Theoretical Computer Science 591 (2015) 106–133

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Studies on finite Sturmian words

Christophe Reutenauer

Laboratoire de combinatoire et d'informatique mathématique, Université du Québec à Montréal, Case postale 8888, succursale Centre-ville, Montréal, Québec H3C 3P8, Canada

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form 1 May 2015
Accepted 3 May 2015
Available online 7 May 2015
Communicated by D. Perrin

Keywords:
Finite Sturmian words
Christoffel words
Sturmian sequences
Palindromes
Special words
Rauzy graph
Periods
Conjugation
Normal form
Circular factors
Enumeration
Markoff numbers

ABSTRACT

Several properties of finite Sturmian words are proved. The inverse of the Richomme–Séébold bijection between factor sets of given length of Sturmian sequences and left special Sturmian words is constructed (using circular factors of Christoffel words) and studied. The Labbé bijection between Christoffel words and left special Sturmian words is introduced and studied. Factor sets are classified by two types: the two classes are characterized by conjugation and periodicity properties. The two classes are related to the Rauzy graph. A normal form for finite Sturmian words is given, following a result of de Luca and De Luca. Several classes of Sturmian words are characterized through this normal form: left or right special words, bispecial words, palindromes, central words. The normal form is used to derive a proof of the Lipatov–Mignosi formula for the number of Sturmian words of given length; to produce a linear algorithm which checks if a word is Sturmian and which computes its normal form; to construct algorithmically the contraction and completion in various classes of Sturmian words; to compute the de Luca palindromic closure of any Sturmian word and prove that it is Sturmian. The completion is related to the de Luca palindromization function and it is shown that the sets of directive words of central words of even and odd length are rational group languages. Christoffel words are characterized in several ways: an extension of a result of Chuan; an extension of a result of Droubay–Pirillo, by counting circular factors; by counting Lyndon factors. A special Christoffel word is constructed having length a given Markoff number. A bivariate count of special Sturmian words is given, using ideas of Bédaride, Domenjoud, Jamet and Rémy. Open problems are given.

1. Introduction

Sturmians sequences appear in several areas of pure and applied mathematics. They appear, without the name, in Markoff's 1880 work on minima of quadratic forms [27,45,46,57], in the 1940 work of Morse and Hedlund in symbolic dynamics, who gave them the name *Sturmian* [52], in arithmetic under their variant called *Beatty sequences* [19] or *spectra* [17,33]. They are the nonperiodic sequences of minimal complexity [38]: for a fixed sequence, this is the function $\mathbb{N} \to \mathbb{N}$ which with n associates the number of factors of the sequence of length n; for a Sturmian sequence, this number is $n+1$.

A factor of a Sturmian sequence is called a *finite Sturmian word* or simply a *Sturmian word* (in the present article, a word is always a finite word, and an infinite word will be called a sequence). Sturmian words appear even before Sturmian sequences, since they appear in the 1875 article of Christoffel [21] (probably one of the last mathematical article written in Latin). In Theoretical Computer Science, Sturmian words appear, under their paradigmatic example, the Fibonacci word,

E-mail address: Reutenauer.Christophe@uqam.ca.

http://dx.doi.org/10.1016/j.tcs.2015.05.003

in the precursory work of Berstel [5] and Aldo de Luca [39]; in particular, the fundamental notions of special factors and palindromes. The terminology “Christoffel words” was introduced by Jean Berstel [6], where is shown among other properties that all the discretizations of a segment, that exist in the literature, are conjugates words. See [2,38,55] for many references and results on Sturmian sequences and words.

The first part of the present article is motivated by the work [10]: a byproduct of the latter article is that the set of factors of given length n in a fixed Sturmian sequence (call such a set a *factor set*) is a basis of the subgroup of the free group F_2 which is the kernel of the homomorphism sending an element onto its algebraic length (that is, letters are sent onto 1 and their inverses onto -1) taken modulo n. Thus, the fact that a factor set is of cardinality $n+1$, known since the work of Morse and Hedlund, appears as a very special case of the Schreier formula (which gives the rank of a subgroup of finite index of a free group). It was therefore tempting to give a more precise description of the factor sets. The number of factor sets of a given length n has been recently given by Richomme and Séébold [58]: it is $\varphi(1)+\cdots+\varphi(n)$, where φ is the Euler function. They show that there is a bijection between factor sets of length n and left special words of length $n-1$; the bijection is defined by $F \mapsto l$, with $al, bl \in F$. We give here the inverse bijection, using the circular factors of a Christoffel word constructed from the left special word by the palindromic closure of de Luca (see Proposition 5.1). Several consequences of these bijections are derived, in particular the Labbé mapping (see Section 6).

In order to describe factor sets, we classify them into two classes. A factor set is of type **C** if it is the union of a conjugation class and another word; factor sets of this type appear already in the work of Zhi Xiong Wen and Zhi Ying Wen [62]. The other factor sets are called of type **P**. The interesting fact is that one may characterize these classes; in particular, type **P** is characterized by a property of periodicity (Proposition 7.3 and Corollary 7.5); the classes of each type may be counted, refining the Richomme–Séébold count, Corollary 7.4. Another result shows that this classification appears in the factor (Rauzy) graph, see Section 10.

In the middle part of the article, we give a normal form for Sturmian words, following a result of de Luca and De Luca [41]: they show that the minimal period of a Sturmian word is the length of some Christoffel factor of this word, a result in the spirit of a conjecture of Duval [30]. This normal form (Proposition 11.1) is very natural and rests on the palindromic factorization of the corresponding Christoffel word. It has many applications. The first one is a proof of the Lipatov–Mignosi formula giving the number of Sturmian words of a given length (Corollary 12.2); we use for its proof an amusing lemma on the number of matrices in $SL_2(\mathbb{N})$ of given total sum (Lemma 12.3). As another application of the normal form, we give characterizations of several classes of Sturmian words: palindromes, special, bispecial, central (Proposition 11.2). Moreover, we use this to characterize the completion and contraction in several classes of Sturmian words: for example, compute the longest palindrome which is the median factor of a given Sturmian palindrome w and the shortest Sturmian palindrome of which w is a median factor, reproving a result of de Luca and De Luca [42], and similar results for Sturmian and special Sturmian words (Section 14). A consequence of these constructions is the result, which seems new, that the palindromic closure of a Sturmian word is Sturmian and has the same minimal period (Corollary 14.7). Another application is a new algorithm to check if a given word is Sturmian; there exist several algorithms in the literature (see for instance [41] and the references therein), but the present one seems conceptually particularly simple: it rests on Duval's algorithm [31] which gives in linear time the factorization of a word into Lyndon words, and it is therefore linear (Section 13).

In the sections which follow, I study several properties of Christoffel words and their conjugation classes (Christoffel classes). Using the notion of *moment* of a word, Wai-Fong Chuan characterizes Christoffel classes [24,25], see also Theorem 16.1; we give a variant of her result, by using the area under the geometric realization of a word, see Proposition 16.2. In the next section, we extend to finite words the characterization by Xavier Droubay and Giuseppe Pirillo of Sturmian sequences, by their palindrome count [28], see also Theorem 17.1; the proof of this finitary version of their result is somewhat similar to their's, see Proposition 17.2. In the next section, I give several results on Christoffel factors of Christoffel words: their position, their number and the number of their occurrences, see Proposition 18.1, Corollary 18.4 and Corollary 18.5. After that, I give a characterization of Markoff numbers, Proposition 19.1, which states that such a number is the length of a special Christoffel word, whose directive word is antipalindromic; it allows to state a problem equivalent to the Markoff injectivity conjecture.

In Section 20, I want to compute the bivariate symmetric function associated with several classes of Sturmian words: all Sturmian words, palindromes, special and bispecial words. This is motivated by the fact that a bivariate count of Sturmian words and of Sturmian palindromes has been performed by the authors of [3]; they give among other results a recursive formula to count the number of Sturmian words, and of Sturmian palindromes, with a given number of a's and b's. Their method may be used for the set of special words, in particular using their function θ, already used in Section 15. In Section 21, tables of numbers are given and in Section 22, several open problems are given. In particular, find a closed formula for the bivariate symmetric function for the classes of Sturmian words; I could do this only for central words, Christoffel words, and bispecial words (following a formula of Fici [32]); the case of all Sturmian words (that is, find a bivariate symmetric function analogue of the Lipatov–Mignosi formula), special words and palindromes is open.

2. Periods and periodic patterns

Recall that a word $w = a_1 \cdots a_n$ has the *period* p if whenever $i, i+p \in \{1, \ldots, n\}$, one has $a_i = a_{i+p}$. The period is *nontrivial* if $p < n$ and we say that w is *periodic* if it has a nontrivial period p.

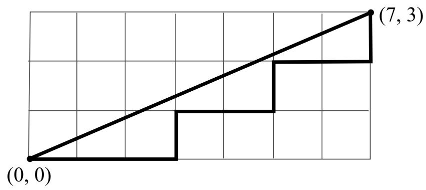

Fig. 1. The lower Christoffel word *aaabaabaab* of slope $\frac{3}{7}$ on the alphabet $\{a < b\}$.

A word w has a *border* if there is a nonempty prefix which is also a suffix; if w has no border, it is called *unbordered*. A word has a border if and only if is periodic.

If $w = a_1 \cdots a_n$ has the (nontrivial) period $p \le n$ ($p < n$), then each factor of the form $a_{i+1} \cdots a_{i+p}$ is called a *(nontrivial) periodic pattern of length* p. Observe that v is a periodic pattern of w if and only if w is a factor of v^∞ and if $|v| \le |w|$, and is moreover nontrivial if and only if $|v| < |w|$. Note also that all periodic patterns of the same length of w are conjugate words.[1]

The next well-known lemma will be useful.

Lemma 2.1. *Let v be a primitive word[2] and $n \ge |v|$; let N be such that the length of $|v^N|$ is at least equal to $n + |v| - 1$; then each factor of length $n + |v| - 1$ of v^N contains as factor the $|v|$ words of length n which have a periodic pattern conjugate to v. In particular, if $n = |v|$, each factor of length $2|v| - 1$ of v^N contains all conjugates of v.*

For example, let $v = aaab$, $n = 5$, $n + |v| - 1 = 8$; the factor *aabaaaba* of length 8 of $(aaab)^\infty$ contains as factors all the 4 words of length 5 which have a periodic pattern equal to *aaab*, *aaba*, *abaa* or *baaa*; these 4 words are *aabaa*, *abaaa*, *baaab*, *aaaba*.

3. Sturmian, Christoffel and special words

Consider *lattice paths*, which are consecutive elementary steps in the plane; each *elementary step* is a segment $[(x, y), (x + 1, y)]$ or $[(x, y), (x, y + 1)]$, with $x, y \in \mathbf{Z}$.

Let p, q be relatively prime integers. Consider the segment from $(0, 0)$ to (p, q) and the lattice path from $(0, 0)$ to (p, q) located below this segment and such that the polygon delimited by the segment and the path has no interior integer point.

Given a totally ordered alphabet $\{a < b\}$, the *lower Christoffel word of slope* q/p is the word in the free monoid $\{a, b\}^*$ coding the above path, where a (resp. b) codes a horizontal (resp. vertical) elementary step. See Fig. 1, where is represented the path with $(p, q) = (7, 3)$ corresponding to the Christoffel word of slope 3/7.

In view of this definition, call *slope* of a word on a totally ordered alphabet $\{a < b\}$ the ratio of its b-degree by its a-degree (that is, the slope of the segment in the plane delimited by the extreme points of the corresponding discrete path). Thus, the slope of a Christoffel word is well-defined.

The *upper Christoffel word of slope* q/p is defined similarly, by using the path located above the segment. Moreover, a *Christoffel word* is a lower or an upper Christoffel word, the order of the alphabet being here immaterial. A Christoffel word is *proper* if its length is at least 2. A Christoffel word is necessarily primitive, since the number of a's (resp. b's) in w is p (resp. q) and since p, q are relatively prime.

The lower Christoffel word w of slope q/p on the alphabet $\{a < b\}$ may be also defined as follows: let $w = a_0 \cdots a_{n-1}$, $n = p + q$, then for $i = 0, \ldots, n - 1$, $a_i = a$ if and only if $iq \mod n < (i + 1)q \mod n$, where $x \mod n$ denotes the remainder of the Euclidean division of x by n.

On a given binary alphabet, there are two lower Christoffel words of length 1, and for $n \ge 2$, the number of lower Christoffel word is $\varphi(n)$, see [38] Corollary 2.2.16. As usual, $\varphi(n)$ is the number of natural numbers between 1 and n which are relatively prime to n.

The *reversal* $\tilde{w}$ of a Christoffel word is also a Christoffel word, and is conjugate to w. More precisely, the reversal of a lower Christoffel word is the upper Christoffel word of the same slope.

Each proper Christoffel word has a *standard factorization* $w = uv$: u, v are the unique Christoffel words such that such a factorization exists. Each proper Christoffel w word has a unique factorization as a product of two palindromes, by [22] Theorem 3.1; we call it its *palindromic factorization*. If $w = uv$ (standard factorization), then the palindromic factorization is $w = v'u'$ with $|u'| = |u|$ and $|v'| = |v|$ and $\tilde{w} = u'v'$, see [16] Proposition 6.1.[3]

Let $n \ge 1$; if $w = a^n b$ (resp. ab^n), then the standard factorization is $a.a^{n-1}b$ (resp. $ab^{n-1}.b$) and the palindromic factorization is $a^n.b$ (resp. $a.b^n$). In all other cases, w has the standard factorization $w = uv$, with u, v proper Christoffel

1 Two words are *conjugate* if they may be written xy and yx for some words x, y.

2 A word is *primitive* if it is not of the form w^p, $p \ge 2$.

3 In the statement of the latter result, there is a small mistake: one has to replace the two first equalities by $w = xnxymy$, $w' = ymyxnx$; the proof is however correct.

words, so that $u = amb$, $v = apb$, m, p palindromes, and the palindromic factorization is $w = (apa)(bmb)$. For example $w = aab.aabaabab = aabaabaa.bab$.

Conjugates of Christoffel words play a special role. These words were called α-words in [22]. Note that the standard words of Aldo de Luca are particular conjugates of Christoffel words; a word on the alphabet $\{a, b\}$ is called *standard* if it may be written mab or mba for some Christoffel word bma or amb (that is, m is central, see below).

A word m is central if and only if amb is a Christoffel word. In this case, assume that the latter has the standard factorization $amb = uv$; then m has the periods $|u|, |v|$. Moreover *the longest palindromic proper prefix (suffix) of m has length* $\max(|u|, |v|) - 2$; this follows indeed from the fact that the smallest period of m is the smallest among the numbers $|u|, |v|$ ([20] Proposition 3). Moreover a palindrome m has the period p if and only if it has a palindromic prefix of length $|m| - p$ (see [40] Lemma 3).

A sequence s is Sturmian if and only if for any n it has exactly $n + 1$ factors of length n. A *Sturmian word* is a finite factor of a Sturmian sequence. It was proved by Serge Dulucq and Dominique Gouyou-Beauchamps that a word w on the alphabet $\{a, b\}$ is Sturmian if and only if it is *balanced*, that is $||u|_a - |v|_a| \leq 1$ for any factors u, v of equal length of w, Théorème 3.1 in [29]. A word is Sturmian if and only if it is a factor of some Christoffel word, see [15,22].

A *standard Sturmian sequence* may be defined geometrically: given a half-line on the positive quadrant, starting at the origin, with irrational slope, consider the infinite path located below the half-line and such that there is no integer point inside the infinite polygon delimited by the half-line and the path, cf. [15]. Then the associated standard Sturmian sequence is obtained by coding the steps of the path by a binary alphabet, and by removing the first letter.

For all this and more on the subject, see [8,37,55]. See also [7] for many equivalent definitions of central, Christoffel and standard words.

4. Factor sets of Sturmian sequences

We call *factor set* the set of words of length n of a sequence s, for some natural number n and some Sturmian sequence s. We call n the *length* of the factor set. The factor set is *proper* if $n \geq 2$.

Factor sets were counted by Gwenaël Richomme and Patrice Séébold, who proved the following theorem (see [58] Prop. 4.4).

Theorem 4.1. *Fix the alphabet $\{a, b\}$ and let $n \geq 1$. The mapping $F \mapsto l$, where F is a factor set of positive length and l is the unique word such that $al, bl \in F$, is a bijection from the set of factors sets of length n onto the set of left special Sturmian words of length $n - 1$. The number of factor sets of length n is $\varphi(1) + \cdots + \varphi(n)$.*

Recall that a *left special Sturmian word* is a word $l \in \{a, b\}^*$ such that al, bl are both Sturmian words. It is useful to denote $l(F)$ the unique word $l = l(F)$ defined in the proposition.

It is of interest to quote here a result of Dominique Perrin and Antonio Restivo, since it says something about the word $l(F)$ in the previous result.

Theorem 4.2. *(See [53] Proposition 2.) Let F be a factor set on the alphabet $\{a < b\}$.*

(i) *For the alphabetical (or lexicographic) order, $al(F)$ is the smallest element of F and $bl(F)$ its largest element;*
(ii) *two words u, v of F of the same length are consecutive in the alphabetical order if and only if $u = rabs$ and $v = rbas$ or if $u = ra$ and $v = rb$.*

5. Inverse of the Richomme–Séébold mapping

We describe now the inverse mapping of the Richomme–Séébold mapping. For this, recall a fundamental construction, due to Aldo de Luca [40] Lemma 5 pp. 59–60: given a word w, there exists a unique shortest word, which is a palindrome, and which has w as prefix; it is called the *right palindromic closure* and denoted $w^{(+)}$; one may compute it by the formula $w^{(+)} = ps\tilde{p}$, where $w = ps$ and s is the longest palindromic suffix of w. For example, $(abbab)^{(+)} = ab\,bab\,ba$, with $p = ab$, $s = bab$. See also [8] p. 24. Note that since a letter is a palindrome, s is at least of length 1, and one has for each nonempty word w

$$|w| \leq |w^{(+)}| \leq 2|w| - 1. \tag{1}$$

We say that a word m is a *circular factor* of the word w, if m is a factor of some conjugate of w. Equivalently, m is a factor of length $\leq |w|$ of the word ww, or of w^∞.

