \ldots$.

For any $k \geq 2$, one can show that $(U_n^{(k)})_{n\geq 0}$ gives the lengths of palindromic prefixes of the k-bonacci word (note that the case $k = 3$ gives the sequence [43, A027084]).

Remark 54. Observe that, by Proposition 51, if $1 \leq n \leq k-1$, then $U_{n-1} = B_n - 1 = L_n - 1$ and if $n = k$, then $U_{k-1} = B_k = L_k$. Moreover, for $n \geq k$, we have $L_n \leq U_{n-1}$ as the case $n = k$ is above, and for $n > k$, we have $L_n^{(k)} = \left(\sum_{i=n-k-1}^{n-1} B_i^{(k)}\right) - 1$. As $L_0 = 1$, this implies that the intervals $[L_n, U_n]$, $n \geq 0$ cover the set of integers $m \geq 1$.

7.2. String attractor profile function

We now study the string attractor profile function of the k-bonacci word $\mathbf{b}^{(k)}$. To do so, we will make use of Proposition 38 therefore we look at prefixes which are fractional powers of another. More specifically, as the string attractors positions will be elements of $(B_n)_{n\geq 0}$, we study fractional powers of the words b_n, $n \geq 0$.

Proposition 55. *For all $n \geq 0$, $\mathbf{b}^{(k)}[1, U_n^{(k)}] = \prod_{i=0}^{n} b_{n-i}^{(k)}$. Moreover, $\mathbf{b}^{(k)}[1, U_n^{(k)}]$ is a fractional power of $b_n^{(k)}$.*

Proof. For $n = 0$, we directly have $\mathbf{b}[1, U_0] = \mathbf{b}[1, 1] = b_0$, so both claims hold in this case. Assume now that the result is true for n and let us prove it for $n+1$. By the induction hypothesis, we have

$$\mu_k(\mathbf{b}[1, U_n]) = \mu_k\left(\prod_{i=0}^{n} b_{n-i}\right) = \prod_{i=0}^{n} b_{n+1-i}.$$

As $\mathbf{b}$ is a fixed point of μ_k, $\mu_k(\mathbf{b}[1, U_n])$ is a prefix of $\mathbf{b}$ and it is followed by the image of a letter. Thus it is followed by the letter 0, and

$$\mathbf{b}[1, U_{n+1}] = \mathbf{b}\left[1, \sum_{i=0}^{n+1} B_i\right] = \left(\prod_{i=0}^{n} b_{n+1-i}\right) \cdot 0 = \prod_{i=0}^{n+1} b_{n+1-i}.$$

Moreover, since $\mathbf{b}[1, U_n]$ is a fractional power of b_n by the induction hypothesis, so is $\mathbf{b}[1, U_n] \cdot a$ for some letter $a \in \{0, 1, \ldots, k-1\}$. By applying the morphism μ_k on both words, we can conclude that $\mathbf{b}[1, U_{n+1}] = \mu_k(\mathbf{b}[1, U_n]) \cdot 0$ is a fractional power of $b_{n+1} = \mu_k(b_n)$. □

Using Proposition 38, we then directly have the following corollary.

Corollary 56. *For all $n \geq 0$, if Γ is a string attractor for $\mathbf{b}^{(k)}[1, L_n^{(k)}]$ and if $B_n^{(k)} \in \Gamma$, then Γ is a string attractor for $\mathbf{b}^{(k)}[1, m]$ for all $m \in [L_n^{(k)}, U_n^{(k)}]$.*

We now exhibit a minimum string attractor of size at most k for each prefix of $\mathbf{b}^{(k)}$ and deduce the string attractor profile function.

