Cyclic Complexity of Words

Julien Cassaigne[1], Gabriele Fici[2,⋆],
Marinella Sciortino[2,⋆], and Luca Q. Zamboni[3,4]

[1] Institut de Mathématiques de Luminy, Marseille, France
cassaigne@iml.univ-mrs.fr
[2] Dipartimento di Matematica e Informatica, University of Palermo, Italy
{fici,mari}@math.unipa.it
[3] Université Claude Bernard Lyon 1, France
[4] Department of Mathematics, University of Turku, Finland
lupastis@gmail.com

Abstract. We introduce and study a new complexity function on words, which we call *cyclic complexity*, which counts the number of conjugacy classes of factors of each given length. We extend the famous Morse-Hedlund theorem to the setting of cyclic complexity by showing that a word is ultimately periodic if and only if it has bounded cyclic complexity. Unlike most complexity functions, cyclic complexity distinguishes between Sturmian words having different slopes. More precisely, we prove that if x is a Sturmian word and y is a word having the same cyclic complexity of x then y is Sturmian and, up to renaming letters, it has the same language of factors of x.

Keywords: Cyclic complexity, factor complexity, Sturmian words, minimal forbidden factor.

1 Introduction

The usual notion of complexity of a discrete system counts the number of distinct patterns of the same size that the system can generate. In the case of sequences (words), this is the number of distinct blocks (factors) of each given length. This measure of complexity is usually called *factor complexity* (or *block complexity*). The words with the "simplest" structure are the periodic ones. They are of the form $x = u^\omega$ (i.e., an infinite concatenation of a same finite block u) called *purely periodic*, or of the form $x = vu^\omega$, called *ultimately periodic*. The non-periodic words are called *aperiodic*. The factor complexity distinguishes between periodic and aperiodic words. In fact, a fundamental result dating back to the late 30's is the famous theorem of Morse and Hedlund [14] stating that a word is aperiodic if and only if it has at least $n+1$ factors of length n for every n. From this result, it is natural to consider those aperiodic words which have minimal factor complexity, i.e., those having exactly $n + 1$ distinct factors of length n for every n. These

⋆ Partially supported by Italian MIUR Project PRIN 2010LYA9RH, "Automi e Linguaggi Formali: Aspetti Matematici e Applicativi".

E. Csuhaj-Varjú et al. (Eds.): MFCS 2014, Part I, LNCS 8634, pp. 159–170, 2014.

are called *Sturmian words* and a vast bibliography exists showing their interest both from the theoretical viewpoint and in applications. For example, Sturmian words code the digital approximations in the plane of Euclidean straight lines with irrational slope, with the property that two Sturmian words have the same slope if and only if they have the same language of factors.

There exist many other measures of complexity of words in literature. For example, a lot of attention has recently been given (see for instance [3,10,16,18,19]) to the *abelian complexity*, which is the function counting the number of factors of each given length up to permutations. Other new measures of complexity of words have been introduced over the time, which are intermediate between factor and abelian complexity (e.g. maximal pattern complexity [7], k-abelian complexity [8], binomial complexity [17]) or involve different definitions that appear naturally in the study of sequences (e.g. periodicity complexity [12], minimal forbidden factor complexity [13], palindromic complexity [5], etc.) For most of these measures, Sturmian words are those aperiodic words of lowest complexity. However, they do not distinguish between two Sturmian words having different slopes.

In this paper we propose a new measure of complexity, *cyclic complexity*, which consists in counting the factors of each given length of a word up to conjugacy. The notion of conjugacy is a basic notion in Combinatorics on Words. Two words are said conjugate if they are equal when read on a circle[1]. That is, the cyclic complexity of a word is the function counting the number of conjugacy classes of factors of each given length.

One of the main results of this paper is that cyclic complexity distinguishes between periodic and aperiodic words. In fact, we prove the following theorem.

Theorem 1. *A word is ultimately periodic if and only if it has bounded cyclic complexity.*

That is, the Morse-Hedlund theorem can be extended to the setting of cyclic complexity. Note that a word is (purely) periodic if and only if there exists an integer n such that all the factors of length n are conjugate. Therefore, the minimum value that the cyclic complexity of an aperiodic word can take is 2. We will prove that Sturmian words have the property that the cyclic complexity takes value 2 infinitely often.