Proposition 5.1. *Let $n \geq 1$ and l be a left special Sturmian word of length $n - 1$. Let $w = al^{(+)}b$. Then w is a Christoffel word and the set of circular factors of length n of w is the inverse image of l in the Richomme–Séébold mapping. The length of w satisfies the inequalities $n + 1 \leq |w| \leq 2n - 1$.*

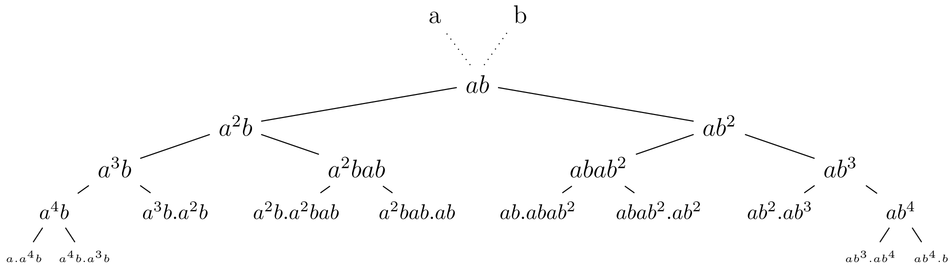

Fig. 2. The Christoffel tree.

Proof. Since l is a left special Sturmian word, $l^{(+)}$ is a central word, see [40] Theorem 3. Hence w in the statement is a Christoffel word and by Eq. (1), its length satisfies the given inequalities: indeed $|w| = |l^{(+)}| + 2 \geq |l| + 2 = n + 1$ and $|w| \leq 2|l| - 1 + 2 = 2(n-1) - 1 + 2 = 2n - 1$.

Since $n < |w|$, it follows from [22], Theorem 6.4 (see also [16] Theorem 4.1), that there are $n+1$ distinct circular factors of length n of w; moreover, since ww is a factor of some Sturmian sequence s, these $n+1$ factors are factors of s. Hence they form a factor set F of length n. Now, since l is a prefix of $l^{(+)}$, al is a prefix of $w = al^{(+)}b$, hence $al \in F$; moreover, $\tilde{w}$ is a circular factor of w (this follows since w and $\tilde{w}$ are conjugate, because w is a Christoffel word); since $l^{(+)}$ is a palindrome, $\tilde{w} = bl^{(+)}a$ and therefore $bl \in F$. Thus, $F \mapsto l$ for the Richomme–Séébold mapping. □

What is the set of Christoffel words obtained for fixed n in the previous proposition? This is answered in the next result.

Proposition 5.2. *Suppose that $n \geq 1$. The mapping $l \mapsto al^{(+)}b$, for l in the set of left special Sturmian words of length $n-1$, is a bijection onto the set of proper Christoffel words w, of length $\geq n+1$, whose standard factorization $w = uv$ satisfies $|u|, |v| \leq n$.*

Proof. We know that $w = al^{(+)}b$ is a proper Christoffel word of length $\geq n+1$; let $w = uv$ be its standard factorization. Note that al is a prefix of w.

Suppose by contradiction that $|u| > n$; then al is a proper prefix of u; $u = amb$ for some palindrome m (since u is of length at least 2, because $n \geq 1$); thus l is a prefix of m and $l^{(+)}$ is not longer than m, too; hence $w = al^{(+)}b$ is not longer than $u = amb$, a contradiction.

The proof that $|v| \leq n$ is symmetric, noting that the mapping of Proposition 5.1 has a right-to-left variant and that the right palindromic closure of l is equal to the left palindromic closure of $\tilde{l}$.

Thus the mapping of the proposition is well-defined. It is injective since it is increasing for the alphabetic order: the words l all have the same length, so have the words al and the order of the words $al^{(+)}b$ is the same, since such a word has al as prefix.

The fact that the mapping is surjective may be proved directly, but a counting argument suffices. We show that the set of Christoffel words described in the proposition has cardinality $\varphi(1) + \cdots + \varphi(n)$.

This follows from the *Christoffel tree* of [9], page 200 (see also [8] page 21). The nodes of this infinite complete binary tree are all proper Christoffel words[4]; each proper Christoffel word w has a standard factorization $w = uv$, where u, v are closer to the root of the tree or $u, v = a$ or b; more precisely if w is not the root ab, then one of u or v is its first ancestor.

Thus the set above is the set of leaves of the finite complete binary subtree of the previous infinite tree, whose internal nodes are all the proper Christoffel words of length at most n (note that each leaf of this subtree has length at least $n+1$); the number of such words is $\varphi(2) + \cdots + \varphi(n)$. Hence the number of leaves is 1 more, that is $\varphi(1) + \cdots + \varphi(n)$. □

The last part of the proof is illustrated in Fig. 2, for $n = 5$. We give for each Christoffel word, which is a leaf, its standard factorization, the cut being represented by a dot. The Christoffel words without dot are all those which are proper and of length ≤ 5.

We denote by $CF_n(w)$ the set of circular factors of length n of the word w.

Corollary 5.3. *Fix $n \geq 1$. The mapping $w \mapsto CF_n(w)$ from the set of Christoffel words, of length $\geq n+1$, whose standard factorization $w = uv$ satisfies $|u|, |v| \leq n$, is a bijection onto the set of factors sets of length n.*

Given a factor set F, we denote by $w(F)$ the unique Christoffel word which corresponds to it by the bijection of the corollary. One has $w(F) = al(F)^{(+)}b$.

[4] Recall the construction of this tree: ab is the root and if $w = uv$ (standard factorization) is a node, then its left (resp. right) successor is uuv (resp. uvv).

6. The Labbé bijection

The number of left special Sturmian words of length n is $\varphi(1) + \cdots \varphi(n+1)$, as it is shown in [43], proof of Theorem 7 (see also [38] proof of Theorem 2.2.36). Since this is also the number of lower Christoffel words of length at most $n+1$, minus 1, a direct bijection seems desirable. Such a bijection was discovered by Sébastien Labbé (private communication), which we give here with its kind permission. In what follows, the alphabet is $\{a < b\}$.

Theorem 6.1. *Let n be fixed, and consider the mapping from the set of lower Christoffel words, excluding b, of length at most $n+1$, which with $w = aw'$ associates the prefix of length n of the word $(w'a)^\infty$. It is a bijection onto the set of left special Sturmian words of length n.*

Proof. It is well-known that a word of the form $(w'a)^N$ is a prefix of a standard Sturmian sequence: indeed, one has simply to increase slightly the slope of w and one obtains a half-line defining such a standard Sturmian sequence. Hence $(w'a)^N$ is left special, by [47] (see also [40] Corollary 1, [38] Proposition 2.1.23). Since a prefix of a left special Sturmian word is also a left special Sturmian word, we conclude that the mapping in the statement is well-defined.

The number of lower Christoffel words of length n, excluding b, is $\varphi(n)$. Hence the set of definition of our mapping has $\varphi(1) + \cdots \varphi(n+1)$ elements. But so has the set of images, by the remark above.

Thus, it is enough to show that the mapping is injective. Suppose that w_1 and w_2 are mapped onto the same word. We may assume that $|w_1| \le |w_2|$; both lengths are $\le n+1$. The prefixes of length n of $(w_1'a)^\infty$ and of $(w_2'a)^\infty$ are equal, with $w_i = aw_i'$. Thus the prefixes of length $n+1$ of w_1^∞ and w_2^∞ are equal. If $|w_1| = |w_2|$, then $w_1 = w_2$. If not, this implies that w_2 has the period $|w_1|$: contradiction, since each Christoffel words, as each Lyndon word,[5] is not periodic. □

We can give the inverse bijection, by using the construction of Section 5.

Proposition 6.2.

(i) *The Labbé bijection is strictly increasing for the alphabetical order.*
(ii) *The inverse bijection of the Labbé bijection is given as follows: let l be a left special Sturmian word of length n on the alphabet $\{a, b\}$. Then the inverse image of l is the first component of the standard factorization of the Christoffel word $al^{(+)}b$.*

Proof. (i) It is enough to show that it is increasing for the alphabetical order, and for this we show that it is the product of three increasing mappings. The first one maps a Christoffel word w onto the infinite word w^∞; it is well-known that if w is a Christoffel word (even a Lyndon word), this mapping is increasing for the alphabetical order. The second mapping takes an infinite word and gives back its prefix of length $n+1$; this mapping is increasing. The third mapping takes a word of length $n+1$ beginning by a and removes this a; this mapping is increasing. Clearly, the Labbé mapping is the product of these three mappings.

(ii) We know by Proposition 5.1 that $w = al^{(+)}b$ is a Christoffel word. Let $w = uv$ be its standard factorization. By Proposition 5.2, $|u| \le n+1$ (since $|l| = n$). It is enough to show that the prefix of length n of $(u'a)^\infty$ is equal to l, where $u = au'$. Write $w = axb$. Suppose first that $v \ne b$. Then v begins by a, so that x begins by $u'a$. Moreover, x has period $|u| = |u'a|$ as follows from Section 3. Hence we obtain that l is a prefix of a power of $u'a$, because l is a prefix of $l^{(+)} = x$.

It remains the case $v = b$. Then $w = ab^N$, which implies that $l = b^n$; thus $l^{(+)} = l = b^n$, $w = ab^{n+1} = ab^n.b$, $u = ab^n$ and the prefix of length n of $(b^n a)^\infty$ is indeed equal to l. □

Proposition 6.3. *The mapping from the set of left special Sturmian words of length n, ordered alphabetically, into $\mathbb{N}$ (resp. $\mathbb{N} \cup \infty$), which with a word associates the numbers of b's (resp. the slope of this word), is increasing.*

This is seen for example for $n = 5$; the set is

$$\{a^5, a^4b, a^3ba, a^2ba^2, aba^2b, ababa, babab, bab^2a, b^2ab^2, b^3ab, b^4a, b^5\}.$$

Proof. Since $j \le j'$ if and only $\frac{j}{n-j} \le \frac{j'}{n-j'}$, it is enough to prove the assertion with the number of b's.

We claim that if w_i, $i = 1, 2$ are two Christoffel words such that $w_1 < w_2$ (alphabetical order), then $|u_1|_b \le |u_2|_b$, where u_i is the prefix of length $n+1$ of w_i^∞. By Theorem 6.1 and Proposition 6.2(i), the claim implies the proposition.

In order to prove the claim we use the geometrical definition of Christoffel words. Let w_i as above and L_i the half-line through the origin whose slope s_i is equal to that of w_i. It follows from [15] that the alphabetical order of Christoffel words coincides with the order of their slopes. Hence $s_1 < s_2$. Hence the half-line L_1 is under L_2. Let P_i the discrete path under

[5] *A Lyndon word* is a word that is primitive and the smallest in its conjugation class; equivalently, it is smaller than any of its nontrivial proper suffix, see [37] Proposition 5.1.2; thus it is necessarily unbordered.

L_i such that there is no integer point between P_i and L_i. Then, likewise, P_1 is under P_2, with possible common parts. Let M_i be the intersection of P_i with the line $x + y = n + 1$. Then the y-coordinate of M_1 is $\leq$ that of M_2. Moreover M_i is the endpoint of the path coded by the word u_i. Thus $|u_1|_b \leq |u_2|_b$ since $|u_i|_b$ is the y-coordinate of M_i. □

7. Classification of factors sets

We say that a factor set (of length n) is of type **C** if it contains a conjugation class of cardinality n. Otherwise we say that it is of type **P**. Note that if F is of type **C**, then it is the union of a conjugation class and of another word.

Assume that the alphabet is $\{a, b\}$. The only factor set of length 0 is $\{1\}$ and the only factor set of length 1 is $\{a, b\}$. Both are of type **C** and we disregard these cases in the sequel.

For example, with $n = 5$, $F_1 = \{a^2ba^2, a^2bab, aba^2b, ababa, ba^2ba, baba^2\}$ is of type **C** and $F_2 = \{a^3ba, a^2ba^2, aba^3, aba^2b, ba^3b, ba^2ba\}$ is of type **P**; indeed, the 5 last words in F_1 are conjugate, and F_2 meets 2 conjugation classes by at least 2 elements. Note that we have given the set in alphabetical order; so, we see that (according to Theorem 4.2) $l_1 = l(F_1) = aba^2$ and $l_2 = l(F_2) = a^2ba$.

In order to illustrate also the function $w(F)$, $l_1^{(+)} = aba^2ba$, since the longest palindromic suffix of l_1 is a^2, so that $w(F_1) = al_1^{(+)}b = a^2ba^2bab$; the reader may verify that the set of circular factors of length 5 of the latter word is indeed F_1. For F_2, we have $l_2^{(+)} = a^2ba^2$, since the longest palindromic suffix of l_2 is aba; then $w(F_2) = al_2^{(+)}b = a^3ba^2b$ and similarly, F_2 is the set of circular factors of length 5 of a^3ba^2b.

Proposition 7.1. *Let the alphabet be $\{a < b\}$. Let F be a factor set of type* **C** *and of length $n \geq 2$. Then for some Christoffel word $x = amb$ of length n, F is the union of the conjugation class of x and of either the word ama or the word bmb.*

In our example F_1 above, $x = a^2bab$, $m = aba$ and $ama = a^2ba^2 \in F$.

Proof. Let F be a factor set of type **C**. Then for some primitive word x, F contains the conjugation class of x. One conjugate of x is a Lyndon word, and therefore a Christoffel word by Theorem 3.2 of [9]. We may assume that it is x itself and that $x = amb$. Note that x is the smallest element of the class, and $\tilde{x} = bma$ the largest. There is one more word in F. Thus x is the smallest element of F or $\tilde{x}$ is the largest. This shows by Proposition 4.2 that $al(F) = amb$ or $bl(F) = bma$. In the first case, $bl(F) = bmb$ is in F and in the second case, $al(F) = ama$ is in F. This concludes the proof, since neither ama nor bmb is conjugate to amb, by counting the letters, and since the cardinality of F is $|x| + 1$. □

Note that the existence of factor sets of type **C** is not at all new. It has been noted by Wen and Wen [62], in their study of *singular factors* of Sturmian sequences, which correspond to the word ama or bmb in the proposition; see also [22] Theorem 6.4 and [37] Exercise 2.2.15.

Corollary 7.2. *If F is of type* **C***, then $l(F) = ma$ or mb for some central word m.*

Proposition 7.3. *Let F be a factor set of type* **P** *and of length $n \geq 2$. Let $w = w(F)$ and uv be its standard factorization. Then u and v are of length $< n$ and F is the set of words of length n having a periodic pattern conjugate to u or v.*

We prove this result in Section 9. In our example F_2 above, $w(F_2) = a^3ba^2b$, $u = a^3b$, $v = a^2b$ and the reader may verify that each word in F_2 has a periodic pattern conjugate to u or v. Moreover, if a word of length 5 has a periodic pattern conjugate to a^3b or to a^2b, then it is a factor of $aaabaaab$ or $aabaaba$, and therefore in F_2, as is easily verified; note that a word may have two periodic patterns conjugate one to a^3b and the other to a^2b, as the word a^2ba^2.

Corollary 7.4. *The number of factor sets of type* **C** *and length n is $2\varphi(n)$ and the number of factor sets of type* **P** *and length n is $\varphi(1) + \cdots + \varphi(n-1) - \varphi(n)$.*

Incidentally, this last number is therefore nonnegative; note that it is also equal to the number of Sturmian words of length $n - 2$ that are left special but not central. Indeed, this follows from [43] Theorem 6.

Proof. The two cases in Proposition 7.1 are mutually exclusive: indeed, the set $\{ama, bmb\}$ is not balanced, so cannot be included in a single factor set. Since the number of Christoffel words of length n is $\varphi(n)$, and since two of them are never conjugate (being Lyndon words), we obtain that the number of factor sets of type **C** is $2\varphi(n)$. The second assertion follows from Theorem 4.1. □

Corollary 7.5. *A factor set is of type* **C** *if and only if it contains an unbordered word (resp. a Christoffel word). A factor set is of type* **P** *if and only if all its members are periodic.*

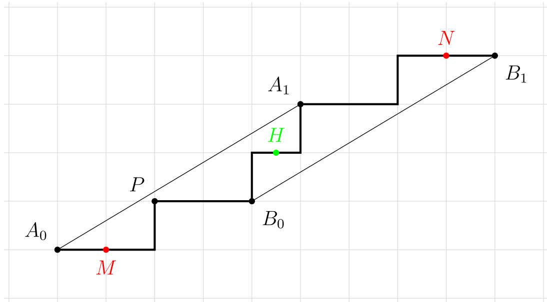

Fig. 3. The Christoffel word *aabaabab*.

Proof. The words in a factor set of type **P** are all periodic, by Proposition 7.3. Moreover, a Christoffel word is unbordered, as are all Lyndon words. This proves the second assertion, since a factor set of type **C** contains a Christoffel word.