Theorem 57. *For all $n \geq 0$, the set*

$$\Gamma_n = \begin{cases} \{B_0^{(k)}, \ldots, B_n^{(k)}\}, & \text{if } n \leq k-1; \\ \{B_{n-k+1}^{(k)}, \ldots, B_n^{(k)}\}, & \text{if } n \geq k; \end{cases}$$

is a minimum string attractor for $\mathbf{b}^{(k)}[1, m]$, for all $m \in [L_n^{(k)}, U_n^{(k)}]$. In particular, the string attractor profile function for $\mathbf{b}^{(k)}$ is given by

$$s_{\mathbf{b}^{(k)}}(n) = \begin{cases} i+1, & \text{if } B_i^{(k)} \leq n < B_{i+1}^{(k)} \text{ for some } i \leq k-2; \\ k, & \text{if } n \geq B_{k-1}^{(k)}. \end{cases}$$

Proof. Using Proposition 51, a simple induction shows that, for all $n \geq 0$, the positions of Γ_n correspond to different letters, which implies that, if Γ_n is a string attractor of a prefix, it is minimum. We prove that it is a string attractor of the length-m prefix, $m \in [L_n, U_n]$, by induction on $n \geq 0$. More precisely, the induction step is divided into three intermediary claims:

1. $\Gamma_{n-1} \cup \{B_n\}$ is a string attractor for $\mathbf{b}[1, L_n]$;
2. Γ_n is a string attractor for $\mathbf{b}[1, L_n]$;
3. Γ_n is a string attractor for $\mathbf{b}[1, m]$ for all $m \in [L_n, U_n]$.

Let us prove the first claim. If $n = 0$, then we take the convention that $\Gamma_{-1} = \emptyset$ and directly conclude that $\{1\}$ is a string attractor for $\mathbf{b}[1, 1]$. Assume now that $n \geq 1$. Then $L_n = U_{n-1} + 1 = B_n$ if $n \leq k-1$, or $L_n \in [L_{n-1}, U_{n-1}]$ if $n \geq k$. Therefore, by the induction hypothesis, $\Gamma_{n-1} \cup \{B_n\}$ is a string attractor for $\mathbf{b}[1, L_n]$.

Let us prove the second claim. If $n \leq k-1$, then $\Gamma_n = \Gamma_{n-1} \cup \{B_n\}$ so the conclusion directly follows from the first claim. Assume then that $n \geq k$ and let us denote $\mathbf{b}[1, L_n] = b_n u$, where $u = \varepsilon$ if $n = k$ or u is b_{n-k-1} without its last letter if $n \geq k+1$. Using the first claim, it remains to show that the position B_{n-k} is not needed in the string attractor, i.e., the factors of $\mathbf{b}[1, L_n]$ that are covered by position B_{n-k} are still covered by Γ_n. As the first position in Γ_n is B_{n-k+1}, it suffices to consider the factor occurrences crossing position B_{n-k} in $\mathbf{b}[1, B_{n-k+1} - 1]$. As

$\mathbf{b}[1, B_{n-k+1} - 1]$ is b_{n-k+1} without its last letter, Proposition 51 implies that they are occurrences in

$$\prod_{i=1}^{k} b_{n-k+1-i} = b_{n-k} b_{n-k-1} \prod_{i=3}^{k} b_{n-k+1-i}.$$

Note that $b_{n-k}u$ is a prefix of this word. We consider two cases: either the considered occurrence is entirely contained in $b_{n-k}u$ or it crosses position $B_{n-k} + B_{n-k-1}$. Observe that, if $n \geq k+1$, these two cases are mutually exclusive.

Case 1. Since b_{n-k} is a suffix of b_n by Proposition 51, the factors having an occurrence in $b_{n-k}u$ crossing position B_{n-k} have an occurrence in $b_n u$ crossing position B_n, so they are covered by Γ_n. See Figure 1.

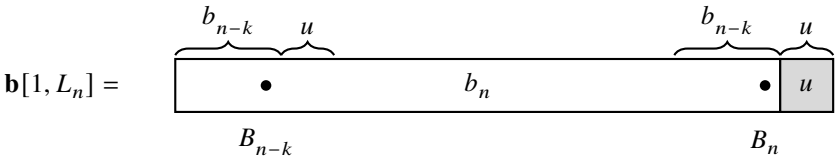

Figure 1: Case 1 in the proof of Theorem 57.