Since the Sturmian words are characterized by having $n+1$ factors of length n for every n, the factor complexity does not distinguish between two Sturmian words with different languages of factors. In contrast, for cyclic complexity, two Sturmian words with different languages of factors have different cyclic complexity. Indeed, we prove something stronger:

Theorem 2. *Let x be a Sturmian word. If a word y has the same cyclic complexity as x then, up to renaming letters, y is a Sturmian word having the same slope of x.*

[1] More formally, u and v are conjugate if and only if one can write $u = w_1w_2$ and $v = w_2w_1$ for some words w_1, w_2.

That is, not only two Sturmian words with different languages of factors cannot have the same cyclic complexity, but the only words which have the same cyclic complexity of a Sturmian word x are those Sturmian words with the same slope of x.

These two results suggest that cyclic complexity can be considered as an interesting refinement of the classical notion of factor complexity and can open new perspectives in the study of complexity of discrete systems.

Note that factor complexity, abelian complexity and cyclic complexity can all be viewed as actions of different subgroups of the symmetric group on the indices of a finite word (respectively, the trivial subgroup, the whole symmetric group and the cyclic subgroup). Since factor and abelian complexity are very well studied, looking at other subgroups of the symmetric group seems a very natural way of investigation.

2 Basics

Given a finite non-empty ordered set A (called the *alphabet*), we let A^* and $A^{\mathbb{N}}$ denote respectively the set of finite words and the set of (right) infinite words over the alphabet A. The order on the alphabet A can be extended to the usual lexicographic order on the set A^*.

For a finite word $w = w_1 w_2 \cdots w_n$ with $n \geq 1$ and $w_i \in A$, the length n of w is denoted by $|w|$. The *empty word* is denoted by ε and we set $|\varepsilon| = 0$. We let A^n denote the set of words of length n and A^+ the set of non-empty words. For $u, v \in A^+$, $|u|_v$ is the number of occurrences of v in u. For instance $|0110010|_{01} = 2$. The *Parikh vector* of w is the vector whose components are the number of occurrences of the letters of A in w. For example, if $A = \{a, b, c\}$, then the Parikh vector of $w = abb$ is $(1, 2, 0)$. The *reverse* (or *mirror image*) of a finite word w is the word obtained by reading w in the reverse order.

Given a finite or infinite word $\omega = \omega_1 \omega_2 \cdots$ with $\omega_i \in A$, we say a word $u \in A^+$ is a *factor* of ω if $u = \omega_i \omega_{i+1} \cdots \omega_{i+n-1}$ for some positive numbers i and n. We let $\text{Fact}(\omega)$ denote the set of all factors of ω, and $\text{Alph}(\omega)$ the set of all factors of ω of length 1. If $\omega = u\nu$, we say that u is a *prefix* of ω, while ν is a *suffix* of ω. A factor u of ω is called *right special* (resp. *left special*) if both ua and ub (resp. au and bu) are factors of ω for distinct letters $a, b \in A$. The factor u is called *bispecial* if it is both right special and left special.

For each factor u of ω, we set

$$\omega\big|_u = \{n \in \mathbb{N} \mid \omega_n \omega_{n+1} \cdots \omega_{n+|u|-1} = u\}.$$

We say ω is *recurrent* if for every $u \in \text{Fact}(\omega)$ the set $\omega\big|_u$ is infinite. We say ω is *uniformly recurrent* if for every $u \in \text{Fact}(\omega)$ the set $\omega\big|_u$ is syndedic, i.e., of bounded gap. A word $\omega \in A^{\mathbb{N}}$ is *(purely) periodic* if there exists a positive integer p such that $\omega_{i+p} = \omega_i$ for all indices i, while it is *ultimately periodic* if $\omega_{i+p} = \omega_i$ for all sufficiently large i. Finally, a word $\omega \in A^{\mathbb{N}}$ is called *aperiodic* if it is not ultimately periodic. For a finite word $w = w_1 w_2 \cdots w_n$, we call p a *period* of u if $w_{i+p} = w_i$ for every $1 \leq i \leq n - p$.

Two finite or infinite words are said to be *isomorphic* if the two words are equal up to a renaming of the letters.

A (finite or infinite) word w over A is *balanced* if and only if for any u, v factors of w of the same length and for every letter $a \in A$, one has $||u|_a - |v|_a| \leq 1$. More generally, w is C*-balanced* if there exists a constant $C > 0$ such that for any u, v factors of w of the same length and for every letter $a \in A$, one has $||u|_a - |v|_a| \leq C$. For example, the word 010111 is not balanced, but it is 2-balanced. Note that if w is C-balanced, then it is C'-balanced for any $C' \geq C$.