If a factor set contains an unbordered word (in particular if it contains a Christoffel word), then it must be of type **C**, by what we have just proved. And conversely, a factor set of type **C** contains a Christoffel word, hence an unbordered word. □

8. Periods of conjugates and factors

The result that follow will help us to characterize in the next section the two classes of factor sets. We determine periods of the conjugates of a Christoffel word, and of its circular factors, and then conversely show that words having certain periods must be circular factors.

Lemma 8.1. *Let $w = aw'$ be a proper Christoffel word with standard factorization $w = uv$. Let p be the prefix of length $|v| - 1$ of w. Then $w'p$ is a palindrome and has v as periodic pattern and thus has the period $|v|$, which is nontrivial if $|w| \geq 3$.*

Proof. Suppose that u, v are both proper Christoffel words. Then $u = amb$, $v = anb$, $w = ambanb$, where $m, n, mban$ are palindromes. Moreover w has the factorization into two palindromes $w = (ana)(bmb)$ (see e.g. [16] Proposition 6.1 or Section 3). Thus $p = an$, $w' = nabmb$, and $w'p = nabmban$ is a palindrome. Now, $w'p$ has $mban$ as a suffix, and $mban$ is a palindrome, as said above. Hence, by [40] Lemma 3 (see also [8] Lemma 4.11), $w'p$ has as period the length of the corresponding prefix, that is nab, which has the same length as $v = anb$. Note that v is a proper suffix of w, hence a factor of $w'p$, which has therefore v as periodic pattern. Now, if $|w| \geq 3$, then $|v| < |v| + |w| - 2 = |p| + |w'| = |w'p|$, and the period is nontrivial.

If u, v are not both proper Christoffel words, then $w = a^n b$ or $w = ab^n$, $n \geq 1$, and the proof is straightforward. □

Observe that the word $w'p$ of the lemma has evidently also the period $|w|$. Since it is of length $|w| - 1 + |v| - 1$, it is the central word associated with the two relatively prime periods $|v|$ and $|w|$, and associated also with the Christoffel word $aw'pb = wv = uvv$.

A geometric illustration of the lemma may be seen in Fig. 3: this figure has a central symmetry (around the point H), as is seen by putting it upside down; this may be used to prove the existence of all the palindromes involved in the proof. Each word is represented by a path (the letter a for a horizontal left to right step and b for a vertical upwards one): $w = A_0A_1$, $w' = MA_1$, $u = A_0P$, $v = PA_1$, $v' = A_0B_0 = A_1B_1$, $u' = B_0A_1$, $p = A_1N$. For more details, see [16] page 24.

We denote by C the *conjugation operator*, which is defined for any word $w = xw'$, with first letter x, by $C(w) = w'x$.

Proposition 8.2. *Let $w = uv$ be a proper lower Christoffel word with its standard factorization. Then for $i = 1, \dots, |v| - 1$, $C^i w$ has a nontrivial periodic pattern conjugate to v; and for $i = |v| + 1, \dots, |w| - 1$, $C^i w$ has a nontrivial periodic pattern conjugate to u.*

For example, $w = aab.aabab$, $|v| = 5$, $C^2 w = baaba.baa$ has the periodic pattern *baaba* conjugate to $v = aabab$; and $C^7 w = baabaaba = (baa)^2 ba$ has the periodic pattern *baa* conjugate to *aab*.

Note that the two missing values of i in the proposition, namely $i = 0$ and $i = |v|$, correspond respectively to w itself and to $\tilde{w}$, which is the corresponding upper Christoffel word. This follows from the fact (see Section 3) that the palindromic factorization $w = v'u'$ satisfies $|v'| = |v|$ and that $\tilde{w} = u'v'$. The two words $w, \tilde{w}$ are not periodic (since, for example, they are Lyndon words for the appropriate order on the two letters, and since Lyndon words are always unbordered). In the example, the upper Christoffel word is $C^5 w = babaabaa$.

Proof. We may assume that $|w| \geq 3$. With the notations of the lemma, it appears that the conjugates $C^i w$, $i = 1, \dots, |v| - 1$ are exactly all factors of length $|w|$ of the word $w'p$. Since the latter has the period $|v|$, with periodic pattern v, these conjugates have the same period, which is nontrivial because $|v| < |w|$, and they have a periodic pattern conjugate to v.

The other statement is left-right symmetric. □

We note that the period given in the proposition is not the smallest one in general. Consider indeed $w = aab.aabaabab$, with $v = aabaabab$ of length 8; then 5 is a period of $m = C^4 w = abaab.abaab.a$, whereas according to the proposition it has also the period 8. A simpler example is the Christoffel word $a^n b$: each conjugate is of the form $a^i ba^j$, $i + j = n$, and it has smallest period $1 + max(i, j)$, whereas $v = a^{n-1}b$ is of length n.

Corollary 8.3. *Let w be a proper Christoffel word with standard factorization $w = uv$. Each circular factor m of w, such that $|u|, |v| < |m| < |w|$, has a periodic pattern conjugate to u or v, and thus a nontrivial period equal to $|u|$ or $|v|$.*

Proof. The word m is a suffix of some conjugate of w. If the conjugate is one of those quoted in the proposition, we see by the condition $|u|, |v| < |m|$ that m has a nontrivial periodic pattern conjugate to u or v.

If m is a suffix of w, then by Lemma 8.1, m has the periodic pattern v, because $|m| < |w|$ hence m is a suffix of w'. Otherwise, m is a suffix of the upper Christoffel word $\tilde{w}$, and it has symmetrically the periodic pattern $\tilde{u}$, since the standard factorization of $\tilde{w}$ is $\tilde{v}\tilde{u}$ and we conclude since $u, \tilde{u}$ are conjugate. □

Note that the periodic pattern given in the corollary may be not the shortest one. For example, let $w = a^2ba^2bab.a^2bab$, of length 13, with the indicated standard factorization. The word aba^2ba^2ba, of length 9, is a circular factor of w; it has the periodic pattern ba^2ba^2ba, conjugate to $u = a^2ba^2bab$, of length 8; but this word is equal to $(aba)^3$ and it has therefore also the period 3.

Proposition 8.4. *Let w be a proper Christoffel word with standard factorization $w = uv$. If a word of length at most $|w| - 1$ is a factor of u^∞ or v^∞, then it is a circular factor of w.*

Proof. It is enough to consider factors of length $|w| - 1$ of v^∞ (the case of u^∞ is symmetric). We show that such a factor is factor of $w'p$, with the notations of Lemma 8.1; this will be enough, since $w'p$ is a factor of ww.

Note that, since v is primitive and since $|w| - 1 \geq |v|$, v^∞ has $|v|$ distinct factors of length $|w| - 1$ and they appear consecutively. Thus, if x is any factor of length $|w| - 1 + |v| - 1$ of v^∞, then x contains as factors all these $|v|$ factors (see Lemma 2.1).

Now, by Lemma 8.1, $w'p$ is a factor of v^∞ of length $|w| - 1 + |v| - 1$. This ends the proof. □

9. Classification: proofs and consequences

Proof of Proposition 7.3. Let F be a factor set of length n and of type **P**. Let $w = w(F) = uv$, its standard factorization. We know that the length of u and v is at most n, by Corollary 5.3. Suppose by contradiction that u or v is of length n, $|v| = n$ say. We use Lemma 8.1 and its notations. The word $w'p$ is of length $|w| - 1 + |v| - 1 = |u| + 2|v| - 2 \geq 2|v| - 1$. Hence, this word has as factor the n conjugates of v. Thus these conjugates are circular factors of w and F is therefore of type **C**, contradiction.

We conclude that the words u, v are of length $< n$. Then by Corollary 8.3, each word in F, being a circular factor of w, of length $n < |w|$, has a periodic pattern conjugate to u or v. Conversely, each word of length n having a periodic pattern conjugate to u or v is by Proposition 8.4 a circular factor of w, hence is in F. □

Corollary 9.1. *Let F be a factor set of length $n \geq 2$. The following conditions are equivalent:*

(i) *F is of type* **C**;
(ii) *on the Christoffel tree, the first ancestor of $w(F)$ is of length n;*
(iii) *the longest palindromic proper prefix of $l(F)$ is of length $n - 2$.*

In the example of Section 7, with $n = 5$, we have $w(F_1) = a^2ba^2bab$ and indeed, this word has in Fig. 2 the first ancestor a^2bab of length 5. Moreover $l(F_1) = aba^2$ and its longest palindromic proper prefix is aba, of length 3.

Proof. We claim that if $y = x^{(+)}$, then x and y have the same longest proper palindromic prefix. This follows from the definition of the right palindromic closure.

(i) implies (ii): suppose that the first ancestor of $w(F)$ is not of length n. By Corollary 5.3, one has the standard factorization $w(F) = uv$ with $|u|, |v| \leq n$; moreover, the first ancestor of $w(F)$ is the longest one of u, v. Thus both are shorter than n. Then by Corollary 8.3, the circular factors of length n of $w(F)$ are periodic, so that F is of type **P** by Corollary 7.5.

(ii) implies (iii): we have $w(F) = al(F)^{(+)}b$. Let uv be its standard factorization. Then the maximum of the length of u and v is n. Hence by a property given in Section 3, the length of the longest proper palindromic prefix of $l(F)^{(+)}$ is $n - 2$. By the claim, this longest proper palindromic prefix is also the longest palindromic proper prefix of $l(F)$.

(iii) implies (ii): by the claim, the longest palindromic proper prefix of $l(F)^{(+)}$ is of length $n - 2$. Hence (by the same property in Section 3), the maximum of the lengths of u and v is n. The longest of them is the first ancestor of $w(F)$.

(ii) implies (i): suppose that the first ancestor x of $w(F)$ is of length n. Let uv be the standard factorization of x. Then $w(F) = uuv$ or uvv contains x as factor. Thus x is a factor of length n of $w(F)$, hence the Christoffel word x is in F. This shows by Corollary 7.5 that F is of type **C**. □

The next result is not new. It was proved by Wai-Fong Chuan [23] Theorem 4.4 (see also [51] page 3 and [34] Theorem 10).

Corollary 9.2. *Nonempty Sturmian words are bordered, except Christoffel words which are unbordered.*

Proof. Christoffel words are Lyndon words, so are unbordered. Each Sturmian word is either a member of a factor set of type **P**, and then is periodic by Proposition 7.3; or a member of factor set of type **C**: hence it is conjugate of a Christoffel word $x = amb$, or it is ama or bmb by Proposition 7.1; in the first case, if it is not a Christoffel word, then by Lemma 8.2, it is periodic; in the second case, it is periodic, too. □

Corollary 9.3. *Let s be a Sturmian sequence and t be the standard Sturmian sequence of the same slope α. For each $n \geq 2$, the set of factors of length n of s is of type* **C** *if and only if the prefix of length $n-2$ of t is a palindrome.*

Proof. Let the alphabet be $\{a < b\}$. It is well-known that s and t have the same factors. Let $n \geq 2$ and F be the set of factors of length n of s. Let l be the prefix of length $n-1$ of t, and u be the longest proper palindromic prefix of l. The word l is a left special Sturmian word, as is each prefix of a standard Sturmian sequence. Thus l is also left special in the set of factors of s, so that $al, bl \in F$. Thus $l = l(F)$. By Corollary 9.1, F is of type **C** if and only if u is of length $n-2$. Thus the corollary follows. □

10. Length n from length $n+1$ and the factor graph

Let $n \geq 0$. Given a factor set F_1 of length $n+1$, there is a factor set F_0 of length n naturally associated with it: it is the set of factors of length n of the words in F_1. We denote $F_0 = \pi(F_1)$. Clearly, the mapping π, from the set of factor sets of length ≥ 1 to the set of all factor sets, is surjective.

Proposition 10.1.

(i) *The restriction of the mapping π to the set of factors sets of type* **P** *is injective;*
(ii) *if F_1 is of type* **C***, then there is exactly another factor set F_1' such that $\pi(F_1) = \pi(F_1')$;*
(iii) *a factor set F_0 of length n is of the form $\pi(F_1)$ for some factor set F_1 of type* **C** *if and only if $l(F_0)$ is a central word. In this case, F_0 is either of type* **P***, or $l(F_0) = a^{n-1}$ (resp. $= b^{n-1}$) and F_0 is the union of the conjugation class of $a^{n-1}b$ (resp. of ab^{n-1}) and of the word a^n (resp. b^n).*

As the proof shows, in case (ii) the two sets differ only by the fact that one set contains ama and the other bmb, with the notations of Proposition 7.1.

Proof. If $l(F_1) = w = w'x$, where x is the last letter of w, then $l(\pi(F_1)) = w'$. For given w' there are at most 2 choices for x. Since the mapping $F \mapsto l(F)$ is a bijection (Theorem 4.1), it follows that the fibers of the mapping π are of cardinality 1 or 2.

Let i (resp. j) denote the numbers of fibers of cardinality 2 (resp. 1) of the restriction of π to the set of factor sets of length $n+1$. Then $2i + j = \varphi(1) + \cdots + \varphi(n+1)$ and, by surjectivity, $i + j = \varphi(1) + \cdots + \varphi(n)$. Thus $i = \varphi(n+1)$. Thus the cardinality of the union of the fibers of cardinality 2 is equal to $2\varphi(n+1)$, which is equal to the number of factors sets of length $n+1$ of type **C**, by Corollary 7.4.

Now, use Proposition 7.1 and its notations, with n replaced by $n+1$: since $x = amb$ is conjugate to $\tilde{x} = bma$, $\pi(F)$ contains am and bm; necessarily, $l(\pi(F)) = m$; thus, by Theorem 4.1, the two factor sets described in Proposition 7.1 have the same image under π. The union of all these 2-element sets has cardinality $2\varphi(n+1)$, which by the previous remark and calculation proves (i) and (ii).

It proves also that if $F_0 = \pi(F_1)$ for some factor set of type **C** and length $n+1$, then $l(F_0) = m$ is a central word. The latter is of length $n-1$. The previous equality happens for $\varphi(n+1)$ sets F_0, since there are $2\varphi(n+1)$ sets F_1 of type **C**, and by what we have just seen; moreover, the number of central words of length $n-1$ is $\varphi(n+1)$. Since each central word is left special, we have proved the first assertion in (iii). Now, if F_0 is not of type **P**, hence of type **C**, then by Corollary 7.2, $l(F_0) = m'a$ or $m'b$ for some central word m'; in the first case, we have $m = m'a$, thus (since m, m' are palindromes) $m' = a^{n-2}$, $l(F_0) = a^{n-1}$ and F_0 contains $bm'a = ba^{n-1}$; in the second case $m' = b^{n-2}$, $l(F_0) = b^{n-1}$ and F_0 contains $am'b = ab^{n-1}$. This implies the last assertion, by Proposition 7.1. □

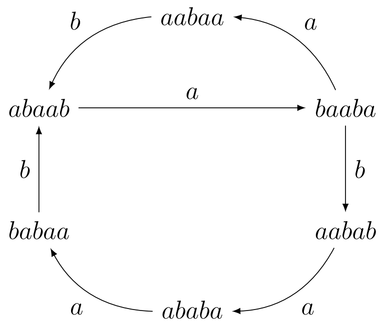

Fig. 4. A factor graph.

A *factor graph* [6] *of Sturmian words of length n* is a directed graph with set of vertices F_0 and set of edges F_1, where F_0 (resp. F_1) is a factor set of length n (resp. $n+1$) such that $F_0 = \pi(F_1)$; the word $w \in F_1$ represents the edge $u \to v$ with $u, v \in F_0$ and u (resp. v) prefix (resp. suffix) of w. For example, the word $w = baab$ represents the edge $baa \to aab$, with $n = 3$.

It is known (see [38] 2.2.3) that a factor graph (F_0, F_1) of Sturmian words has the following properties: let g (resp. d) be the unique word in F_0 which has two left (resp. right) extensions in F_1 (note that $g = l(F_1)$). Then the graph is composed of three paths: the first is from g to d, and the two others from d to g. For convenience, let us call the former path the *direct path* and the two others the *return paths*. The length of a path being the number of edges, the direct path may be of length 0 (that is, a loop around $g = d$, and this happens if and only if this equality holds); a return path may be of length 1, but only one of them, since a path of length 1 is completely determined by its two vertices. See Fig. 4, where $g = abaab$, $d = baaba$.

There is a canonical way to label the edges of a factor graph, so that it becomes an automaton over the alphabet, $\{a, b\}$ say: each edge determined by the word $w \in F_1$ is labelled by the last letter of w; in other words, this edge is $u \to v$ with $w = ux = yv$, $x, y \in \{a, b\}$, and the label is x, which is the last letter of w, and also of the endpoint v of the edge. This labelling induces by product a labelling, by a word in $\{a, b\}^*$, of each path in the graph.

Proposition 10.2. *Consider the factor graph determined by the factor set F_1. Then the two simple closed paths in the graph are, up to conjugation, labelled u and v, where $w(F_1)$ has the standard factorization uv.*

As an example, see Fig. 4: the closed paths are labelled aab and $abaab$; the factor set F_1 of the edges is the set of ux, where u, x are as above; then one sees that $l(F_1) = abaab$, since $babaab, aabaab \in F_1$. Thus $w(F_1) = al(F_1)^{(+)}b = aabaabab$ and $u = aab$, $v = aabab$.

Proof. Let $n+1$ be the length of the words in F_1. By symmetry, it is enough to show that there is a simple closed path in the graph whose label is conjugate to v.

Consider a word m with the following properties: it has length $n + |v|$, its factors of length $n+1$ are all in F_1, and it has the period $|v|$. Then the sequence of its $|v|+1$ consecutive factors of length n is the sequence of the vertices of a path in the graph; this path is closed because, by periodicity, the first and last factors are equal; the sequence of last letters of these factors, excluding the last factor, is conjugate to v, so that the label of the path is conjugate to v; moreover, the path is simple, since v is primitive.