Case 2. Similarly, by Proposition 51, $b_{n-k}b_{n-k-1}$ is a suffix of b_{n-1} and $\prod_{i=3}^{k} b_{n-k+1-i} = \prod_{i=1}^{k-2} b_{n-k-1-i}$ is a prefix of b_{n-k-1}, so of b_{n-2} (as the finite k-bonacci words are prefixes of each other). As $b_{n-1}b_{n-2}$ is a prefix of b_n, we conclude that the factors having an occurrence in $\mathbf{b}[1, B_{n-k+1} - 1]$ crossing position $B_{n-k} + B_{n-k-1}$ have an occurrence in b_n crossing position B_{n-1}, so they are covered by Γ_n. See Figure 2. This ends the proof of the second claim.

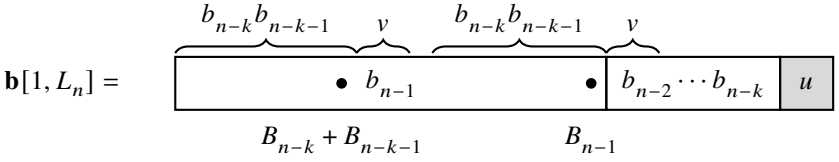

Figure 2: Case 2 in the proof of Theorem 57 with $v = \prod_{i=3}^{k} b_{n-k+1-i}$.

The third claim is a direct consequence of the second claim and of Corollary 56, and this ends the proof that, for all $n \geq 0$ and for all $m \in [L_n, U_n]$, Γ_n is a string attractor of the length-m prefix of $\mathbf{b}$. Finally, the string attractor profile function follows from Remark 54. □

Remark 58. Observe that, in the Tribonacci case, the elements in our string attractors are the same as in [41, Theorem 6]. The corresponding intervals of prefix lengths are also linked. Indeed, our sequence $(U_n^{(3)})_{n\geq 0}$ is related to the sequence $(W_n)_{n\geq 4}$ defined in [41, Theorem 6] as follows: we have $W_{n+3} = U_{n+1}^{(3)}$ for all $n \geq 1$. Therefore, our upper bounds and that of Schaeffer and Shallit coincide. However, our lower bounds are smaller than theirs. On the other hand, the string attractors obtained for the palindromic prefixes in [17] are different from ours. For instance, the case $k = 3$ is treated in [17, Example 8].

7.3. Leftmost complexity

We can further prove that the string attractor from Theorem 57 is actually a leftmost string attractor. For the purpose of the next few results, we set $U_{-1}^{(k)} = 0$.

Proposition 59. *The leftmost complexity of* $\mathbf{b}^{(k)}$ *satisfies* $lm_{\mathbf{b}^{(k)}}(m) = B_n^{(k)}$ *for all* $n \geq 0$ *and* $m \in [U_{n-1}^{(k)} + 1, U_n^{(k)}]$.

Proof. We show that the factor $\mathbf{b}[B_n, U_{n-1} + 1]$ does not occur in $\mathbf{b}$ before position B_n. This implies that, for all $m \geq U_{n-1} + 1$, any string attractor of $\mathbf{b}[1, m]$ contains a position at least equal to B_n and, combined with Theorem 57, proves the claimed leftmost complexity.

The claim is direct for $n = 0$ as $B_0 = 1 = U_{-1} + 1$. Assume now that it is true for n and let us prove it for $n + 1$. By construction, the B_{n+1}th letter of $\mathbf{b}$ is the last letter of the image of the B_nth letter under μ_k, and, by Proposition 55, $\mathbf{b}[B_{n+1} + 1, U_n + 1]$ is the image of $\mathbf{b}[B_n + 1, U_{n-1} + 1]$, potentially followed by a letter 0 (this occurs when $\mathbf{b}[B_n, U_{n-1} + 1]$ ends with the letter $k - 1$). Therefore, each occurrence of $\mathbf{b}[B_{n+1}, U_n + 1]$ in $\mathbf{b}$ is associated with the image of an occurrence of $\mathbf{b}[B_n, U_{n-1} + 1]$. Using the induction hypothesis, we conclude that $\mathbf{b}[B_{n+1}, U_n + 1]$ does not occur before position B_{n+1}. □

7.4. Span complexity

For the k-bonacci words $\mathbf{b}^{(k)}$, the factor complexity function is given by $p_{\mathbf{b}^{(k)}}(n) = (k-1)n+1$. Therefore, when $k \geq 3$, Proposition 41 implies that the span complexity is linear. However, the string attractors described in Section 7.2 do not have the smallest difference between their extreme positions. In what follows, we compute the span for infinitely many prefixes and describe string attractors (of unbounded size) having that span.