The *factor complexity* of an infinite word ω is the function

$$p_\omega(n) = |\,\mathrm{Fact}(\omega) \cap A^n|,$$

i.e., the function that counts the number of distinct factors of length n of ω, for every $n \geq 0$. The factor complexity is a standard measure of the complexity of an infinite word. By Morse-Hedlund theorem, words with bounded factor complexity are precisely ultimately periodic words and aperiodic words with minimal factor complexity have linear factor complexity. In the binary case, aperiodic words with minimal factor complexity have factor complexity equal to $n+1$, i.e., they are Sturmian words. An example of word achieving maximal factor complexity over an alphabet of size $k > 1$ can be obtained by concatenating the k-ary expansions of non-negative integers. For example, if $k = 2$, one obtains the so called Champernown word $0110111001011101111000\cdots$

The factor complexity counts the factors appearing in the word. A dual point of view consists in counting the shortest factors that *do not* appear in the word. This leads to another measure of complexity, described below.

Let w be a (finite or infinite) word over an alphabet A. A finite non-empty word v is a *minimal forbidden factor* for w if v does not belong to $\mathrm{Fact}(w)$ but every proper factor of v does. We denote by $\mathrm{MF}(w)$ the set of all minimal forbidden words for w. The *minimal forbidden factor complexity* of an infinite word ω is the function

$$mf_\omega(n) = |\mathrm{MF}(\omega) \cap A^n|\,,$$

i.e., the function that counts the number of distinct minimal forbidden factors of length n of ω, for every $n \geq 0$.

We now introduce a new measure of complexity. The idea is to count the factors of each given length that are different up to a rotation. Recall that two finite words u, v are *conjugate* if there exist words w_1, w_2 such that $u = w_1 w_2$ and $v = w_2 w_1$. The conjugacy relation is an equivalence over A^*, which is denoted by $\sim$, whose classes are called *conjugacy classes*. Note that two words belonging to the same conjugacy class must have the same Parikh vector.

The *cyclic complexity* of an infinite word ω is the function

$$c_\omega(n) = \left|\frac{\mathrm{Fact}(\omega) \cap A^n}{\sim}\right|,$$

i.e., the function that counts the number of distinct conjugacy classes of factors of length n of ω, for every $n \geq 0$.

Observe that, by the definition, $c_\omega(n) \leq p_\omega(n)$ for every n. Moreover, if a word ω has maximal cyclic complexity, then it has maximal factor complexity. In fact, let $w \in A^*$ be any word. We want to show that $w \in \mathrm{Fact}(\omega)$. Consider the word ww. From the maximality of the cyclic complexity of ω, some conjugate of ww is an element of $\mathrm{Fact}(\omega)$. But every conjugate of ww contains w as a factor, hence $w \in \mathrm{Fact}(\omega)$.

Since a word having maximal factor complexity clearly also has maximal cyclic complexity, we have the following proposition.

Proposition 1. *An infinite word has maximal cyclic complexity if and only if it has maximal factor complexity.*

The cyclic complexity, as well as the other mentioned complexity functions, can be naturally extended to any factorial language. Recall that a language is any subset of A^*. A language L is called *factorial* if it contains all the factors of its words, i.e., if $uv \in L \Rightarrow u, v \in L$. The cyclic complexity of L is defined by

$$c_L(n) = \left| \frac{L \cap A^n}{\sim} \right|.$$

The cyclic complexity is an invariant for several operations on languages. For example, it is clear that if two languages are isomorphic (i.e., one can be obtained from the other by renaming letters), then they have the same cyclic complexity. Furthermore, if L is a language and $\tilde{L}$ is obtained from L by reversing (mirror image) each word in L, then L and $\tilde{L}$ have the same cyclic complexity.

3 Cyclic Complexity Distinguishes between Periodic and Aperiodic Words

In this section we give a proof of Theorem 1. The following lemma connects cyclic complexity to balancedness.

Lemma 1. *Let $\omega \in A^{\mathbb{N}}$ and suppose that there exists a constant C such that $c_\omega(n) \leq C$ for every n. Then ω is C-balanced.*

Proof. For every n, there are at most C conjugacy classes of factors of length n in ω. This implies that there are at most C different Parikh vectors for the factors of ω of length n, that is, ω has abelian complexity bounded by C. It can be proved (see [16]) that this implies that the word ω is C-balanced. □

Lemma 2. *Let $\omega \in A^{\mathbb{N}}$ be aperiodic and let $v \in A^+$ be a factor of ω which occurs in ω an infinite number of times. Then, for each positive integer K there exists a positive integer n such that ω contains at least $K + 1$ distinct factors of length n beginning in v.*