Now, take the notations of Lemma 8.1 with $w = w(F_1)$. We apply the previous property to the suffix m of length $n + |v|$ of $w'p$. Note that it exists since the length of w is at least $n+2$ by Proposition 5.1 (with n replaced by $n+1$); hence, the length of w' is at least $n+1$ and that of $w'p$ is at least $n+1+|v|-1$. □

Corollary 10.3. *Let G be a factor graph of Sturmian words, with set of vertices F_0 of length n and set of edges F_1 of length $n+1$.*

(i) *The set of edges is uniquely determined by the set of vertices if and only $l(F_0)$ is not a central word;*
(ii) *if $l(F_0)$ is central, then there are two possible graphs, which differ by one edge. This happens if and only if one of the two return paths in the graph is of length 1.*

Proof. Clearly, $F_0 = \pi(F_1)$. Proposition 10.1 implies the corollary except its final assertion. But it is easily seen that if $F_0 = \pi(F_1) = \pi(F_1')$ for two distinct factor sets F_1, F_1', then, with the notations of Proposition 7.1, one has to cases: F_1 (resp. F_1') is the union of the conjugation class of amb and of ama (resp. and of bmb); since bma is conjugate to amb, if F_1 is the set

[6] It is also called *Rauzy graph*.

of edges, the graph contains the edge $am \to ma$ and $g = ma, d = am$, so that there is a return path of length 1; if the set of edges is F'_1, then there is the edge $bm \to mb$ and $g = mb, d = bm$.

Conversely, suppose that there is a return path of length 1. Since F_0 has $n + 1$ elements, the description above of the factor graph shows that there is simple closed path of length $n + 1$. Let w the label of this path. By Proposition 10.2, w is primitive, since it is the conjugate of a Christoffel word. Moreover, F_1 contains the conjugation class of w. Thus F_1 is of type **C** and $l(F_0)$ is by Proposition 10.1 a central word. □

11. Minimal periodic pattern of a Sturmian word

In [41] Theorem 1, Aldo de Luca and Alessandro De Luca show (using a result of Filippo Mignosi and Luca Zamboni [51] on Duval extensions) that the shortest periodic pattern of each nonempty Sturmian word is conjugate to a Christoffel word and they prove that this property characterizes Sturmian words. Actually their proof pages 121–122 prove more: if p is the smallest period of the Sturmian word m, then it has a periodic pattern of length p which is a Christoffel word.

Note that if a word m has a periodic pattern w which is unbordered (recall, nonperiodic), then necessarily the length of w is the smallest period of m (since the smallest period of m is also a period of w). The result of de Luca and De Luca recalled above shows that for Sturmian words, one has always this ideal situation: the smallest period of m coincides with the length of the longest unbordered factor of m. This kind of problem was first studied by Jean-Pierre Duval [30], who gives among other results families of words which do not satisfy the latter property (page 41). See also [34,51].

Note that it is not true that for each period p of a Sturmian word, the latter has a periodic pattern which is conjugate to a Christoffel word of length p; for example, $m = abaaaba$ (which is Sturmian since it is a conjugate of the Christoffel word a^3ba^2b) has the (nonminimal) period 6, but none of its two factors of length 6, $abaaab$ and $baaaba$, is conjugate to a Christoffel word.

From the previous result of de Luca and De Luca, we easily deduce a normal form for Sturmian words.

Proposition 11.1. *Each Sturmian word, which is not the power of a letter, has a unique factorization $m = sw^n p$, where w is a proper Christoffel word with palindromic factorization $w = v'u'$, $n \geq 1$, s is a proper suffix of u' and p is a proper prefix of w. A word of this form is Sturmian and has $|w|$ as smallest period.*

For example, $m = aba^2baba^2 = ab.a^2bab.a^2$, with $n = 1$, and $w = a^2bab = a^2.bab$ (palindromic factorization), ab a proper suffix of bab, a^2 a proper prefix of w.

We call *left normal form* the normal form given in the previous proposition. There is also a *right normal form* $sw^n p$, where this time s is a proper suffix of w and p is a proper prefix of v'.

We say that the left normal form $m = sw^n p$ is *short* if p is a proper prefix of v'; for such a word m, left and right normal form coincide. Otherwise, we say that the left normal form is *long*; then the right normal form is $(sv')(\tilde{w})^n p_1$, where $u'v' = \tilde{w}$ and where $p = v'p_1$ and p_1 is a proper prefix of u'; note also that the left normal form of $\tilde{m}$ is then $\tilde{p}_1 w^n(v'\tilde{s})$, since u', v' are palindromes.

The *reversal* of the left (resp. right) normal form $m = sw^n p$ is $\tilde{p}\tilde{w}^n\tilde{s}$: it is the right (resp. left) normal form of $\tilde{m}$, since $\tilde{w}$ is a Christoffel word with palindromic factorization $u'v'$.

Proof. Let m be a Sturmian word, on the alphabet $\{a, b\}$, and suppose that m is not a power of a nor b. By the result of de Luca and De Luca recalled above, m admits a factor w which is a Christoffel word, and of length equal to the smallest period of m; we choose the leftmost occurrence of such a factor; since w is a periodic pattern of m, we may write $m = sw^n p$, where $n \geq 1$, s is a proper suffix of w and p is a proper prefix of w. Now, w has the palindromic factorization $w = v'u'$, and $u'v'$ is the conjugate Christoffel word $\tilde{w}$. This implies that s must be a proper suffix of u': otherwise u' is a suffix of s, $s = s_1u'$ and $\tilde{w} = u'v'$ is a factor of $m = s_1u'(v'u')w^{n-1}p$, which is at the left of w, contradiction.

For uniqueness, suppose that $m = sw^n p$ with the indicated conditions. It follows from the hypothesis that the smallest period of m is $\pi = |w|$ and that the periodic patterns of length π are the conjugates of w. It is enough to show that w is the leftmost occurrence in m of a Christoffel word of length π. The only Christoffel words conjugate to w are w and $\tilde{w}$. Note that in a product $\cdots v'u'v'u'v'u' \cdots$, the only occurrences of $\tilde{w} = u'v'$ are the apparent ones, w being primitive. Thus, since s is a proper suffix of u', there is no occurrence of $\tilde{w}$ in $m = sw^n p$ at the left of the leftmost occurrence of w.

The last assertion follows from the fact that each power of a Christoffel word is a Sturmian word, as is well-known; moreover, the period cannot be smaller, since w is unbordered. □

Some particular classes of Sturmian words may be characterized by this normal form.

Proposition 11.2. *Let $sw^n p$ be the left normal form of some Sturmian word m, which is not the power of a letter, with $w = v'u'$ (palindromic factorization). Then*

(i) *m is a palindrome if and only if $p = v'\tilde{s}$;*
(ii) *m is a left special Sturmian word if and only if s is the longest proper suffix of u';*

(iii) *m is a central word if and only if s (resp. p) is the longest proper suffix (resp. prefix) of u' (resp. of w);*
(iv) *m is a right special Sturmian word if and only if and only if p is the longest proper prefix of v' or that of w;*
(v) *m is a bispecial Sturmian word if and only if s is the longest proper suffix of u' and if p is the longest proper prefix of v' or that of w;*
(vi) *m is a power of some Christoffel word if and only if $p = s = 1$;*
(vii) *m is the conjugate of some power with exponent ≥ 2 of some Christoffel word, without being a power of a Christoffel word, if and only if $ps = w$.*

Recall that a *bispecial* Sturmian word is a word that is left and right special.

Proof. (i) Suppose that $p = v'\tilde{s}$. Then $m = s(v'u')^n v'\tilde{s}$ is a palindrome.

Conversely, if m is a palindrome, then its left normal form cannot be short: otherwise, it is equal to its right normal form, and therefore the left normal form of $\tilde{m}$ is $\tilde{p}\tilde{w}^n\tilde{s}$; thus $\tilde{m}$ cannot be equal to m, by unicity of the left normal form and because $w \neq \tilde{w}$. Hence the left normal form satisfies $p = v'p_1$, p_1 a proper prefix of u', and the left normal form of $\tilde{m}$ is $\tilde{p}_1 w^n(v'\tilde{s})$ by a remark after Proposition 11.1. Thus, by uniqueness and since $\tilde{m} = m = sw^n p$, we must have $p = v'\tilde{s}$.

(ii) By Theorem 6.1, we know that a left special Sturmian word is of the form $m = w'w^n p$, for some Christoffel word w, where w' is the longest proper suffix of w, $n \geq 0$ and p is a proper prefix of w. Let $w = v'u'$ be the palindromic factorization of w (w is not a letter, since m is not a power of a letter).

Suppose first that $n \geq 1$ or $|p| \geq |v'|$; then the length of $w^n p$ is at least equal to that of v'. We may write $v' = av_1$, $w' = v_1 u'$. Note that m is a prefix of $w'w^\infty = v_1 u'(v'u')^\infty = v_1(u'v')^\infty$. Moreover the length of m is at least $|v_1| + |u'| + |v'|$ (indeed, either $n \geq 1$ and then the length of m is at least that of $w'w$, or $|p| \geq |v'|$ and the length of m is $\geq |w'p|$). Thus we have $m = v_1 u'(v'u')^n p$ and: either p is a proper prefix of v', $n \geq 1$, and $u'p$ is a proper prefix of $u'v'$; or v' is a prefix of p, $p = v'p_1$, p_1 proper prefix of u', and $m = v_1(u'v')^{n+1}p_1$. Thus we have $m = v_1(u'v')^k p'$, where v_1 is the maximal proper suffix of v' and p' is a proper prefix of the Christoffel word whose palindromic factorization is $u'v'$. Hence m has the required left normal form.

Suppose now that $n = 0$ and $|p| < |v'|$. Then by Lemma 8.1, $m = w'p$ has the period $|v|$, where $w = uv$, its standard factorization. We may write $w = u_1 v^k$, with $k \geq 1$ maximum, so that v has the standard factorization $v = u_1 v_1$ and the palindromic factorization $v = v'_1 u'_1$ with $|u'_1| = |u_1|$. Let $u_1 = ax_1$. Then $m = x_1 v^k p$ and x_1 is a suffix of v (because m has the period $|v|$) of length $|u_1| - 1$, hence it is the longest proper suffix of u'_1. Thus m has the required normal form, since p is a proper prefix of v (because m has the period $|v|$ and $|p| < |v'| = |v|$).

Conversely, suppose that $m = sw^n p$, in normal form, with s the maximal proper suffix of u'. Write $u' = xu_1$, x the first letter of u'. Then m is a prefix of $u_1 w^\infty = u_1(v'u')^\infty = u_1 v'(u'v')^\infty$. Hence m is a prefix of the word $u_1 v'(u'v')^N$ for some N. Since $u'v' = xu_1 v'$ is a Christoffel word, the former word is a left special Sturmian word by Theorem 6.1, hence m too.

(iii) Since a word is central if and only if it is a palindromic left special Sturmian word (see [42] Proposition 9), (iii) follows from (i) and (ii). Indeed, if s is the maximal proper suffix of u', then $\tilde{s}$ is the maximal proper prefix of u', hence $v'\tilde{s}$ is the maximal proper prefix of $w = v'u'$.

(iv) Suppose first that the left normal form is short; then the left and right normal forms coincide, so that (iv) follows by symmetry from (ii). Now, suppose that the left normal form is long; then the left normal form of $\tilde{m}$ is $\tilde{p}_1 w^n(v'\tilde{s})$, with the notations after Theorem 11.1; note that m is right special if and only if $\tilde{m}$ is left special; now by (ii), $\tilde{m}$ is left special if and only if $\tilde{p}_1$ is the longest proper suffix of u'; this is equivalent to the fact that p_1 is the longest proper prefix of u', hence that $p = v'p_1$ is the longest proper prefix of $v'u' = w$.

(v) This follows from (ii) and (iv).

(vi) and (vii) are evident. □

Note that normal forms for Sturmian words have been obtained previously: the representation of Sturmian words by triples of natural numbers, see Berstel–Pocchiola [11], Berenstein et al. [4].

12. Application of the normal form: counting

Since powers of letters are Sturmian words, we obtain the following consequence of Proposition 11.2.

Corollary 12.1. *In the algebra of noncommutative formal series in a, b over the integers, the sum of all Sturmian words is equal to*

$$1 + a^+ + b^+ + \sum_w S(u')w^+ P(w),$$

where the sum is over all proper Christoffel words w with palindromic factorization $w = v'u'$, with the following notations: $x^+ = \sum_{n>0} x^n$, $S(x)$ (resp. $P(x)$) is the sum of the proper suffixes (resp. prefixes) of the word x.

We recover the well-known counting formula for Sturmian words (Evgeny Petrovitch Lipatov [36], and independently Filippo Mignosi [48]).

Corollary 12.2. *Let s_n denote the number of Sturmian words of length n. Then*

$$\sum_n s_n x^n = x^* + (x^*)^2 \sum_{n\geq 1} \varphi(n) x^n.$$

We use the notation $x^* = \sum_{n\geq 0} x^n = (1-x)^{-1}$. In other words, for any $n \geq 0$,

$$s_n = 1 + \sum_{i+j=n} (j+1)\varphi(i)$$

with the convention that $\varphi(0) = 0$.

We use in the proof the following certainly well-known lemma.

Lemma 12.3. *Let $n \geq 2$. The number of 2 by 2 matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of determinant 1 over the natural numbers of total sum $a+b+c+d = n$ is $\varphi(n)$.*

In order to prove the lemma, we use a method using Christoffel words; the facts that we recall now will also serve in the proof of the counting formula.

Let the alphabet be ordered: $a < b$. It is known (see e.g. [13] Corollary 3.2) that there is a bijection between the set of proper lower Christoffel words on this alphabet and the set 2 by 2 matrices over the natural numbers of determinant 1; in this bijection the sum of the first (resp. the second) column of the matrix is equal to the length of u (resp. v), where $w = uv$ is the standard factorization of the corresponding Chritoffel word. Recall that then the palindromic factorization of w is $v'u'$ with lengths $u' = |u|$ and $|v'| = |v|$, and that the corresponding upper Christoffel word is $\tilde{w} = u'v'$.

Proof of Lemma 12.3. The number of lower Christoffel words of length n is $\varphi(n)$. By what has been just recalled, there is a bijection between them and the set described in the lemma. The latter follows. □

Proof of Corollary 12.2. In Corollary 12.1, by sending each letter onto a single one, x say, we obtain, using the remarks above, taking into account that w may be a lower or an upper Christoffel word,

$$\sum_n s_n x^n = 1 + \frac{2x}{1-x} + \sum_M (1 + x + \cdots + x^{a+c-1} + 1 + x + \cdots + x^{b+d-1}) \\ \times \frac{x^{a+b+c+d}}{1-x^{a+b+c+d}} (1 + x + \cdots + x^{a+b+c+d-1}),$$

where the sum is over all matrices $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of natural numbers with determinant 1. The latter expression may be simplified and is equal to

$$x^* + xx^* + \sum_M (1 + x + \cdots + x^{a+c-1} + 1 + x + \cdots + x^{b+d-1}) x^{a+b+c+d} x^*$$

If we subtract x^* from the formula that we have to prove and then multiply by $(1-x)^2$, we see that we must prove that

$$\sum_{n\geq 1} \varphi(n) x^n = x(1-x) + \sum_M (1 - x^{a+c} + 1 - x^{b+d}) x^{a+b+c+d}.$$

The right-hand side is equal to

$$x - x^2 + \sum_M (2x^{a+b+c+d} - x^{2a+b+2c+d} - x^{a+2b+c+2d}).$$

By Lemma 12.3, the number of matrices M as above, of total sum $a + b + c + d = n$, is equal to $\varphi(n)$ $(n \geq 2)$. Thus $\sum_M x^{a+b+c+d} = \sum_{n\geq 2} \varphi(n) x^n$. Moreover, it is well-known that the set of these matrices is a free monoid generated by the two matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Since $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & a+b \\ c & c+d \end{pmatrix}$ (of total sum $2a+b+2c+d$) and $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} a+b & b \\ c+d & d \end{pmatrix}$ (of total sum $a+2b+c+2d$), one has $\sum_M x^{a+b+c+d} = x^2 + \sum_M (x^{2a+b+2c+d} + x^{a+2b+c+2d})$ and therefore

$$\begin{aligned}\sum_{n\geq 1}\varphi(n)x^n &= x + \sum_{n\geq 2}\varphi(n)x^n \\ &= x + 2\sum_M x^{a+b+c+d} - (x^2 + \sum_M (x^{2a+b+2c+d} + x^{a+2b+c+2d})) \\ &= x - x^2 + \sum_M (2x^{a+b+c+d} - x^{2a+b+2c+d} - x^{a+2b+c+2d}). \qquad \square\end{aligned}$$

13. Application of the normal form: algorithms

Recall that each word m on a totally ordered alphabet has a unique factorization of the form $m = l_1^{n_1} \cdots l_k^{n_k}$ where the l_j's are Lyndon words such that $l_1 > \ldots > l_n$ and where the n_i's are positive natural numbers, see [37] Theorem 5.1.5. This is called the *Lyndon factorization* of m. Recall also that each lower Christoffel word is a Lyndon word.