We first make the following observation which gives a lower bound on the span. Recall that we have set $U_{-1}^{(k)} = 0$.

Proposition 60. *Let* $k \geq 2$. *For all* $n \geq 2$, *the factors* $\mathbf{b}^{(k)}[i, i + U_{n-3}^{(k)}]$ *are distinct for all* $i \in [B_{n-2}^{(k)} + 1, B_n^{(k)}]$.

Proof. Let us prove the result by induction on n. For $n = 2$, we need to consider the letters in $u = \mathbf{b}[2, B_2]$. If $k = 2$, then $u = 10$ and if $k \geq 3$, then $u = 102$ so all the letters are indeed distinct.

Let us now assume that the claim is true for $n \geq 2$ and let us prove it for $n + 1$. We proceed by contradiction and assume that there exist $i, j \in [B_{n-1} + 1, B_{n+1}]$ minimal such that $i < j$ and $\mathbf{b}[i, i + U_{n-2}] = \mathbf{b}[j, j + U_{n-2}]$. As $B_{n-1} + 1$ marks the beginning of the image of a letter in $\mathbf{b}$ and i and j are taken minimal, we know that the factor $u = \mathbf{b}[i, i + U_{n-2}] = \mathbf{b}[j, j + U_{n-2}]$ begins with 0. We may also assume that it does not end with 0. Indeed, otherwise, we consider the word $u = \mathbf{b}[i, i + U_{n-2} - 1] = \mathbf{b}[j, j + U_{n-2} - 1]$ instead.

As the word u starts with 0, there exist $i' < j'$ such that $\mu_k(\mathbf{b}[1, i' - 1]) = \mathbf{b}[1, i - 1]$ and $\mu_k(\mathbf{b}[1, j' - 1]) = \mathbf{b}[1, j - 1]$. Moreover, as $U_{n-2} + 1 \geq 2$ and as u does not end with a 0, it can be uniquely desubstituted (i.e., its preimage under μ_k is unique). There thus exists ℓ such that $\mathbf{b}[i', i' + \ell] = \mathbf{b}[j', j' + \ell]$ and $\mu_k(\mathbf{b}[i', i' + \ell]) = u$.

As $|\mu_k(\mathbf{b}[1, i' - 1])| = i - 1 \in [B_{n-1}, B_{n+1} - 1]$, we have $i' \in [B_{n-2} + 1, B_n]$. The same holds for j'. Therefore, by the induction hypothesis, we have $\mathbf{b}[i', i' + U_{n-3}] \neq \mathbf{b}[j', j' + U_{n-3}]$. Let us take $\ell' \in [\ell, U_{n-3} - 1]$ maximal such that $\mathbf{b}[i', i' + \ell'] = \mathbf{b}[j', j' + \ell']$ and let us denote $v = \mathbf{b}[i', i' + \ell']$. By maximality of ℓ', v is right special. Moreover, the set of factors of $\mathbf{b}$ is stable under reversal [12, Theorem 5], i.e., the reversal of any factor of $\mathbf{b}$ is also a factor. In

particular, v^R is a left special factor of $\mathbf{b}$. Furthermore, the left special factors of $\mathbf{b}$ are exactly its prefixes [12, Proposition 5], so v^R is a prefix of $\mathbf{b}$ and also of $\mathbf{b}[1, U_{n-3}]$ as $\ell' \leq U_{n-3} - 1$. However, we have

$$|\mu_k(v^R)| = |\mu_k(v)| \geq |\mu_k(\mathbf{b}[i', i' + \ell])| \geq U_{n-2}$$

by definition of ℓ. This is a contradiction as, by Proposition 55, we have $|\mu_k(v^R)| \leq |\mu_k(\mathbf{b}[1, U_{n-3}])| < U_{n-2}$. □

We now describe a new string attractor for prefixes of the k-bonacci word.