Proof. Suppose to the contrary that for some K, ω has at most K distinct factors of each length n which begin in v. Since ω is aperiodic and v occurs infinitely often in ω, there exist $K+1$ distinct suffixes of ω (say $y_0, y_1, \ldots, y_K$) beginning in v. By the pigeonhole principle, for each positive integer n there exist $0 \leq i < j \leq K$ such that y_i and y_j begin in the same prefix of length n. Again by the pigeonhole principle, there exist $0 \leq i < j \leq K$ such that y_i and y_j begin in the same prefix of length n for infinitely many distinct values of n. Hence, $y_i = y_j$, a contradiction. □

Proof of Theorem 1. If ω is ultimately periodic, then it has bounded factor complexity by Morse-Hedlund theorem, hence it must have bounded cyclic complexity.

Let us now prove that if ω is aperiodic, then for any fixed positive integer M, $c_\omega(n) \geq M$ for some n. Short of replacing ω by a suffix of ω, we can suppose that each letter occurring in ω occurs infinitely often in ω. First, suppose that for each positive integer C, ω is not C-balanced. Then, by Lemma 1, the cyclic complexity of ω is unbounded and we are done. Next, suppose that each $u \in A^+$ is a factor of ω. In this case, ω would have full complexity, whence the cyclic complexity of ω is again unbounded. Thus, we can suppose that ω is C-balanced for some positive integer C, and that some $u \in A^+$ is not a factor of ω. Since ω is C-balanced, there exists a positive integer N such that each factor of ω of length N contains an occurrence of each $a \in \mathrm{Alph}(\omega)$. As u is a forbidden factor of ω, it follows that u is a forbidden factor of each suffix of ω. Since each letter occurring in ω occurs infinitely often, it follows there exist a suffix ω' of ω, a letter $a \in A$ and a word $v \in A^+$ such that av is a forbidden factor of ω' and v occurs in ω' infinitely often. By Lemma 2, there exists a positive integer $n_0 \geq 2|v|$ such ω' contains at least MN distinct factors of length n_0 beginning in v. We denote these factors by $u_1, u_2, \ldots, u_{MN}$. There exist $v_1, v_2, \ldots, v_{MN}$, each in A^N, such that $u_i v_i$ are factors of ω' (of length $n_0 + N$) for each $1 \leq i \leq MN$. Since each v_i contains an occurrence of a, it follows there exists $n \geq n_0$ such that ω' contains at least M distinct factors of length n beginning in v and terminating in a. Since av is a forbidden factor of ω', no two of these factors are conjugate to one another. Hence, $c_{\omega'}(n) \geq M$ and thus $c_\omega(n) \geq M$. □

4 Cyclic Complexity Distinguishes between Sturmian Words with Different Languages

In this section we exhibit results on the cyclic complexity of Sturmian words and give a sketch of the proof of Theorem 2.

There exists a vast bibliography on Sturmian words (see for instance the survey papers [1,2], [9, Chap. 2], [15, Chap. 6] and references therein).

A Sturmian word is an infinite word having exactly $n+1$ distinct factors of length n, for every $n \geq 0$. That is, a word x is Sturmian if and only if $p_x(n) = n+1$ for every $n \geq 0$. Note that an immediate consequence of the definition is that $|\mathrm{Alph}(x)| = 2$, so a Sturmian word is a binary word. In this section we fix the alphabet $A = \{0, 1\}$.

A very well known instance of Sturmian words is the Fibonacci word $F = 010010100100101001 \cdots$, obtained as the limit of the substitution $0 \mapsto 01$, $1 \mapsto 0$.

Sturmian words have a multitude of combinatorial properties that make them fundamental objects in the field of Combinatorics on Words. By Morse-Hedlund Theorem, Sturmian words are those aperiodic words with minimal factor complexity. We recall some other characterizations in the next proposition.

Proposition 2. *Let $x \in A^{\mathbb{N}}$. The following conditions are equivalent:*

1. *x is Sturmian;*
2. *x is balanced and aperiodic;*
3. *x has exactly one right (resp. left) special factor for each length.*

Recall that the *slope* of a finite binary word w over the alphabet A is defined as $s(w) = \frac{|w|_1}{|w|}$. The slope of an infinite binary word, when it exists, is the limit of the slopes of its prefixes. A Sturmian word can also be defined by considering the intersections with a squared-lattice of a semi-line having a slope which is an irrational number. A horizontal intersection is denoted by the letter 0, while a vertical intersection is denoted by the letter 1. Note that the slope of the Sturmian word is exactly the slope of such a semi-line. For example, the slope of the Fibonacci word is $(1+\phi)^{-1}$, where $\phi = (1+\sqrt{5})/2$ is the golden ratio.