Proposition 13.1. *Let m be a Sturmian word on the totally ordered alphabet $\{a < b\}$ which has the lower Christoffel word w as periodic pattern. Let $m = l_1^{n_1} \cdots l_k^{n_k}$ be the Lyndon factorization of m. Then for some j, $w = l_j$, $l_1^{n_1} \cdots l_{j-1}^{n_{j-1}}$ is a proper prefix of w, $l_{j+1}^{n_{j+1}} \cdots l_k^{n_k}$ is a proper suffix of w and l_j is longer than any other l_i.*

Proof. The proof is obtained by using the methods of the proof of Proposition 7 in [18]. By hypothesis we may write $m = sw^k p$ where s (resp. p) is a proper suffix (resp. prefix) of w. We know that m is a factor of some power of the Christoffel word w. Hence the word m determines a discrete path in the plane, which is obtained by discretizing from below a segment of the straight line of slope equal to that of the Christoffel word w. Since a line is convex, it follows, as in [18] that the Lyndon factors of m are all Christoffel words, and one of them is w, since the line contains the endpoints of a path determined by w.

Thus for some j, $w = l_j$, $n = n_j$, $s = l_1^{n_1} \cdots l_{j-1}^{n_{j-1}}$ and $p = l_{j+1}^{n_{j+1}} \cdots l_k^{n_k}$.

Now, s and p are shorter than w, so that the l_i are shorter than l_j if $i \neq j$. $\square$

From the proposition, we easily deduce an algorithm which checks if a given word is a Sturmian word. Indeed, we know by the de Luca–De Luca theorem (see the beginning of Section 11) that each Sturmian word on the totally ordered alphabet $\{a < b\}$ has a periodic pattern which is a Christoffel word; it is either a lower Christoffel word, or an upper Christoffel word, in which case the word m' obtained by exchanging a and b in m, has as periodic pattern a lower Christoffel word; note that m and m' are simultaneously Sturmian or not (it can happen that only one of them has a lower Christoffel word as periodic pattern). Hence the algorithm does in turn for m and m' the following steps:

1. Find the Lyndon factorization $m = l_1^{n_1} \cdots l_k^{n_k}$.
2. Check if there is a unique longest Lyndon word. If no, then stop. If yes, let it be l_j.
3. Check if l_j is a Christoffel word. If no, then stop.
4. Check if $l_1^{n_1} \cdots l_{j-1}^{n_{j-1}}$ is a suffix of l_j and if $l_{j+1}^{n_{j+1}} \cdots l_k^{n_k}$ is a prefix of l_j. If no, then stop. If yes, then m is a Sturmian word.

If the algorithm, applied to m and m', always stops, then m is not a Sturmian word.

The algorithm is linear with respect to the length of m. Indeed, finding the Lyndon factorization is linear, by a beautiful algorithm of Duval [31] (see also [56] Section 7.4). Moreover, checking if a given word is a Christoffel word is linear too: one has simply to count the number of a and b's in the word and then generate the corresponding Christoffel word by the Cayley graph algorithm (see [8] Section 1.2 and [18] p. 2243), and then check if the two words coincide.

A closer look at this algorithm shows that it may be used to compute the left normal form of m. Indeed, if m is a Sturmian word, the algorithm outputs a factorization $m = sw^n p$, where w is a Christoffel word and where s (resp. p) is a proper suffix (resp. prefix) of w. Write $w = v'u'$, the palindromic factorization. If s is a proper suffix of u', the left normal form is $m = sw^n p$; if not, then $s = s'u'$ and the left normal form is $m = s'(u'v')^n(u'p)$.

One must of course know how to generate in linear time the palindromic factorization of Christoffel word. This is done using as above the Cayley graph definition of a Christoffel word: the factorization is obtained by cutting the word at the maximum value of the Cayley graph (geometrically it corresponds to the integer point on the path that is furthest from the segment A_0A_1, see Fig. 3).

14. Application of the normal form: completion and contraction

The set of Sturmian words on a fixed alphabet, ordered by the prefix ordering, has the following properties: there is a unique minimal element (the empty word); each element is covered[7] by one or two elements (the elements covered by two elements are the right special Sturmian words); each element has an upper bound which is a right special Sturmian word.

Several subsets of the set of Sturmian words have similar properties. It is therefore convenient to introduce the following definition: say that a poset is an *S-poset* if the three properties below are satisfied:

- there are finitely many minimal elements;
- each element is covered by one or two elements;
- call *special* an element which is covered by two elements; then each element has an upper bound which is special.

Thus we see that the set of Sturmian words on $A = \{a, b\}$ with the prefix order is an S-subset.

Given an element x in an S-poset, we call *completion* of x the smallest special element which is larger than x or equal to it; and we call *contraction* of x the greatest special element which is smaller than x or equal to it.

Consider now the set of left special Sturmian words with the prefix order: it is also an S-subset. Indeed, it is proved in [43] Lemma 8, that each left special Sturmian word m may be extended by one letter at the right into a left special Sturmian word. It may happen that they are two such extensions, in which case this word is *right special with respect to the set of left special Sturmian words*. It is also shown that this happens if and only if m is a central word, see [43] (Lemma 6 and Corollary 1, noting that a word is *strictly bispecial* if and only if it is a left special Sturmian word having two right extensions by a letter into a left special Sturmian word; see also [38] Proposition 2.2.34). Finally, each left special Sturmian word w is prefix of some central word: indeed, an already mentioned result of de Luca asserts that $w^{(+)}$ is central [40] Theorem 3.

As third example, consider the set of Sturmian palindromes on A with the median factor order. Recall that u is a *median factor* of v if $v = xuy$ where x, y have equal length. We know by [42], Proposition 7 and Lemma 10, that for each Sturmian palindrome w, awa or bwb is Sturmian and that both are exactly when w is central. We know also by [42] Corollary 4, that each Sturmian palindrome is the median factor of some central word. Thus the set of Sturmian palindromes with the median factor order is an S-subset. Note that there are three minimal elements, namely $1, a$ and b.

Thus, in the three previous S-posets, we know what are the special elements. In the three propositions below, we show how to compute the completion and the contraction of each element, using the left normal form, and we compare the minimal periods of these elements.

Recall the left normal form sw^np of the Sturmian word m: $w = v'u'$ is the palindromic factorization of the Christoffel word w, s is a proper suffix of u' and p is a proper prefix of w.

Proposition 14.1. *Let $m = sw^np$ be a Sturmian word written in left normal form. We assume that w is not the power of a letter.*

(i) *The shortest right special Sturmian word of which m is a prefix is sw^np_1, where p_1 is the longest proper prefix of v' if p is a proper prefix of v', and where p_1 is the longest proper prefix of w otherwise.*
(ii) *Suppose that m is not right special (that is, p is neither the longest proper prefix of w nor that of v'). The longest prefix of m which is a right special Sturmian word is sw^np_1 where p_1 is the longest proper prefix of v' if v' is a prefix of p; and if p is a proper prefix of v', it is $sw^{n-1}p_2$, where p_2 is the longest proper prefix of w.*

Proof. Recall that, by Proposition 11.2(iv), m is right special if and only either p is the longest proper prefix of w or if p is the longest proper prefix of v'. From this remark, (i) follows immediately. Likewise, (ii) follows if we verify that $sm^{n-1}p_2$ is right special: this is surely true if $n - 1 \geq 1$, by the same remark. Suppose that $n = 1$; we have seen in the first part of the proof of Theorem 6.1 that if $w = aw'$ is a Christoffel word with first letter a, then $(w'a)^N$ is left special, together with all its prefixes; symmetrically, we have that the word sp_2 is right special, since it is a suffix of $(bw'')^2$, writing $w = w''b$. □

Proposition 14.2. *Let $m = sw^np$ be a Sturmian palindrome written in left normal form.*

(i) *The shortest central word of which m is a median factor is $s_1w^np_1$, where s_1 (resp. p_1) is the longest suffix of u' (resp. prefix of w).*
(ii) *If m is not a central word, then the longest central word which is a median factor of m is $s_2(u'v')^{n-1}u'\tilde{s}_2$, where s_2 is the longest proper suffix of v'.*

Proof. Recall that, by Proposition 11.2(i), $p = v'\tilde{s}$. By the same result, the proposition will follow provided we show that $s_2(u'v')^{n-1}u'\tilde{s}_2$ is a central word. This is clear if $n > 1$, by the same result again. Suppose that $n = 1$. The word to consider is $s_2u'\tilde{s}_2$. By the remark after Lemma 8.1, this word is central (because it is the word $w'p$ with the notations of this remark). □

[7] In a poset, an element x is *covered by* y if $x < y$ and if there is not element lying strictly between them.

Proposition 14.3. *Let $m = sw^n p$ be a left special Sturmian word written in left normal form.*

(i) *The shortest central word of which m is a prefix is $sw^n p_1$ where p_1 is the longest proper prefix of w;*
(ii) *the longest central word which is prefix of m is $sw^{n-1} p_1$, with p_1 as above.*

Proof. By Proposition 11.2(ii), s is the longest proper prefix of u'. From (iii) in the same proposition, we deduce (i). The statement (ii) follows similarly if we show that $sw^{n-1}p_1$ is central. This is the symmetrical statement of a fact shown in the proof of the previous proposition. □

Corollary 14.4. *In the previous three S-posets, each word and its completion have the same minimal period. Moreover, a word $m = sw^n p$ and its contraction x have the same period, except in the three following cases:*

- *In the case of all Sturmian words, $n = 1$, p is a proper prefix of v' and m is not right special; then the smallest period of x is $|u'|$.*
- *In the case of Sturmian palindromes, $n = 1$; then the smallest period of x is $|v'|$.*
- *In the case of left special Sturmian words, $n = 1$; then the smallest period of x is $|u'|$.*

The statements on Sturmian palindromes in the corollary have been already proved in [42], Theorem 23. To see this, we have to verify that $|w| = |x| - |v'| + 2$. This is easily seen in the second case of the proof below.

Proof. The assertion on completions follows from Proposition 11.1, since in each of the three propositions the left normal form of the completion is given. For the contraction, we argue similarly, except for the cases listed in the corollary, which we examine closer now. Recall that the standard factorization of w is $w = uv$, with $|u| = |u'|$ and $|v| = |v'|$.

- The S-poset is the set of all Sturmian words, $n = 1$, p is a proper prefix of v' and m is not right special. Then $x = sp_2$, p_2 is the longest proper prefix of w and recall that s is a proper prefix of u'. It follows then from the dual version of Lemma 8.1 that x has the periodic pattern u; hence the smallest period of x is $|u|$.
- The S-poset is the set of all Sturmian palindromes and $n = 1$. Then $x = s_2 u' \tilde{s}_2$ where s_2 is the longest proper suffix of v'. Then, since $w = v'u'$, we have $s_2 u' = w'$, the longest proper suffix of w. By Lemma 8.1, we deduce that $x = w' \tilde{s}_2$ has the periodic pattern v, hence has the smallest period $|v|$. Note that $|x| = |w| - 2 + |v'|$.
- The S-poset is the set of all left special Sturmian words and $n = 1$; then $x = sp_1$, where p_1 is the longest proper prefix of w. We argue as in the first case, using that s is the longest proper suffix of u'. □

Recall that the right palindromic closure $m^{(+)}$ of a word m has been defined in Section 5.

Proposition 14.5. *Let $m = sw^n p$ be a Sturmian word written in left normal form. Then $m^{(+)}$ is equal to $sw^n v' \tilde{s}$ if p is a prefix of $v'\tilde{s}$ and to $sw^{n+1} v' \tilde{s}$ otherwise.*

Lemma 14.6. *Suppose that $w = v'u'$ is a Christoffel word with its palindromic factorization. Then in the bi-infinite sequence $\ldots v'u'v'u'v'u' \ldots$ the only occurrences of the factors u' and v' are those indicated.*

Proof. The bi-infinite sequence has a natural interpretation as bi-infinite path; see Fig. 3, with $w = aabaabab$, $v' = aabaa$, $u' = bab$. This path lies between two parallel straight lines, and the indicated factors u' and v' are the subpaths going from one line to the other: see the figure, where the paths from A_0 to B_0 and from A_1 to B_1 each give a factor v' and the path from B_0 to A_1 gives the factor u'.

Suppose that there were another factor u': it must then start from a point lying strictly between the two lines, so that its endpoint could not lie between the two lines, contradiction. □

Proof of Proposition 14.5. Suppose first that p is a prefix of $v'\tilde{s}$. Then p is a prefix of v' or it has v' as proper prefix. In the first case, m is shown as the first case of Fig. 5, where the right-hand end of m is the right-hand end of the curly bracket. The latter represents a palindromic suffix of m, since u', v' are palindromes. Note that the rightmost u' is sent by central symmetry around the centre of this palindrome onto the leftmost u', both shown on the figure. If it were not the longest palindromic suffix of m, then in a longer one, the rightmost u' would be sent by central symmetry of this longer palindrome on a factor u' of m which is strictly at the left of the leftmost u' indicated on the figure. This contradicts Lemma 14.6. Thus the longest palindromic suffix is the one indicated by the curly bracket and it follows that the palindromic closure of m is $sw^n v' \tilde{s}$, by the construction of $m^{(+)}$ recalled in Section 5.

In the second case, one argues similarly, with the rightmost and leftmost v' shown in the second case of Fig. 5.

Suppose now that p is not a prefix of $v'\tilde{s}$. Then m is shown in the last case of Fig. 5, terminating again at the right end of the curly bracket. The latter represents a palindromic suffix of m. By central symmetry around the centre of this palindrome, the two extreme v' in the bracket correspond each to another. If there was a longer palindromic suffix, the just mentioned rightmost v' should match by symmetry inside this new palindrome with the leftmost v' in the figure. But this

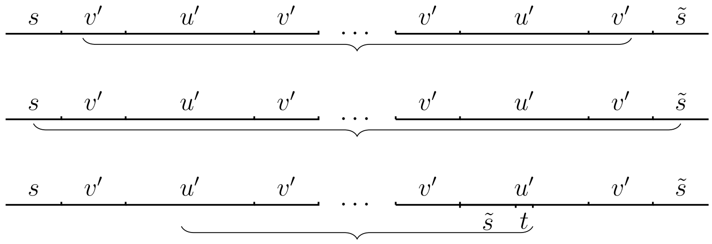

Fig. 5. Longest palindromic suffix of m.

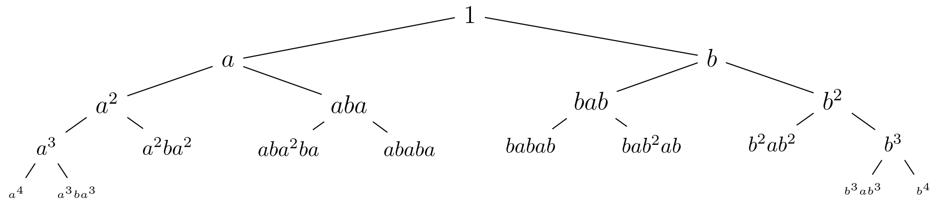

Fig. 6. The tree of central words.

is impossible because the word is too short: the last factor, indicated by a t, would fall outside the word by this symmetry. Thus we have found the longest palindromic suffix and we conclude as previously. □

The following result seems to be new.

Corollary 14.7. *The right palindromic closure of a Sturmian word is a Sturmian word, which has the same minimal period.*

15. Several infinite complete binary trees

The set of central words has naturally a structure of infinite complete binary trees, of which this words are the nodes. This follows from a wonderful construction due to Aldo de Luca, namely his function *Pal*, the *iterated palindromic closure*, see [40] Proposition 8 where this function is denoted ψ. This function is defined, using the palindromic closure $w^{(+)}$ defined in Section 5, recursively as follows: $Pal(1) = 1$ and for each word u and each letter x: $Pal(ux) = (Pal(u)x)^{(+)}$. It is then shown in [40] that the function *Pal* is a bijection from $\{a, b\}^*$ onto the set of central words on the alphabet $\{a, b\}$. If $w = Pal(u)$, then u is called the *directive word* of w.

Now, consider the complete infinite binary tree and encode each node by a word in $\{a, b\}^*$ in such a way that a means "left subtree" and b means "right subtree"; we call this word the *natural encoding* of the node. Label each node, which is naturally encoded by u, by the word $Pal(u)$. We obtain the *tree of central words*; see Fig. 6, where for example the node naturally encode by aba is labelled $Pal(aba) = aba^2ba$.

The tree of central words is a variant of the *Christoffel tree* of [9] Figure 3: replace in the tree of central words each label w by awb; see Fig. 2. Another variant is the *standard tree* [9] Figure 3; see also [38] Figure 2.5.

Recall the notion of S-poset of the previous section. Given a minimal element of such a poset, we may associate with it naturally an infinite complete binary tree: the nodes are the special elements of the poset; the two children of a special element x, covered by x_0, x_1, are y_0, y_1, where y_i is the special least upper bound of x_i.

Taking for instance the S-poset of left special Sturmian words, which has 1 as unique minimal element, the associated tree has as nodes the central words; the two children of the central word x are the shortest central words y_0, y_1 such that xa is a prefix of y_0 and xb is a prefix of y_1. Now, it is easily seen, by de Luca's construction above, that this tree is nothing else than the tree of central words. This tree may be also described as the Hasse diagram of the set of central words, ordered by prefix order. Note that, since these words are palindromes, the same tree is the Hasse diagram for the suffix order.

In the S-poset of palindromes, we have three minimal element $1, a, b$, thus three complete infinite binary trees; denote them T_1, T_a, T_b. The union of their nodes is disjoint and is the set of all central words. We describe now these trees.