Proposition 61. *Let $k \geq 2$. For all $n \geq 1$ and for all $m \in [U_{n-1}^{(k)} + 1, U_n^{(k)}]$, $\Gamma_n = \{U_{n-2}^{(k)} + 1, U_{n-2}^{(k)} + 2, \ldots, B_n^{(k)}\}$ is a string attractor of $\mathbf{b}^{(k)}[1, m]$.*

Proof. We proceed by induction on $n \geq 1$. For the base case $n = 1$, the interval $[U_{1-1} + 1, U_2]$ becomes $[2, 3]$, and $\Gamma_1 = \{1, 2\}$, so the conclusion follows.

Now assume that the result is true for $n \geq 1$ and we show it also holds for $n + 1$. To do so, we will use the following observation. From Proposition 55 and [2, Proposition 4.4], one may prove that $\mathbf{b}[1, U_n]$ is a palindrome for all $n \geq -1$. By the induction hypothesis, Γ_n is a string attractor for $\mathbf{b}[1, U_n]$. As this word is a palindrome, it also has the string attractor

$$\Gamma_n^R = \{U_n + 1 - B_n, \ldots, U_n + 1 - U_{n-2} - 1\} = \{U_{n-1} + 1, \ldots, B_n + B_{n-1}\}.$$

In particular, $\Gamma_{n+1} \supseteq \Gamma_n^R$ is a string attractor of $\mathbf{b}[1, U_n]$ when $B_{n+1} \leq U_n$. If $B_{n+1} > U_n$, then $n \leq k - 1$ and $B_{n+1} = U_n + 1$, so Γ_{n+1} is a string attractor of $\mathbf{b}[1, U_n + 1]$. In both cases, Propositions 38 and 55 imply that Γ_{n+1} is a string attractor of $\mathbf{b}[1, m]$ for all $m \in [U_n+1, U_{n+1}]$. □

Corollary 62. *Let $k \geq 2$. For all $n \geq 2$ and for all $m \in [U_n^{(k)} - B_{n-1}^{(k)} - B_{n-2}^{(k)}, U_n^{(k)}]$, we have $\mathsf{span}_{\mathbf{b}^{(k)}}(m) = B_n^{(k)} - U_{n-2}^{(k)} - 1$. In particular, for infinitely many prefixes, there is a factor length for which the bound given by Proposition 28 is tight.*

Proof. Using Propositions 60 and 28, we know that for $m \geq B_n + U_{n-3}$, we have $\mathsf{span}_{\mathbf{b}}(m) \geq B_n - B_{n-2} - U_{n-3} - 1 = B_n - U_{n-2} - 1$. Observe that $B_n + U_{n-3} = U_n - B_{n-1} - B_{n-2}$. On the other hand, using Proposition 61, we know that for $m \in [U_{n-1} + 1, U_n]$, we have $\mathsf{span}_{\mathbf{b}}(m) \leq B_n - U_{n-2} - 1$.

If $k \geq 3$, then $B_n + U_{n-3} \geq U_{n-1} + 1$ so, for all $m \in [U_n - B_{n-1} - B_{n-2}, U_n]$, we have $\mathsf{span}_{\mathbf{b}}(m) = B_n - U_{n-2} - 1$, as desired. It remains to consider $k = 2$. In that case, $B_n - U_{n-2} - 1 = 1$ which does not depend on n. Therefore $\mathsf{span}_{\mathbf{b}}(m) \geq 1$ for all $m \geq B_2 + U_{-1} = 3$ and $\mathsf{span}_{\mathbf{b}}(m) \leq 1$ for all $m \geq U_0 + 1 = 2$. Therefore, the conclusion follows for all $m \geq 3$. □

Observe that, for the Fibonacci word, we once again obtain that $\mathsf{span}_{\mathbf{b}^{(2)}} = 1$, as in Theorem 43.