An important property of Sturmian words is that their factors depend on their slope only, i.e., we have the following result (see [14]).

Proposition 3. *Let x, y be two Sturmian words. Then* $\mathrm{Fact}(x) = \mathrm{Fact}(y)$ *if and only if x and y have the same slope.*

A fundamental role in the study of factors of Sturmian words is played by the *central words*. A word is central if it has coprime periods p and q and length $p+q-2$. There are several characterizations of central words (see [1] for a survey). Here we recall the following ones.

Proposition 4. *Let w be a word over A. The following conditions are equivalent:*

1. *w is a central word;*
2. $0w1$ *and* $1w0$ *are conjugate;*
3. *w is a bispecial factor of some Sturmian word;*
4. *w is a palindrome and the words $w0$ and $w1$ are balanced;*
5. $0w1$ *is balanced and is the least element (w.r.t. the lexicographic order) in its conjugacy class;*
6. *w is a power of a letter or there exist central words p_1, p_2 such that $w = p_1 01 p_2 = p_2 10 p_1$. Moreover, in this latter case $|p_1|+2$ and $|p_2|+2$ are coprime periods of w and* $\min(|p_1|+2, |p_2|+2)$ *is the minimal period of w.*

Let w be a central word with coprime periods p and q and length $p+q-2$. The words $0w1$ and $1w0$, which, by Proposition 4, are conjugate, are called *Christoffel words*. Let $r = |0w1|_0$ and $s = |0w1|_1$. It can be proved that r and s are the multiplicative inverses of p and q modulo $p+q$, respectively. Moreover, the conjugacy class of $0w1$ and $1w0$ contains exactly $|w|+2$ words. If we sort these words lexicographically and arrange them as rows of a matrix, we obtain a square matrix with remarkable combinatorial properties (see [4,6,11]). We call this matrix the (r,s)-*Christoffel array* and denote it by $\mathcal{A}_{r,s}$ (see Figure 1 for an example). Two consecutive rows of $\mathcal{A}_{r,s}$ differ only by a swap of two consecutive positions. Moreover, the columns are also conjugate and in particular the first one is 0^r1^s, while the last one is 1^s0^r.

Every aperiodic word (and therefore, in particular, every Sturmian word) contains infinitely many bispecial factors. If w is a bispecial factor of a Sturmian word x, then w is central by Proposition 4 and there exists a unique letter $a \in A$ such that w', the shortest palindrome beginning with wa, is a bispecial factor of x. Moreover, if p and q are the coprime periods of w such that $|w| = p+q-2$, then the word w' is central and its coprime periods p' and q' verifying $|w'| = p'+q'-2$ satisfy either $p' = p+q$ and $q' = p$, or $p' = p+q$ and $q' = q$, depending on the letter a. For example, 010 is a bispecial factor of the Fibonacci word F and has coprime periods 3 and 2 (and length $3+2-2$). The successive (in length order) bispecial factor of F is 010010, which is the shortest palindrome beginning in $010 \cdot 0$ and has coprime periods 5 and 3 (and length $5+3-2$). There exist other Sturmian words having 010 as a bispecial factor and for which the successive bispecial factor is 01010 (i.e., the shortest palindrome beginning with $010 \cdot 1$) whose coprime periods are 5 and 2.

These combinatorial properties of central words and the bispecial factors of a Sturmian word will be needed in our proof of Theorem 2.

Sturmian words have unbounded cyclic complexity (by Theorem 1) but their cyclic complexity takes value 2 for infinitely many n. More precisely, we have the following result.

Lemma 3. *Let x be a Sturmian word. Then $c_x(n) = 2$ if and only if $n = 1$ or there exists a bispecial factor of x of length $n-2$.*

The value 2 is the minimal possible for an aperiodic word. In fact, it is well known that a word ω is (purely) periodic if and only if there exists $n \geq 1$ such that all the factors of length n of ω are conjugate.