Proposition 15.1. *Let $w = Pal(u)$ be a central word on the alphabet $\{a, b\}$. The shortest central word which contains awa as median factor is $Pal(v)$, with:*

(i) *if $u = a^i$ then $v = a^{i+2}$;*
(ii) *if $u = xba^i$, then $v = xab^{i+1}$.*

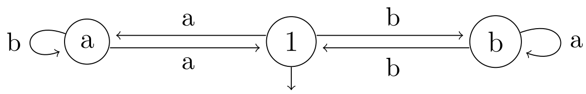

Fig. 7. A finite automaton.

Of course, there is a similar result by exchanging a and b, whose statement we omit. Both results have a geometric interpretation, which appears on the tree of central words, see Fig. 6; for example, if $w = b^2 = Pal(b^2)$, with $x = b$, $i = 0$, then awa has as shortest median central extension the word $Pal(bab) = bab^2ab$ which appears, one level lower, on the left side of the face on which right side w lies. We leave the general statement to the reader.

We need the following lemma, where we use the function θ of [3]. Let $A = \{a, b\}$. Then θ is the mapping from A^* into itself defined recursively by: $\theta(1) = 1$, $\theta(a^{n+1}w) = a^n\theta(w)$ if $w \in A^* \setminus aA^*$, $\theta(bw) = b\theta(w)$. In other words, θ erases one a in each maximal block of a's. For example, $\theta(a^5) = a^4$, $\theta(aababaaa) = abbaa$ and $\theta(baab) = bab$. For later use, note that if $w \in aA^*$, then $\theta(aw) = a\theta(w)$.

Lemma 15.2. *Let $v \in \{a, b\}^*$. Then $\theta(Pal(av)) = Pal(v)$.*

Proof. We use a formula due to Jacques Justin [35] (Lemma 2.1; see also lemma 4.7 in [8]) (actually a symmetric version of it): $Pal(av) = aR_a(Pal(v))$ where R_a is the monoid endomorphism of $\{a, b\}^*$ sending a onto a and b onto ba.

Now, if v contains no b, then the formula of the lemma is evident, since in this case $Pal(v) = v$. Suppose that v contains b; then so does $Pal(v)$ and we may write $Pal(v) = a^{i_0}ba^{i_1}b \cdots ba^{i_n}$. Then $Pal(av) = aR_a(Pal(v)) = a^{i_0+1}ba^{i_1+1}b \cdots ba^{i_n+1}$. Hence $\theta(Pal(av)) = a^{i_0}ba^{i_1}b \cdots ba^{i_n} = Pal(v)$. □

Proof of Proposition 15.1. We prove first that if u and v are as indicated, then $Pal(u)$ is a median factor of $Pal(v)$. This is clear in case (i). Now assume that $u = xba^i$ and $v = xab^{i+1}$ and argue by induction on $|x|$. If x is the empty word, then $Pal(u) = Pal(ba^i) = (ba)^ib$ and $Pal(v) = Pal(ab^{i+1}) = (ab)^{i+1}a = a(ba)^{i+1} = a(ba)^iba = aPal(u)a$ and therefore $Pal(u)$ is a median factor of $Pal(v)$.

Assume now that x is nonempty. If x begins by a, $x = ay$ say, then we use the previous lemma and we have: $\theta(Pal(u)) = \theta(Pal(ayba^i)) = Pal(yba^i)$ and $\theta(Pal(v)) = \theta(Pal(ayab^{i+1})) = Pal(yab^{i+1})$; by induction, $\theta(Pal(u)) = Pal(yba^i)$ is a median factor of $\theta(Pal(v)) = Pal(yab^{i+1})$; we deduce that $Pal(u)$ is a median factor of $Pal(v)$, since on one hand, both words begin by a, and on the other, if two palindromes begin by a and if the image of the first under θ is a median factor of the image of the second, then the first is a median factor of the second.

If x begins by b, we argue similarly by using the function obtained from θ by exchanging a and b.

Now, the proposition follows, since the smallest subset of $\{a, b\}^*$ containing $1, a, b$ and closed under the operations $c^n \mapsto c^{n+2}$, $c = a$ or b, $n \in \mathbb{N}$; $xcd^i \mapsto xdc^{i+1}$, $\{c, d\} = \{a, b\}$, $i \in \mathbb{N}$, is $\{a, b\}^*$, as may be seen easily by using the geometric interpretation mentioned above. □

Corollary 15.3. *The set of directive words of central words in $\{a, b\}^*$ of even length (equivalently, which have the empty word as median factor) is the rational language recognized by the automaton of Fig. 7 with initial state 1. The set of directive words of central words which have a (resp. b) as median factor is the rational language recognized by the automaton in the same figure with initial state a (resp. b).*

It is interesting to note that these languages are group languages: their syntactic monoid is the symmetric group of order 3.

Proof. For the first assertion, it is enough to prove that the smallest subset of $\{a, b\}^*$ containing 1 and which is closed under the operations mentioned in the previous proof, is the language L recognized by the automaton with initial state 1. Suppose for instance that xab^i is in L. Then there is a path from 1 to 1 labelled by this word. With classical notations, since the automaton is deterministic: $1.xab^i = 1$. If $i = 2j$, then, since the automaton is bi-deterministic, we must have $1.x = a$, so that $1.xba^{i+1} = a.ba^{i+1} = a.a^{i+1} = 1$ as is seen on the automaton, using the fact that $i + 1$ is odd; if $i = 2j + 1$, then we must have $1.x = b$, so that $1.xba^{i+1} = b.ba^{i+1} = 1$.

All other verifications are similar and left to the reader. □

16. On a theorem of Chuan

Following [24,25], call *moment* of the word $w = a_1 \cdots a_n$ of length n on the alphabet $\{0, 1\}$ the quantity $M(w) = \sum_i (n + 1 - i)a_i$.

Theorem 16.1. *(See Chuan [24,25].) The following conditions are equivalent for a word w of length n:*

(i) *w is conjugate to a Christoffel word;*
(ii) *the set of moments of the conjugates of w is an interval of length n in $\mathbb{N}$;*
(iii) *the difference between the largest and the smallest moment of the conjugates of w is $n - 1$.*

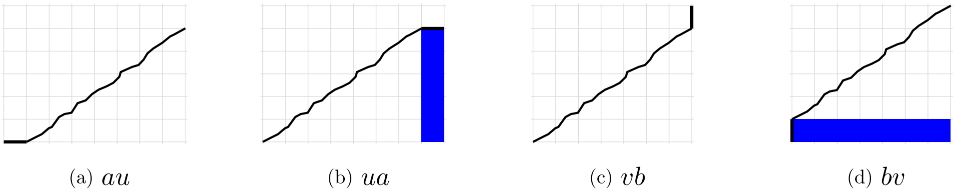

Fig. 8. Areas.

We give below a variant of this theorem of Chuan. The proof is also close to her's, but somewhat simpler. Note that although our result has a strong similarity with Remark 4.5 in [24], it is not the same result: indeed, the path associated with a word is not the same as ours.

Recall that a word on a totally ordered binary alphabet has a canonical representation by a discrete path, starting from the origin, in the discrete plane. We call *area* of the word the area located below the path and above the horizontal axis. For example, the area of the word *aabaabab* is 4, as is seen on Fig. 3.

Proposition 16.2. *The following conditions are equivalent, for a word w of length n on the totally ordered binary alphabet $\{a < b\}$:*

(i) *w is conjugate to a Christoffel word;*
(ii) *the set of areas of the conjugates of w is an interval of length n in $\mathbb{N}$;*
(iii) *the difference between the largest and the smallest area of the conjugates of w is $n - 1$.*

We use a simple lemma, whose proof is left to the reader; see Fig. 8.

Lemma 16.3. *Let the alphabet be $\{a < b\}$ and u, v be two words on this alphabet. Then $area(ua) = area(au) + |ua|_b$ and $area(vb) = area(bv) - |bv|_a$.*

Proof of the proposition. Let $p = |w|_a$ and $q = |w|_b$.

(i) implies (ii): it has been proved in Corollary 5.1 of [16] that for each element u in the set of conjugates of a Christoffel word, ordered alphabetically, if u is not the largest element, then $u = xaby$ and the next element in this set is $xbay$. Clearly, the area of $xbay$ is one more than the area of $xaby$. Hence (ii) follows.

(ii) implies (iii) is evident

(iii) implies (ii): the lemma implies that modulo $n = p + q$ we have $area(C(x)) \equiv area(x) + q$ for any conjugate x of w (C has been defined in Section 8). By assumption, there exist some conjugate x and some natural number s such that the difference between the areas of $C^s(x)$ and of x is $n - 1$. We deduce that $sq \equiv -1$, hence q is relatively prime to n. Since q generates the cyclic group $\mathbb{Z}/n\mathbb{Z}$, the areas of the words in the conjugation class are all distinct, because they are distinct modulo n. Thus, by the hypothesis of (iii), the set of areas is an interval of length n in $\mathbb{N}$.

(ii) implies (i): We may assume that w has the smallest area among all its conjugates. Define $f(x) = area(x) - area(w)$. Then the image of f is the interval $\{0, 1, \ldots, n - 1\}$ and f is injective.

If x is some conjugate beginning by a, then by the Lemma, $f(x)$ must lie in $\{0, \ldots, p - 1\}$, otherwise its conjugate $C(x)$ would have an image larger than $n - 1$. And if x begins by b, then $f(x)$ lies in $\{p, \ldots, n - 1\}$.

Write $w = a_0 \cdots a_{n-1}$. Then by the Lemma, $f(a_i \cdots a_n a_1 \cdots a_{i-1}) \equiv iq \mod n$, since $q \equiv -p \mod n$. If $a_i = a$, then by what we have seen above, $f(a_i \cdots a_n a_1 \cdots a_{i-1})$ lies in $\{0, \ldots, p - 1\}$, hence $iq \mod n < (i + 1)q \mod n$; if $a_i = b$, then similarly $iq \mod n > (i + 1)q \mod n$.

This proves that w is a lower Christoffel word, by the definition given in Section 3. □

By Chuan's theorem, the set of moments of the conjugates of a Christoffel word is an interval in $\mathbb{N}$. This may be generalized to factor sets, as follows.

Proposition 16.4. *The set of moments of a factor set is an interval in $\mathbb{N}$.*

Proof. Arguing as in the proof of "(i) implies (ii)" in Proposition 16.2, one obtains the result as a consequence of Theorem 4.2(ii). □

17. On a theorem of Droubay and Pirillo

In [28], Xavier Droubay and Giuseppe Pirillo have proved the following result.

Theorem 17.1. *A sequence is Sturmian if and only if for any n, it has 1 (resp. 2) palindromic factor of length n if n is even (resp. odd).*

We may prove the following finitary version of this result.

Proposition 17.2. *A word w is conjugate to a Christoffel word if and only if for any $n = 0, 1, \ldots, |w| - 1$, w has 1 (resp. 2) palindromic circular factors of length n if n is even (resp. odd).*

Lemma 17.3. *If for any $n = 0, 1, \ldots, |w| - 1$, w has 1 (resp. 2) palindromic circular factors of length n if n is even (resp. odd), then w is primitive.*

Proof. Suppose that w is not primitive. Then $w = u^l$ for some primitive word u and some integer $l \geq 2$. If $|u|$ is even, then by hypothesis, since u is a circular factor of w of length $< |w|$, we may assume that u is a palindrome up to conjugation of w; thus $u = v\tilde{v}$. Then u and $\tilde{v}v$ are both circular factors of w. By the hypothesis, they must be equal, so that $v = \tilde{v}$ and u is not primitive, contradiction.

Thus $|u|$ is odd. There are two palindromic circular factors of w of length $|u|$. They are both conjugate to u. Now, associate with u a regular $|u|$-gon, with vertices labelled with the letters of u, written clockwise (say). Since u has two palindromic conjugates, this figure has two axial symmetries; hence it is preserved by some rotation and u is not primitive, contradiction. □

The proof of the proposition is in two parts: the "only if" part is an easy consequence of the theorem of Droubay and Pirillo. The "if" part follows closely their proof, with some modifications.

Proof of the proposition. 1. Suppose that w is conjugate to a Christoffel word. It follows from [22], Theorem 6.4 (see also [16] Theorem 4.1), that for $n = 0, 1, \ldots, |w| - 1$, there are $n + 1$ distinct circular factors of length n of w. This holds also for the factors of length n of ww. Since ww is a factor of some Sturmian sequence, it follows that the set of circular factors of length n of w is a factor set. This implies by the theorem of Droubay and Pirillo that the number of palindromic circular factors of length n of w is 1 if n is even, and 2 if n is odd.

2. For the reverse implication, we follow and adapt step by step the clever proof by Droubay and Pirillo. Define the following properties of a word $w \in \{a, b\}^*$ and an integer n with $0 < n < |w|$:

- $B_n(w)$: for any circular factors u, v of length at most n of w, and any letter x, one has $||u|_x - |v|_x| \leq 1$.
- $S_n(w)$: there is a unique circular factor u of length $n - 1$ of w such that ua, ub are circular factors of w (in other words, u is right special with respect to the set of circular factors of w).
- $R_n(w)$: for any circular factor u of length n of w, $\tilde{u}$ is also a circular factor of w.

We prove by induction that for any $n \leq |w| - 1$, the three properties hold. First, it is easily verified that they hold for $n = 1$. We assume that they hold for some $n < |w| - 1$ and show that they hold also for $n + 1$.

$B_{n+1}(w)$: suppose that this property is false. Then, since $B_n(w)$ holds, there exists, by a result of Coven and Hedlund (Lemma 3.06 in [26]; see also [38] Proposition 2.1.3), a palindrome p such that apa, bpb are circular factors of length $n + 1$ of w. By hypothesis on w, we deduce that $n + 1$ is odd. Since p is a palindromic circular factor of w of length $n - 1$, there is another one, m say. There exists a letter x such that mx is a circular factor of w. Since mx is of length n, $R_n(w)$ implies that xm is a circular factor of w. Note that pa, pb are circular factors of w, so that by $S_n(w)$, ma, mb are not both circular factors of w. It follows that the unique right extension of xm as a circular factor of w is xmx: thus we have 3 palindromic circular factors of length $n + 1$ (namely apa, bpb, xmx, noting that $m \neq p$), contradicting the hypothesis. Thus $B_{n+1}(w)$ must hold.

$S_{n+1}(w)$: recall from [16] Lemma 4.1, that for any primitive word w, w has at least $n + 1$ circular factors of length n for any $n = 0, 1, \ldots, |w| - 1$. By the lemma, w is primitive. This implies that for any $n = 0, 1, \ldots, |w| - 2$, there exists at least one circular factor of w that is special. Suppose that property $S_{n+1}(w)$ does not hold. Since $n < |w|$, we deduce that there exists two distinct special circular factors of length n of w, u and v say. We may write $u = u'as$, $v = v'bs$, where s is the longest common suffix of u, v. Then u, v being special, $u'asa, v'bsb$ are circular factors of length $n + 1$ of w. Thus asa, bsb are circular factors of length $\leq n + 1$ of w, which contradicts property $B_{n+1}(w)$.

$R_{n+1}(w)$: since $S_{n+1}(w)$ holds, the number of circular factors of length $n + 1$ is $n + 2$. If $n + 1$ is even, the number of nonpalindromic circular factor of length $n + 1$ is $n + 1$, and if $n + 1$ is odd, it is n; in both cases, this number is even. So the number of circular factors of length $n + 1$ such that their reversal is not a circular factor, is even. If $R_{n+1}(w)$ does not hold, we must therefore have at least 2 words whose reversal is not a circular factor, u_1, u_2 say. Since $n + 1 \geq 2$, we may write $u_1 = xmx'$, $u_2 = ypy'$ for some letters x, x', y, y'. By $R_n(w)$, the word $x'\tilde{m}$ is a circular factor. Since $\tilde{u}_1 = x'\tilde{m}x$ is not a circular factor, the word $u_3 = x'\tilde{m}\hat{x}$ is a circular factor, where $\hat{a} = b$ and $\hat{b} = a$. Similarly $u_4 = y'\tilde{p}\hat{y}$ is a circular factor.

By $R_n(w)$ again, $\tilde{m}x$ is a circular factor. Thus $\tilde{m}$ is special. Similarly $\tilde{p}$ is special. From $S_n(w)$ it follows that they are equal. Thus $xmx' = u_1 \neq u_2 = ymy'$. If $x' = y'$, then $x \neq y$, that is $x = \hat{y}$ and we see that $u_4 = x'\tilde{m}x = \tilde{u}_1$, contradiction, since $\tilde{u}_1$ is not a circular factor.

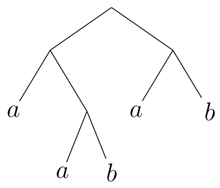

Fig. 9. The tree associated with *aabab*.

Thus we must have $x' \neq y'$, which means $y' = \hat{x}'$. We have 4 cases, and we indicate in each case two words among u_1, u_2, u_3, u_4 which contradict $B_{n+1}(w)$:

- $x = y$ and $x = x'$: $u_1 = xmx$, $u_4 = \hat{x}\tilde{m}\hat{x}$.
- $x = y$ and $x = \hat{x}'$: $u_2 = xmx$, $u_3 = \hat{x}\tilde{m}\hat{x}$.
- $x = \hat{y}$ and $x = x'$: $u_1 = xmx$, $u_2 = \hat{x}m\hat{x}$.
- $x = \hat{y}$ and $x = \hat{x}'$: $u_3 = \hat{x}\tilde{m}\hat{x}$, $u_4 = x\tilde{m}x$.