8. Conclusions

In this paper, we have shown the close relationship between string attractor based measures and classical notions of repetitiveness on infinite words, like the factor complexity and the recurrence function. In particular, we identify some of the combinatorial properties that an infinite word needs in order to have a bounded string attractor profile function. Nonetheless, a complete characterization of these words is still missing. Furthermore, we have used the new leftmost and span complexities to obtain novel characterizations of infinite words, such as

periodic and aperiodic words, and the families of Sturmian and quasi-Sturmian words. We wonder if other measures based on the distribution of the positions of a string attractor can be used to characterize other combinatorial properties or families of words. Finally, for the k-bonacci words we have shown how to construct for each prefix a string attractor with minimum size, minimum leftmost measure, or minimum span. The methods presented here rely on the properties that k-bonacci words inherit from their morphic construction. A future direction of research could be a generalization of such a strategy to extend the construction of a smallest string attractor to other families of morphic sequences.

Acknowledgements

We thank Julien Leroy for the fruitful discussion on the S-adic words which led to the construction of the infinite word from Example 15.

Funding: This work was supported by the Fonds de la Recherche Scientifique – FNRS [grant number 1.C.104.24F]; and Italian Ministry of University and Research - MUR [grant number 2022YRB97K].

References

[1] Allouche, J., Shallit, J.O., 2003. Automatic Sequences - Theory, Applications, Generalizations. Cambridge University Press.

[2] Ambrož, P., Masáková, Z., Pelantová, E., Frougny, C., 2006. Palindromic complexity of infinite words associated with simple parry numbers, in: Annales de l'institut Fourier, pp. 2131–2160.

[3] Béal, M., Perrin, D., Restivo, A., 2021. Decidable problems in substitution shifts. doi:10.48550/arXiv.2112.14499.

[4] Cassaigne, J., 1997a. Complexité et facteurs spéciaux. Bulletin of the Belgian Mathematical Society-Simon Stevin 4, 67–88.

[5] Cassaigne, J., 1997b. Sequences with grouped factors, in: Developments in Language Theory, Aristotle University of Thessaloniki. pp. 211–222.

[6] Cassaigne, J., 2001. Recurrence in infinite words, in: STACS, Springer. pp. 1–11.

[7] Cassaigne, J., Nicolas, F., 2010. Factor complexity, in: Berthé, V., Rigo, M. (Eds.), Combinatorics, Automata and Number Theory. Cambridge University Press. volume 135, pp. 163–247.

[8] Castiglione, G., Restivo, A., Sciortino, M., 2008. Hopcroft's algorithm and cyclic automata, in: LATA, Springer. pp. 172–183.

[9] Castiglione, G., Restivo, A., Sciortino, M., 2009. Circular sturmian words and hopcroft's algorithm. Theor. Comput. Sci. 410, 4372–4381.

[10] Constantinescu, S., Ilie, L., 2007. The lempel–ziv complexity of fixed points of morphisms. SIAM J. Discret. Math. 21, 466–481.

[11] Dolce, F., 2023. String attractors for factors of the thue-morse word, in: WORDS, Springer. pp. 117–129.

[12] Droubay, X., Justin, J., Pirillo, G., 2001. Episturmian words and some constructions of de luca and rauzy. Theor. Comput. Sci. 255, 539–553.

[13] Durand, F., 2000. Linearly recurrent subshifts have a finite number of non-periodic subshift factors. Ergodic Theory and Dynamical Systems 20, 1061 – 1078.

[14] Durand, F., 2003. Corrigendum and addendum to: "Linearly recurrent subshifts have a finite number of non-periodic subshift factors" [Ergodic Theory Dynam. Systems **20** (2000), no. 4, 1061–1078; MR1779393 (2001m:37022)]. Ergodic Theory Dynam. Systems 23, 663–669. URL: https://doi.org/10.1017/S0143385702001293, doi:10.1017/S0143385702001293.

[15] Durand, F., Host, B., Skau, C., 1999. Substitutional dynamical systems, Bratteli diagrams and dimension groups. Ergodic Theory and Dynamical Systems 19, 953–993.

[16] Durand, F., Perrin, D., 2022. Dimension Groups and Dynamical Systems: Substitutions, Bratteli Diagrams and Cantor Systems. Cambridge Studies in Advanced Mathematics, Cambridge University Press.

[17] Dvořáková, L., 2022. String attractors of episturmian sequences. doi:10.48550/arXiv.2211.01660.