Since a Sturmian word contains infinitely many bispecial factors, the previous result implies that for a Sturmian word x one has that $\liminf c_x(n) = 2$. However, this is not a characterization of Sturmian words. In fact, there exist non-Sturmian aperiodic words with minimal cyclic complexity (in the sense of having limit inferior of the cyclic complexity equal to 2). Consider for example the morphism $\mu : 0 \mapsto 00, 1 \mapsto 01$. It is possible to prove that in the word $\mu(F) = 0001000001000100001 \cdots$, image of the Fibonacci word F under μ, there are exactly 2 conjugacy classes of factors of length n for every n that is the double of a Fibonacci number[2], so that $\liminf c_{\mu(F)}(n) = 2$. However, the word $\mu(F)$ is not Sturmian (it contains the factors 00000 and 10001 and therefore is not balanced). We show in Table 1 the first values for the cyclic complexity of F and $\mu(F)$.

[2] Recall that Fibonacci numbers are defined by: $F_0 = 1$, $F_1 = 1$, and $F_n = F_{n-1}+F_{n-2}$ for every $n > 1$.

Table 1. The initial values of the cyclic complexity for the Fibonacci word F and its morphic image $\mu(F)$

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
$c_F(n)$	2	2	2	3	2	4	4	2	7	4	5	8	2	9	9	4	13	5	9	14	2	16
$c_{\mu(F)}(n)$	2	2	2	2	3	2	3	3	3	2	5	4	5	4	6	2	7	7	7	4	9	5

We now give a sketch of the proof of Theorem 2.

Proof of Theorem 2 (Sketch). Since y has the same cyclic complexity of x, we have that in particular $2 = c_x(1) = c_y(1)$, so y is a binary word. Since x is aperiodic, by Theorem 1 c_x is unbounded. Since x and y have the same cyclic complexity we have, always by Theorem 1, that y is aperiodic.

We want to prove that y and x have the same factors. By contradiction, let $n+2$ be the least length for which x and y have different factors. This implies that x and y have a same bispecial factor w of length n. Let p' and q', with $p' > q'$, be the two coprime periods of w such that $n = |w| = p' + q' - 2$. Let w_x (resp. w_y) be the successive (in length order) bispecial factor of x (resp. of y). It can be proved that $\{|w_x|, |w_y|\} = \{2p' + q' - 2, p' + 2q' - 2\}$ and that w_x and w_y cannot have the same length.

Suppose $|w_x| < |w_y|$. Then, by Lemma 3, y would have cyclic complexity equal to 2 at length $|w_x| + 2$, which is impossible since between $|w|$ and $|w_y|$ the word y behaves as a Sturmian word and so by Lemma 3 it should have a bispecial factor of length $|w_x| + 2$. Hence, we can suppose that $|w_x| > |w_y|$, so that w_x has periods $p' + q'$ and p' and length $2p' + q' - 2$, while w_y has periods $p' + q'$ and q' and length $p' + 2q' - 2$.

To ease notation, we set $p = p' + q'$ and $q = p'$, so that $|w_y| = 2p - q - 2$ and $|w_x| = p + q - 2$. Let us consider the set of factors of y of length $2p - q$. Since $|w| + 2 < 2p - q < |w_x| + 2$, we know by Lemma 3 that $c_x(2p - q) > 2$. So, $c_y(2p - q) > 2$.

If there was a Sturmian word y' such that $\mathrm{Fact}(y') \cap A^{2p-q} = \mathrm{Fact}(y) \cap A^{2p-q}$, then $2p - q$ would be the length of a bispecial factor plus 2 of a Sturmian word and then, by Lemma 3, we would have $c_y(2p - q) = 2$, a contradiction. This implies that w_y is a bispecial factor of y that behaves differently from a bispecial factor of a Sturmian word. More precisely, we must have that $0w_y$ and $1w_y$ are both right special factors of y. Therefore, $0w_y0$ and $1w_y1$ are in two different conjugacy classes and all the other factors of y of length $2p - q$ are in a third conjugacy class. In other words, we have $c_y(2p - q) = 3$. Thus, in order to have a contradiction we are left to prove that $c_x(2p - q) \geq 4$.

It is known that among the $p + q + 1$ factors of x of length $p + q$, there is one factor with a Parikh vector and the remaining $p + q$ factors with the other Parikh vector, these latter being in a same conjugacy class, which is in fact the conjugacy class of the Christoffel word $0w_x1$.

Let $r = |0w_x1|_0$ and $s = |0w_x1|_1$. Without loss of generality, we can suppose that $r > s$, i.e., we can suppose that 11 does not appear as a factor in x. Therefore, we can build the (r,s)-Christoffel array $\mathcal{A}_{r,s}$. The factors of length $2p - q$ of x can be obtained by removing the last $2q - p$ columns from $\mathcal{A}_{r,s}$ (of course, in this way some rows can be equal and therefore some factors appear more than once). We refer to the matrix so obtained as to $\mathcal{A}'_{r,s}$.