Thus, finally, there is a unique special circular factor for each $n = 0, 1, \ldots, |w| - 2$, which implies that the number of circular factors is $n + 1$, for each $n = 0, 1, \ldots, |w| - 1$. From this and Theorem 4.1 in [16], it follows that w is conjugate to a Christoffel word. □

18. Christoffel factors of Christoffel words

Proposition 18.1. *Let w be a lower Christoffel word on the alphabet $\{a < b\}$. Then each lower Christoffel word which is a circular factor of w is either equal to w or a factor of u or v, where $w = uv$ is the standard factorization.*

A lower Christoffel word w on the alphabet $\{a < b\}$ is in particular a Lyndon word. Hence the standard factorization applied iteratively associates with w a complete parenthesization, equivalently a complete binary tree whose leaves form the word w. The proposition says that each circular factor of w which is Christoffel word corresponds to a subtree. See Fig. 9; the factors are the letters and the words *aabab*, *aab*, *ab*.

Lemma 18.2. *Let m, w be Lyndon words such that m is a circular factor of w. Then m is a factor of w.*

Proof. Since m is a circular factor of w, it is a factor of ww. By contradiction, suppose that $m = xy$ for some nontrivial suffix x (resp. prefix y) of w. Then, since w is a Lyndon word, we have $w \leq x$. Thus $y \leq w \leq x < m$ (the last inequality is strict since $x \neq m$, because $y \neq 1$). But m is a Lyndon word, hence $m \leq y$, contradiction. □

Lemma 18.3. *Let m, w be Lyndon words such that m is a proper factor of w. Then m is a factor of either the longest proper prefix of w which is a Lyndon word, or of the longest proper suffix of w which is a Lyndon word.*

Proof. We show a little bit more: let $w = uv$ where u is the longest proper prefix of w which is a Lyndon word; then m is either a factor of u, or of v, or is a suffix of w. This will imply the result, since it is known that v is a Lyndon word (by [61] p. 15; see also [37] Exercise 5.1.6 or [59] Lemma 6).

Assume the contrary. Then m is not a suffix of w and m has the nontrivial factorization $m = m_1 m_2$ where m_1 is a nontrivial suffix of u and m_2 is a nontrivial prefix of v. By [59] Lemma 6 or [54] Lemma 3, the word $w^\flat$ obtained from w by suppressing its last letter has the periodic pattern u; since m is a factor of this word, m has the period $|u|$. But m is a Lyndon word, so it has no nontrivial period, hence we must have $|m| \leq |u|$. Hence $|m_2| < |u|$. Since $w = uv$ and $w^\flat$ has the periodic pattern u (and is a prefix of u^∞) and since m_2 is a prefix of v, we deduce that m_2 is a proper prefix of u. Thus

$$m_2 < u \leq m_1 < m,$$

the second inequality because u is a Lyndon word and m_1 a suffix of u. We deduce that $m_2 < m$, and since m_2 is a nontrivial suffix of m, m cannot be a Lyndon word. □

Recall that if w is a Lyndon word of length at least two, then its *right* (resp. *left*) *standard factorization* is $w = uv$ where v (resp. u) is the longest proper suffix (resp. prefix) of w that is a Lyndon word.

We use below a result of Guy Melançon [49] (see also [50] Theorem 3): following [8] Section 6.3, let us say that a Lyndon word w is *balanced*$_2$ if its left and right standard factorizations coincide, $w = uv$ say, and if moreover u, v are balanced$_2$; the result of Melançon is that a Lyndon word on the alphabet $\{a < b\}$ is balanced$_2$ if and only if it is a lower Christoffel word.

Proof of the proposition. Let m be a lower Christoffel word which is a circular factor of w.

By the first lemma, m is a factor of w. We know that w has the standard factorization $w = uv$ where, by the result of Guy Melançon above, u is the longest proper prefix of w which is a Lyndon word and where v is the longest proper suffix of w which is a Lyndon word. Thus, by the second lemma, m is a factor of either u or v. $\square$

Each proper (that is, of length at least 2) Lyndon word w has a factorization $w = uv$ where u, v are Lyndon words. It follows inductively that each Lyndon word of length n has at least $n-1$ occurrences of factors which are proper Lyndon words.

The next result shows that among Lyndon words, lower Christoffel words have the minimum number of Lyndon factors, when they are counted with their multiplicities.

Corollary 18.4. *A Lyndon word w on the alphabet $\{a < b\}$ of length $n \geq 1$ is a lower Christoffel word if and only if it has $n-1$ occurrences of factors which are proper Lyndon words.*

For example, the occurrences of factors, which are proper Lyndon words, of the Christoffel word *aabab* are the 4 occurrences: w itself, *aab* and *ab*, the latter having 2 occurrences. As counter-example, take *aabb*, which is not a Christoffel word: the occurrences of factors, which are Lyndon words of length ≥ 2, are the four words: *aabb*, *aab*, *abb*, *ab*.

Proof. The "only if" part follows inductively from Proposition 18.1: see the remark following it, using the fact that each Lyndon word which is a factor of a Christoffel word must be a Lyndon word [9], and that a complete binary tree with n leaves has $n-1$ internal nodes.

Now suppose that w is a Lyndon word on this alphabet which has $n-1$ occurrences of factors which are proper Lyndon words. We may assume that w is of length at least 3. Suppose that the left and right standard factorizations of w do not coincide. We then have $w = uv = u'v'$, where u, v, v, v' are Lyndon words and $u \neq u'$, $v \neq v'$. Note that $v' \neq u$ and $u' \neq v$ since w is unbordered.

Suppose that u' is of maximum length among these four words. Then, as observed above, u has at least $|u|-1$ (resp. $|v|-1$) occurrences of factors which are proper Lyndon words; note that u' does not appear in these multisets of factors (since it is too long and since $u' \neq v, u' \neq u$), and is a proper Lyndon word by the length hypothesis on w. Thus, counting also w itself, we have found $1 + 1 + |u| - 1 + |v| - 1$ occurrences of factors of w which are proper Lyndon words. This contradicts the hypothesis.

The proof when v' is of maximum length among the four words u, v, u', v' is similar, and the two other cases follow by symmetry.

Thus, the two standard factorizations of w coincide, $w = uv$ say. The previous argument also shows that u (resp. v) has $|u|-1$ (resp. $|v|-1$) occurrences of factors which are proper Lyndon words. By induction, u, v are lower Christoffel words, hence are balanced$_2$, by Melançon's result. Thus w is balanced$_2$ and by the same result, it is a lower Christoffel word. $\square$

The following result extends Lemma 9 in [59].

Corollary 18.5. *The Lyndon words that are factors of a given proper lower Christoffel word w are w, a, b and all the ancestors of w on the Christoffel tree. The number of distinct Lyndon words which are factors of w is 3 + the depth of w in the Christoffel tree.*

Note that, by [9], a Lyndon word which is a factor of a lower Christoffel word is necessarily a lower Christoffel word.

Proof. Let l be a Lyndon word which is proper factor of $w = uv$ (standard factorization). By Proposition 18.1 l must be a factor of u or v. Note that if $|w| \geq 3$, u is a prefix of v or v is a suffix of u: this follows for example from the construction of the Christoffel tree, see Section 5. Since the longest of both is the first ancestor of w in the Christoffel tree, the corollary follows. $\square$

Related results and conjectures are given by Kalle Saari [59].

19. On Markoff numbers

Markoff numbers have been introduced by Markoff in his two articles [45,46], in his study of minima of undefinite binary quadratic forms and approximations by continued fractions. See the recent book by Martin Aigner [1].

For purpose of simplicity, we take the following definition: a natural number m is a Markoff number if for some lower Christoffel word u on the alphabet $\{a < b\}$, $m = \mu(u)_{12}$, where μ is the homomorphism from the free monoid $\{a, b\}^*$ into $SL_2(\mathbf{Z})$, defined by $\mu a = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mu b = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$. See [14] Theorems 23 and 26, [8] Theorem 8.4, or [1] Theorem 4.13.

Recall that a word $w \in \{a, b\}^*$ is *antipalindromic* if it is equal to the reversal of the word obtained by exchanging a and b in w.

Proposition 19.1. *For each Markoff number m, there exists an antipalindromic word $w \in \{a,b\}^*$ such that m is equal to the length of the Christoffel word $aPal(w)b$.*

The point here is that w is antipalindromic, since Christoffel words of each length exist. As an example the Markoff number 13 is equal to the length of the Christoffel word $aPal(baba)b = ababbababbabb$.

Proof. Let $= \mu(u)_{12}$ as above. Note that $u \neq a$ (otherwise $m = 1$). If $u = b$, we take $w = 1$. We assume now that u is proper. Let α be the homomorphism from the free monoid $\{a,b\}^*$ into $SL_2(\mathbf{Z})$, defined by $\alpha a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\alpha b = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Then $\mu a = \alpha(ab)$ and $\mu b = \alpha(aabb)$. Thus $\mu = \alpha\phi$ where ϕ is the monoid endomorphism of $\{a,b\}^*$ sending a onto ab and b onto $aabb$. Let $\phi(u) = awabb$: w exists since $u \in a\{a,b\}^*b$, so that $\phi(u) \in ab\{a,b\}^*aabb$. One has $u = apb$ where p is a palindrome. Hence $\phi(u) = ab\phi(p)aabb$ and $w = b\phi(p)a$. Since $\phi(a)$ and $\phi(b)$ are antipalindromes and since p is a palindrome, $\phi(p)$ is an antipalindrome; hence so is w.

Now, let $\alpha(w) = \begin{pmatrix} h & i \\ j & k \end{pmatrix}$; then we have

$$\mu(u) = \alpha\phi(u) = \alpha(awabb) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} h & i \\ j & k \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} h+j & i+k \\ j & k \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} * & h+i+j+k \\ * & * \end{pmatrix}.$$

Thus $m = h + i + j + k$, which shows by [13] Corollary 3.2, that m is equal to the length of the Christoffel word $aPal(w)b$. □

The *Markoff injectivity conjecture* (Frobenius 1913, see the book [1]) asks, in one of its equivalent forms, if for each number m, the Christoffel words u, as at the beginning of this section, is unique.

It is shown in the previous proof that the word w is equal to $b\phi(p)a$ for some central word p, where the endomorphism ϕ is described in the proof. Thus, reversing the construction, the conjecture is equivalent to the question: is the mapping from the set of central words into the set of Markoff numbers, $p \mapsto 2 + |Pal(b\phi(p)a)|$, injective?

20. Bivariate counting

We consider below a, b as commuting variables. Denote by $Ce_n(a,b)$ the bivariate generating functions of the set of central Sturmian words of length n, and by $Ch_n(a,b)$ that of the lower Christoffel words of length n. It is clear that for $n \geq 2$,

$$Ch_n(a,b) = abCe_{n-2}(a,b).$$

Moreover $Ch_1(a,b) = a + b$ and for $n \geq 2$

$$Ch_n(a,b) = \sum a^i b^j,$$

where the sum is over all pairs i, j of relatively prime natural numbers of sum n. This follows indeed from the natural bijection between lower Christoffel words and this set of pairs, using for example the slope of the word, see Section 3. We have also the following result, where μ is the Möbius function.

Proposition 20.1. *One has $\sum_{n=de} Ch_d(a^e, b^e) = h_n(a,b)$ and consequently $Ch_n(a,b) = \sum_{n=de} \mu(d) h_e(a^d, b^d)$.*

Here h_n is the *homogeneous symmetric function*; that is $h_n(a,b) = \sum_{i+j=n} a^i b^j$, the sum of all monomials of total degree n.

Proof. The first identity expresses the fact that each point (i, j) in the first quadrant of the discrete plane, such that $i + j = n$, different from the origin (that is, $n \neq 1$), is of the form $e(i', j')$ for some unique point (i', j') whose coordinates are relatively prime and such that $i' + j'$ divides n. The second formula follows by Möbius inversion. □

The polynomial $Ch_n(q, 1) = \sum_{n=de} \mu(d) \frac{1-q^n}{1-q^d}$ is a classical q-analogue of the Euler function φ, see [60] page 46.

Gabriele Fici ([32] Theorem 2) has given a nice characterization of bispecial Sturmian words; he deduces from it that the number of these words is $2(n + 1) - \varphi(n + 2)$ ([32] Corollary 1). We use his characterization to prove the following result, which extends his cardinality result.

Proposition 20.2. *The bivariate commutative image of the set of bispecial Sturmian words of length n is equal to $2h_n(a,b) - Ce_n(a,b)$.*

1
1 1
1 2 1
1 3 3 1
1 4 4 4 1
1 5 6 6 5 1
1 6 8 6 8 6 1
1 7 11 8 8 11 7 1
1 8 13 12 8 12 13 8 1
1 9 17 13 12 12 13 17 9 1
1 10 20 16 16 10 16 16 20 10 1

Fig. 10. Bivariate counting of Sturmian words, following [3].

For example, we have for $n = 6$: $a^6 + 2a^5b + a^4b^2 + 2a^3b^3 + a^2b^4 + 2ab^5 + b^6 = 2a^6 + 2a^5b + 2a^4b^2 + 2a^3b^3 + 2a^2b^4 + 2ab^5 + 2b^6 - (a^6 + a^4b^2 + a^2b^4 + b^6) = 2h_6 - Ce_6(a, b)$.

Proof. The characterization of Fici is as follows: a Sturmian word w on the letters a, b is bispecial if and only if it is obtained from some power of some *proper* Christoffel word by removing its first and last letter (which are a and b, not necessarily in this order). For example, *aba* is obtained from the Christoffel word *aabab* (and also from *babaa*), and *ababaaba* is obtained from *aababaabab*, the square of the previous Christoffel word. Now there are two mutually excluding cases: (i) awb and bwa are Christoffel words, and in this case w is a central word; (ii) awb or bwa is a proper power of a Christoffel word, but not both.

There is a canonical surjective function from the set of powers of proper Christoffel words into the set of pairs of positive natural numbers, which associates with each such word u the pair $(|u|_a, |u|_b)$. Each pair is obtained twice, since there is the lower and the upper Christoffel word. Taking into account the two cases above, we deduce that the generating function of bispecial Sturmian words is:

$$2(1/ab)(h_{n+2}(a,b) - a^{n+2} - b^{n+2}) - Ce_n(a,b) = 2h_n(a,b) - Ce_n(a,b). \qquad \square$$

Denote by $\sigma(i, j)$ (resp. $t(i, j), \pi(i, j)$) the number of Sturmian (resp. of left special Sturmian, resp. of palindromic Sturmian) words on the alphabet $A = \{a, b\}$ having i (resp. j) occurrences of the letter a (resp. b). Note that all these functions are symmetric in i, j.

Basic results for bivariate enumeration of Sturmian words are due to Nicolas Bédaride, Éric Domenjoud, Damien Jamet and Jean-Luc Rémy.

They define the mapping θ (see Section 15) and use it to give a recursion formula for the function $\sigma(i, j)$: actually, they consider rather the function $s(L, h) = \sigma(L - h, h)$ for $L \geq h$, that is, the number of words of length L and b-degree h. Furthermore they give also a recursive formula for the function $p(L, h) = \pi(L - h, h)$.

Theorem 20.3. *(See [3].)*

(i) *Let w be a word in $A^* \setminus A^*bbA^*$. Then w is Sturmian if and only if $\theta(w)$ is Sturmian.*
(ii) *For all $L, h \in \mathbb{N}$ satisfying $0 \leq h \leq L/2$, one has: $s(L, h) = s(L - h - 1, h) + s(L - h, h) - s(L - 2h - 1, h) + s(h - 1, L - 2) + s(h - 1, L - 1)$.*
(iii) *$p(L, h) = 0$ if $L < 0$ or $L = 0$ and $h \neq 0$; $p(L, h) = 1$ if $L \geq 0$ and either $h = 0$ or $h = L$; $p(L, h) = p(L, h \mod L)$ if $L > 0$ and either $h < 0$ or $h > L$; $p(L, h) = p(L - h - 1, h) + p(h - 1, L - 1)$ otherwise.*

Here, the function $s(L, h)$ is extended to all integers by $s(L, h) = 0$ if $L < 0$; $s(0, h) = 0$ if $h \neq 0$; $s(L, h) = s(L, h \mod L)$ if $L > 0$. Note that, by symmetry, if $L/2 < h \leq L$, then $s(L, h) = s(L, L - h/2)$ and therefore the formula of Bédaride–Domenjoud–Jamet–Rémy allows to compute all $s(L, h)$. The numbers $\sigma(i, j)$ are shown in Fig. 10, up to $i + j = 10$.

Similarly, the function $p(L, h)$ is extended to all integers by $p(L, h) = 0$ if $L < 0$; $p(0, h) = 0$ if $h \neq 0$; $p(L, h) = p(L, h \mod L)$ if $L > 0$. The numbers $\pi(i, j)$ are shown in Fig. 12.

It can be shown that θ preserves Christoffel words, is compatible with normal forms, with palindromes and with special words. We shall not insist on that, but it may simplify the next proof.

We say that the letter x is *isolated* in the word w if x^2 is not a factor of w.

Proposition 20.4. *For $x \in \{a, b\}$, let $t_{a,a}(i, j)$ (resp. $t_{a,b}(i, j)$) denote the number of left special Sturmian words which begin by a and end by a (resp. and end by b), in which b is isolated and such that $|w|_a = i$, $|w|_b = j$. Then $t(0, 0) = t(1, 0) = t(0, 1) = 1$; $t(i, j) = t(j, i)$; $t(i, i) = 2$; $t(i, j) = t_{a,a}(i, j) + t_{a,b}(i, j)$ if $i > j$; $t_{a,a}(i, j) = t(i - j - 1, j)$ if $i > j$ and $= 0$ otherwise; $t_{a,b}(i, j) = t_{a,b}(i - j, j) + t_{a,a}(j, i - j)$ if $i \geq j$ and $= 0$ otherwise.*

It should be clear that these formulas allow to compute the values $t(i, j)$. They were used to compute these numbers up to $i + j = 10$ in Fig. 11.