[18] Dvořáková, L., Hendrychová, V., 2023. String attractors of Rote sequences. doi:10.48550/arXiv.2308.00850.

[19] Frosini, A., Mancini, I., Rinaldi, S., Romana, G., Sciortino, M., 2022. Logarithmic equal-letter runs for BWT of purely morphic words, in: DLT, Springer. pp. 139–151.
[20] Gheeraert, F., Romana, G., Stipulanti, M., 2023. String attractors of fixed points of k-bonacci-like morphisms, in: WORDS, Springer. pp. 192–205.
[21] Heinis, A., 2004. Languages under substitutions and balanced words. Journal de théorie des nombres de Bordeaux 16, 151–172.
[22] Holub, S., 2014. Words with unbounded periodicity complexity. Int. J. Algebra Comput. 24, 827–836.
[23] Kempa, D., Prezza, N., 2018. At the roots of dictionary compression: string attractors, in: STOC, ACM. pp. 827–840.
[24] Knuth, D.E., Jr., J.H.M., Pratt, V.R., 1977. Fast pattern matching in strings. SIAM J. Comput. 6, 323–350.
[25] Kociumaka, T., Navarro, G., Prezza, N., 2023. Toward a definitive compressibility measure for repetitive sequences. IEEE Trans. Inf. Theory 69, 2074–2092.
[26] Kutsukake, K., Matsumoto, T., Nakashima, Y., Inenaga, S., Bannai, H., Takeda, M., 2020. On repetitiveness measures of thue-morse words, in: SPIRE, Springer. pp. 213–220.
[27] Lempel, A., Ziv, J., 1976. On the complexity of finite sequences. IEEE Trans. Inf. Theory 22, 75–81.
[28] Lothaire, M., 2002. Algebraic combinatorics on words. volume 90. Cambridge University Press.
[29] de Luca, A., Mignosi, F., 1994. Some combinatorial properties of sturmian words. Theor. Comput. Sci. 136, 361–285.
[30] Mantaci, S., Restivo, A., Romana, G., Rosone, G., Sciortino, M., 2021. A combinatorial view on string attractors. Theor. Comput. Sci. 850, 236–248.
[31] Mantaci, S., Restivo, A., Sciortino, M., 2003. Burrows-wheeler transform and sturmian words. Inf. Process. Lett. 86, 241–246.
[32] Morse, M., Hedlund, G.A., 1938. Symbolic dynamics. American Journal of Mathematics 60, 815–866.
[33] Mousavi, H., 2016. Automatic theorem proving in Walnut. doi:10.48550/arXiv.1603.06017.
[34] Navarro, G., 2022a. The compression power of the BWT: technical perspective. Commun. ACM 65, 90.
[35] Navarro, G., 2022b. Indexing highly repetitive string collections, part I: repetitiveness measures. ACM Comput. Surv. 54, 29:1–29:31.
[36] Navarro, G., 2022c. Indexing highly repetitive string collections, part II: compressed indexes. ACM Comput. Surv. 54, 26:1–26:32.
[37] Pansiot, J., 1984. Complexité des facteurs des mots infinis engendrés par morphimes itérés, in: ICALP, Springer. pp. 380–389.
[38] Pansiot, J., 1986. Decidability of periodicity for infinite words. RAIRO Theor. Informatics Appl. 20, 43–46.
[39] Rauzy, G., 1985. Mots infinis en arithmetique, in: Nivat, M., Perrin, D. (Eds.), Automata on infinite word. Springer Berlin Heidelberg. volume 192, pp. 164–171.
[40] Restivo, A., Romana, G., Sciortino, M., 2022. String attractors and infinite words, in: LATIN, Springer. pp. 426–442.
[41] Schaeffer, L., Shallit, J., 2021. String attractors for automatic sequences. doi:10.48550/arXiv.2012.06840.
[42] Sciortino, M., Zamboni, L.Q., 2007. Suffix automata and standard sturmian words, in: Developments in Language Theory, Springer. pp. 382–398.
[43] Sloane, N.J.A., 1964. The On-Line Encyclopedia of Integer Sequences. URL: http://oeis.org.