The cases $s = 1, 2, 3$ can be proved separately. Here we give the sketch of the proof when $s > 3$. Recall that $\{r, s\} = \{p^{-1}, q^{-1}\} \mod (p + q)$. Suppose that $s = p^{-1} < q^{-1}$. In this case, one can prove that the last three rows in $\mathcal{A}'_{r,s}$ are distinct and start and end with 1. Therefore, each of these rows is unique in its conjugacy class. Since any other row correspond to a factor with a different Parikh vector, this implies that there are at least 4 conjugacy classes and we are done.

The other case is when $s = q^{-1} < p^{-1}$. This case can be proved analogously by considering the first four rows of the matrix $\mathcal{A}'_{r,s}$. In fact, one can prove that the factors appearing in the first four rows of the matrix $\mathcal{A}'_{r,s}$ are pairwise distinct and neither is conjugate to another. □

Example 1. Consider the Fibonacci word F and its bispecial factor $w = 010010$, which has periods $p = 5$ and $q = 3$. We have $s = q^{-1} = 3 < 5 = r = p^{-1}$. In Figure 1 we show the $(5,3)$-Christoffel array $\mathcal{A}_{5,3}$. The rows are the lexicographically sorted factors of F with Parikh vector $(5,3)$. The other factor of length 8 of F is 10100101. The factors of F of length $2p - q = 7$ can be obtained by removing the last column of the matrix. Notice that the first 4 rows (once the last character has been removed) are pairwise distinct and neither is conjugate to another.

To end this section, we compare the cyclic complexity to the minimal forbidden factor complexity for the special case of Sturmian words.

In [13] the authors proved the following result.

Theorem 3. *Let x be a Sturmian word and let y be an infinite word such that for every n one has $p_x(n) = p_y(n)$ and $mf_x(n) = mf_y(n)$, i.e., y is a word having the same factor complexity and the same minimal forbidden factor complexity as x. Then, up to isomorphism, y is a Sturmian word having the same slope as x.*

$$\mathcal{A}_{5,3} = \begin{pmatrix} 0\,0\,1\,0\,0\,1\,0\,1 \\ 0\,0\,1\,0\,1\,0\,0\,1 \\ 0\,1\,0\,0\,1\,0\,0\,1 \\ 0\,1\,0\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,0\,1\,0 \\ 1\,0\,0\,1\,0\,0\,1\,0 \\ 1\,0\,0\,1\,0\,1\,0\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0 \end{pmatrix}$$

Fig. 1. The matrix $\mathcal{A}_{5,3}$ for the Fibonacci word F for $p = 5$ and $q = 3$

Note that Theorem 2 is much stronger than Theorem 3, because in this latter the fact that y is a Sturmian word follows directly from the hypothesis that y has the same factor complexity as x.

Indeed, the cyclic complexity is more fine than the minimal factor complexity. Let x be an infinite binary word such that $\mathrm{MF}(x) = \{11, 000\}$ and y an infinite binary word such that $\mathrm{MF}(y) = \{11, 101\}$. Then x and y have the same minimal forbidden factor complexity, but it is readily checked that $c_x(5) = 3$ while $c_y(5) = 4$. Note that x contains 7 factors of length 5 corresponding to 3 cyclic classes $(00100, 00101, 01001, 10010, 10100, 10101)$ while y contains the factors $00000, 10000, 10010, 10001$ no two of which are cyclically conjugate.

5 Conclusions and Further Developments

We introduced a new measure of complexity of words, cyclic complexity. We showed that for this measure of complexity the Morse-Hedlund theorem can be extended, that is, a word is ultimately periodic if and only if it has bounded cyclic complexity (Theorem 1). The aperiodic words with minimal cyclic complexity can be defined as those having exactly 2 conjugacy classes of factors of length n for infinitely many values of n. Among these we have Sturmian words (which are the aperiodic words with minimal factor complexity), but we also exhibited a non-Sturmian example which, however, is a morphic image of a Sturmian word. We leave as an open problem that of characterizing the aperiodic words with minimal cyclic complexity.

Contrarily to other measures of complexity, cyclic complexity characterizes the language of a Sturmian word, in the sense that two Sturmian words with different languages of factors have different cyclic complexities. More precisely, we proved that a word having the same cyclic complexity as a Sturmian word must be Sturmian and have the same slope (Theorem 2). A natural question is therefore the following: Given two infinite words x and y with the same cyclic complexity, what can we say about their languages of factors?