1
1 1
1 2 1
1 2 2 1
1 3 2 3 1
1 3 2 2 3 1
1 4 3 2 3 4 1
1 4 3 3 3 3 4 1
1 5 3 4 2 4 3 5 1
1 5 4 3 3 3 3 4 5 1
1 6 4 4 5 2 5 4 4 6 1

Fig. 11. Bivariate counting of left special Sturmian words.

Proof. Clearly $t(0,0) = t(1,0) = t(0,1) = 1$ and by symmetry, $t(i,j) = t(j,i)$. We see that $t(i,i) = 2$, since it corresponds to the two words $(ab)^i$ and $(ba)^i$.

If w is a Sturmian word such that $|w|_a = i$, $|w|_b = j$ with $i > j$, then b must be isolated (since a^2 and b^2 are not simultaneously factor of a Sturmian word); moreover, if w is left special, then bw is Sturmian, and therefore w does not begin by b (by a similar argument). Thus we have $t(i,j) = t_{a,a}(i,j) + t_{a,b}(i,j)$.

Suppose that w is a left special Sturmian word that begins and ends by a, such that $|w|_a = i$, $|w|_b = j$ and in which b is isolated. Then we must have $i > j$. Note that $\theta(w)$ is Sturmian by Theorem 20.3 and one has $|\theta(w)|_a = i - j - 1$ and $|\theta(w)|_b = j$; it is left special, by applying the same theorem to the words $\theta(aw) = a\theta(w)$ (this is true since w begins by a) and $\theta(bw) = b\theta(w)$. Moreover, let u be any left special Sturmian word with $|u|_a = i - j - 1$ and $|u|_b = j$. Then there is a unique word w such that $\theta(w) = u$, $|w|_a = i$, $|w|_b = j$; w is Sturmian and similarly left special. We deduce that $t_{a,a}(i,j) = t(i - j - 1, j)$ if $i > j$.

Suppose now that w is a left special Sturmian word beginning by a and ending by b, such that $|w|_a = i$, $|w|_b = j$ and in which b is isolated. Then we must have $i \geq j$. If we apply θ to this word, we find a Sturmian word u which is left special (same argument as previously), which ends by b, but begins either by a or by b, and such that $|u|_a = i - j$ and $|u|_b = j$; note that in the first case, w begins by a^2 and aw is Sturmian since w is left special, hence the a-runs in w are of length at least 2, so that in $u = \theta(w)$, b is isolated; in the second case, w begins by ab, and $aw = aab\cdots$ and $bw = bab\cdots$ are Sturmian, hence the a-runs in w are of length 1 and 2, so that in u, a is isolated. Conversely, if u is a left special Sturmian word with $|u|_a = i - j$ and $|u|_b = j$, ending by b, such that either u begins by a and b is isolated, or u begins by b and a is isolated, there is a unique w as above. It follows that we have $t_{a,b}(i,j) = t_{a,b}(i - j, j) + t_{a,a}(j, i - j)$, since the u's in the second case are counted by $t_{a,a}(j, i - j)$ (by symmetry). □

The set of all conjugates of all powers of Christoffel words (also called α-words by Chuan) is a *cyclic language*. Recall that, according to [12], a language (that is, a set of words) is cyclic if: (i) for any words u, v, one has $uv \in L$ if and only if $vu \in L$; and (ii) for any word w and any $n \geq 1$, $w \in L$ if and only $w^n \in L$. The *zeta function* of a cyclic language is the series $\sum_{n>0} exp(a_n x^n/n)$ where a_n is the number of words of length n in the language. If α_n is the number of Lyndon words of length n in the language, assumed to be cyclic, then the zeta function has the infinite product expansion $\prod_{n>0} 1/(1 - x^n)^{\alpha_n}$, which shows that its coefficients are natural numbers, [12] Proposition 1. Thus, since the Lyndon words among conjugates of powers of Christoffel words are exactly the lower Christoffel words, it follows that the zeta function of this language is

$$\frac{1}{1-x} \prod_{n>0} 1/(1 - x^n)^{\varphi(n)}$$

since there are $\varphi(n)$ lower Christoffel words of length n for $n > 1$ and there are 2 Christoffel words of length 1. Equivalently, the number of conjugates of powers of Christoffel words of length n is $\sum_{d|n} d\varphi'(d)$, where φ' coincides with φ except that $\varphi'(1) = 2$.

21. Numerology

The world of Sturmian words contains some interesting numbers. In Fig. 10 are shown the numbers $\sigma(i,j)$, with each row corresponding to $i + j$ equal $0, 1, \ldots, 10$. Note that the row sums count the number of Sturmian words of a given length; for $n = 0, \ldots, 10$, these numbers are (thanks to the Lipatov–Mignosi formula $1 + \sum_{i+j=n}(j+1)\varphi(i)$): $1, 2, 4, 8, 14, 24, 36, 54, 76, 104, 136$.

In Fig. 11 are shown the numbers $t(i,j)$, which count the left (or right) special Sturmian words. The row sums are the number of left special words of a given length, which are, for $n = 0, \ldots, 10$: $1, 2, 4, 6, 10, 12, 18, 22, 28, 32, 42$, according to the de Luca–Mignosi formula $\varphi(1) + \cdots + \varphi(n+1)$, [43] Theorem 6.

In Fig. 12 are shown the numbers $p(i,j)$, which count the Sturmian palindromes. The row sums are the number of Sturmian palindromes of a given length, which are, for $n = 0, \ldots, 10$: $1, 2, 2, 4, 4, 8, 6, 14, 10, 20, 14$ according to the de Luca–De Luca formula $1 + \sum_{n-2i>0} \varphi(n - 2i)$, [42] Theorem 11.

1
1 1
1 0 1
1 1 1 1
1 0 2 0 1
1 1 2 2 1 1
1 0 2 0 2 0 1
1 1 3 2 2 3 1 1
1 0 3 0 2 0 3 0 1
1 1 3 3 2 2 3 3 1 1
1 0 4 0 2 0 2 0 4 0 1

Fig. 12. Bivariate counting of Sturmian palindromes following [3].

22. Open problems

1. There is one item missing in Proposition 11.2, namely a characterization using the normal form of words which are conjugate to some Christoffel word, without being a Christoffel word.
2. A problem related to the previous one is to characterize the minimal periods of the conjugates of a Christoffel word w; compare Proposition 8.2 and the remark following its proof. It may be that it is always the length of some Christoffel factor of w.
3. Find a combinatorial description of the function which associates with each Sturmian word w the shortest right special Sturmian word having w as prefix (one may think of the palindromic closure of de Luca: it associates a central word with each left special word). Describe the infinite binary tree associated as in Section 15 to the S-poset of all Sturmian words; in other words, determine the function which with a right special Sturmian word u associates the shortest right special Sturmian word which has ua (resp. ub) as prefix.
4. There is a generalization of Christoffel words to larger alphabets, called Christoffel–Lyndon words, see [50]. Does Proposition 18.4 also characterize Christoffel–Lyndon words?
5. As suggested on the figures, it seems that on each row in Fig. 10 and Fig. 12, the maximum is attained for the third entry, whereas in Fig. 11, it is the second one.
6. Having in mind the Lipatov–Mignosi formula (see Corollary 12.2), an interesting problem is to give some similar closed formula for the bivariate symmetric function which counts the Sturmian words of given length. The same problem is open for left Sturmian words and for Sturmian palindromes.
   Remark: recall that Schur positivity is a basic property of symmetric function: the fact that a symmetric function is Schur-positive means that it encodes a representation of the symmetric group and of the general linear group. It may be of interest for the interested reader to know that the bivariate symmetric function counting finite Sturmian words of a given length is not Schur-positive: indeed, for $n = 6$, it is equal to $s_6 + 5s_{51} + 2s_{42} - 2s_{33}$, with the Macdonald notations (see [44] for this).
7. It was shown by [3] that the series $\sum_L s(L,h)x^L$ is rational, and even has for $h \geq 2$ the remarkable denominator $(1-x^{L-1})(1-x^L)(1-x^{L+1})$. They give a similar result for the function $p(L,h)$. One may ask the following question for the functions $\sigma(i,j)$, $t(i,j)$ and $\pi(i,j)$: let i, j, u, v be natural numbers; show that the series $\sum_{n\geq 0} t(i+nu, j+nv)x^n$ are rational and find a denominator; similarly for $\sigma(i,j)$ and $\pi(i,j)$.

Acknowledgements

Many thanks to Sébastien Labbé, Damien Jamet, Dominique Perrin, Christian Kassel, François Bergeron and Srećko Brlek for useful discussions. Thanks to the anonymous referee, who indicated a lot of corrections in the first version.

References

[1] M. Aigner, Markov's Theorem and 100 Years of the Uniqueness Conjecture, Springer-Verlag, 2013.
[2] J.-P. Allouche, J. Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, 2003.
[3] N. Bédaride, É. Domenjoud, D. Jamet, J.-L. Rémy, On the number of balanced words of given length and height over a two-letter alphabet, Discrete Math. Theor. Comput. Sci. 12 (3) (2010) 41–62.
[4] C.A. Berenstein, L.A. Kanal, D. Lavine, E.C. Olson, A geometric approach to subpixel registration accuracy, Comput. Vis. Graph. Image Process. 40 (1987) 334–360.
[5] J. Berstel, Mots de Fibonacci, Séminaire d'Informatique théorique, Institut de Programmation, Paris, 1980.
[6] J. Berstel, Tracé de droites, fractions continues et distances discrètes, in: M. Lothaire (Ed.), Mots, Mélanges offerts à Schützenberger, Hermès, Paris, 1990, pp. 298–309.
[7] J. Berstel, Sturmian and episturmian words (a survey of some recent results), in: Algebraic Informatics, in: Lecture Notes in Comput. Sci., vol. 4728, Springer, Berlin, 2007, pp. 23–47.
[8] J. Berstel, A. Lauve, C. Reutenauer, F. Saliola, Combinatorics on Words: Christoffel Words and Repetitions in Words, CRM Monograph Series, American Mathematical Society, 2008.
[9] J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1997) 171–203.
[10] J. Berstel, C. De Felice, D. Perrin, C. Reutenauer, G. Rindone, Bifix codes and Sturmian words, J. Algebra 369 (2012) 146–202.
[11] J.-P. Berstel, M. Pocchiola, Random generation of finite Sturmian words, Discrete Math. 153 (1996) 29–39.

[12] J.-P. Berstel, C. Reutenauer, Zeta functions of formal languages, Trans. Amer. Math. Soc. 321 (1990) 533–546.
[13] V. Berthé, A. de Luca, C. Reutenauer, On an involution of Christoffel words and Sturmian morphisms, European J. Combin. 29 (2008) 535–553.
[14] E. Bombieri, Continued fractions and the Markoff tree, Expo. Math. 25 (2007) 187–213.
[15] J.-P. Borel, F. Laubie, Quelques mots sur la droite projective réelle, J. Théor. Nombres Bordeaux 5 (1993) 23–51.
[16] J.-P. Borel, C. Reutenauer, On Christoffel classes, RAIRO Theor. Inform. Appl. 40 (2006) 15–27.
[17] M. Boshernitzan, A. Fraenkel, A linear algorithm for nonhomogeneous spectra of numbers, J. Algorithms 5 (1984) 187–198.
[18] S. Brlek, J.-O. Lachaud, X. Provencal, C. Reutenauer, Lyndon+Christoffel = digitally convex, Pattern Recognit. 42 (2009) 2239–2246.
[19] N.G. de Bruijn, Updown generation of Beatty sequences, Indag. Math. 43 (1989) 385–407.
[20] A. Carpi, A. de Luca, Codes of central Sturmian words, Theoret. Comput. Sci. 340 (2005) 220–239.
[21] E.B. Christoffel, Observatio arithmetica, Ann. Mat. 6 (1875) 145–152.
[22] W.-F. Chuan, α-Words and factors of characteristic sequences, Discrete Math. 177 (1997) 33–50.
[23] W.-F. Chuan, Unbordered factors of the characteristic sequences of irrational numbers, Theoret. Comput. Sci. 205 (1998) 337–344.
[24] W.-F. Chuan, Characterizations of α-words, moments and determinants, Fibonacci Quart. 41 (2003) 194–208.
[25] W.-F. Chuan, Moments of conjugacy classes of binary words, Theoret. Comput. Sci. 310 (2004) 273–285.
[26] E.M. Coven, G.A. Hedlund, Sequences with minimal block growth, Math. Syst. Theory 7 (1973) 138–153.
[27] T.W. Cusick, M.E. Flahive, The Markoff and Lagrange Spectra, American Mathematical Society, 1989.
[28] X. Droubay, G. Pirillo, Palindromes and Sturmian words, Theoret. Comput. Sci. 223 (1999) 73–85.
[29] S. Dulucq, D. Gouyou-Beauchamps, Sur les facteurs des suites de Sturm, Theoret. Comput. Sci. 71 (1990) 381–400.
[30] J.-P. Duval, Relationship between the period of a finite word and the length of its unbordered segments, Discrete Math. 40 (1982) 31–44.
[31] J.-P. Duval, Factorizing words over an ordered alphabet, J. Algorithms 4 (1983) 363–381.
[32] G. Fici, On the structure of bispecial Sturmian words, J. Comput. System Sci. 80 (2014) 711–719.
[33] R.L. Graham, D. Knuth, O. Patashnik, Concrete Mathematics, 2nd edition, Addison Wesley, 1994.
[34] T. Harju, D. Nowotka, Minimal Duval extensions, Internat. J. Found. Comput. Sci. 15 (2004) 349–354.
[35] J. Justin, Episturmian morphisms and a Galois theorem on continued fractions, Theor. Inform. Appl. 39 (1) (2005) 207–215.
[36] E.P. Lipatov, On a classification of binary collections and properties of homogeneity classes (in Russian), Problemy Kibernet. 39 (1982) 67–84.
[37] M. Lothaire, Combinatorics on Words, Addison Wesley, 1983, 2nd ed., Cambridge University Press, 1997.
[38] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, 2002.
[39] A. de Luca, A combinatorial property of the Fibonacci words, Inform. Process. Lett. 12 (1981) 193–195.
[40] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics, Theoret. Comput. Sci. 183 (1997) 45–82.
[41] A. de Luca, A. De Luca, Some characterizations of finite Sturmian words, Theoret. Comput. Sci. 356 (2006) 118–125.
[42] A. de Luca, A. De Luca, Combinatorial properties of Sturmian palindromes, Internat. J. Found. Comput. Sci. 17 (2006) 557–573.
[43] A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, Theoret. Comput. Sci. 136 (1994) 361–385.
[44] I.G. Macdonald, Symmetric Functions and Hall Polynomials, Oxford University Press, 1979, 2nd ed., 1995.
[45] A.A. Markoff, Sur les formes quadratiques binaires indéfinies, Math. Ann. 15 (1879) 381–496.
[46] A.A. Markoff, Sur les formes quadratiques binaires indéfinies (second mémoire), Math. Ann. 17 (1880) 379–399.
[47] F. Mignosi, Infinite words with linear subword complexity, Theoret. Comput. Sci. 65 (1989) 221–242.
[48] F. Mignosi, On the number of Sturmian factors, Theoret. Comput. Sci. 82 (1991) 71–84.
[49] G. Melançon, Visualisation de graphes et Combinatoire des mots, Habilitation à diriger des recherches, Université de Bordeaux 1, 1999.
[50] G. Melançon, C. Reutenauer, On a class of Lyndon words extending Christoffel words and related to a multidimensional continued fraction algorithm, J. Integer Seq. 16 (9) (2013) 13.9.7, 30 pp.
[51] F. Mignosi, L.Q. Zamboni, A note on a conjecture of Duval and Sturmian words, Theor. Inform. Appl. 36 (2002) 1–3.
[52] M. Morse, G.A. Hedlund, Symbolic dynamics II: Sturmian trajectories, Amer. J. Math. 62 (1940) 1–42.
[53] D. Perrin, A. Restivo, A note on Sturmian words, Theoret. Comput. Sci. 429 (2012) 265–272.
[54] D. Perrin, A. Restivo, Enumerative combinatorics on words, in: Handbook of Enumerative Combinatorics, Chapman and Hall/CRC, 2015.
[55] N. Pytheas Fogg, Substitutions in dynamics, arithmetics and combinatorics, in: V. Berthé, S. Ferenczi, C. Mauduit, A. Siegel (Eds.), Lecture Notes in Mathematics, vol. 1794, Springer-Verlag, Berlin, 2002.
[56] C. Reutenauer, Free Lie Algebras, Oxford University Press, 1993.
[57] C. Reutenauer, On Markoff's property and Sturmian words, Math. Ann. 336 (2006) 1–12.
[58] G. Richomme, P. Séébold, On factorially balanced sets of words, Theoret. Comput. Sci. 412 (2011) 5492–5497.
[59] K. Saari, Lyndon words and Fibonacci numbers, J. Combin. Theory Ser. A 121 (2014) 34–44.
[60] J. Sándor, B. Crstici, Handbook of Number Theory II, Kluwer Academic Publishers, 2004.
[61] G. Viennot, Algèbres de Lie libres et monoïdes libres, Lecture Notes in Mathematics, vol. 691, Springer-Verlag, Berlin, 1978.
[62] Z.X. Wen, Z.Y. Wen, Some properties of the singular words of the Fibonacci word, European J. Combin. 15 (1994) 587–598.