First, there exist two periodic words having same cyclic complexity but whose languages of factors are not isomorphic nor related by mirror image. For example, let τ be the morphism: $0 \mapsto 010$, $1 \mapsto 011$ and consider the words $x = \tau((010011)^\omega)$ and $x' = \tau((101100)^\omega)$. One can verify that x and x' have the same cyclic complexity up to length 17 and, since each has period 18, the cyclic complexities agree.

Furthermore, it is easy to show that even two aperiodic words can have same cyclic complexity but different languages of factors. For example, let x be an infinite binary word such that $\mathrm{MF}(x) = \{000111\}$ and y an infinite binary word such that $\mathrm{MF}(y) = \{001111\}$. Then the languages of factors of x and y are not isomorphic, nor related by mirror image, yet the two words have the same cyclic complexity. However, we do not know if this can still happen with the additional hypothesis of linear complexity, for example. In every case, these examples show that cyclic complexity does not determine, in general, the language of factors. So, our Theorem 2 is very special to Sturmian words.

In conclusion, we believe that the new notion of complexity we introduced in this paper, cyclic complexity, can open new perspectives in the study of complexity of words and languages.

References

1. Berstel, J.: Sturmian and episturmian words (a survey of some recent results). In: Bozapalidis, S., Rahonis, G. (eds.) CAI 2007. LNCS, vol. 4728, pp. 23–47. Springer, Heidelberg (2007)
2. Berstel, J., Lauve, A., Reutenauer, C., Saliola, F.: Combinatorics on Words: Christoffel Words and Repetitions in Words. CRM monograph series, vol. 27. American Mathematical Society (2008)
3. Blanchet-Sadri, F., Fox, N.: On the asymptotic abelian complexity of morphic words. In: Béal, M.-P., Carton, O. (eds.) DLT 2013. LNCS, vol. 7907, pp. 94–105. Springer, Heidelberg (2013)
4. Borel, J.-P., Reutenauer, C.: On Christoffel classes. RAIRO Theor. Inform. Appl. 40(1), 15–27 (2006)
5. Brlek, S., Hamel, S., Nivat, M., Reutenauer, C.: On the palindromic complexity of infinite words. Internat. J. Found. Comput. Sci. 15(2), 293–306 (2004)
6. Jenkinson, O., Zamboni, L.Q.: Characterisations of balanced words via orderings. Theoret. Comput. Sci. 310(1-3), 247–271 (2004)
7. Kamae, T., Zamboni, L.: Maximal pattern complexity for discrete systems. Ergodic Theory Dynam. Systems 22(4), 1201–1214 (2002)
8. Karhumäki, J., Saarela, A., Zamboni, L.Q.: On a generalization of abelian equivalence and complexity of infinite words. J. Comb. Theory, Ser. A 120(8), 2189–2206 (2013)
9. Lothaire, M.: Algebraic Combinatorics on Words. Cambridge University Press, Cambridge (2002)
10. Madill, B., Rampersad, N.: The abelian complexity of the paperfolding word. Discrete Math. 313(7), 831–838 (2013)
11. Mantaci, S., Restivo, A., Sciortino, M.: Burrows-Wheeler transform and Sturmian words. Inform. Process. Lett. 86(5), 241–246 (2003)
12. Mignosi, F., Restivo, A.: A new complexity function for words based on periodicity. Internat. J. Algebra Comput. 23(4), 963–988 (2013)
13. Mignosi, F., Restivo, A., Sciortino, M.: Words and forbidden factors. Theoret. Comput. Sci. 273(1-2), 99–117 (2002)
14. Morse, M., Hedlund, G.A.: Symbolic dynamics. Amer. J. Math. 60, 1–42 (1938)
15. Fogg, N.P.: Substitutions in Dynamics, Arithmetics and Combinatorics. Lecture Notes in Math., vol. 1794. Springer (2002)
16. Richomme, G., Saari, K., Zamboni, L.: Abelian complexity of minimal subshifts. J. Lond. Math. Soc. 83(1), 79–95 (2011)
17. Rigo, M., Salimov, P.: Another generalization of abelian equivalence: Binomial complexity of infinite words. In: Karhumäki, J., Lepistö, A., Zamboni, L. (eds.) WORDS 2013. LNCS, vol. 8079, pp. 217–228. Springer, Heidelberg (2013)
18. Saarela, A.: Ultimately constant abelian complexity of infinite words. J. Autom. Lang. Comb. 14(3-4), 255–258 (2010)
19. Turek, O.: Abelian complexity and abelian co-decomposition. Theoret. Comput. Sci. 469, 77–91 (2013)