Some results on digital segments and balanced words ☆

Alessandro De Luca *, Gabriele Fici

A R T I C L E   I N F O

A B S T R A C T

We exhibit combinatorial results on Christoffel words and binary balanced words that are motivated by their geometric interpretation as approximations of digital segments. We give a closed formula for counting the exact number of balanced words with a zeroes and b ones. We also study minimal non-balanced words.

1. Introduction

The study of digital approximations of lines in the plane by discrete paths in the grid $\mathbb{Z}^2$ is a classical topic at the frontier between digital geometry, arithmetic, and combinatorics on words. For example, it is well known that digital (infinite) rays with irrational slope are encoded by infinite *Sturmian words*. A word is a sequence of letters drawn from a finite alphabet. A (right) infinite word is Sturmian if it has $n+1$ distinct factors (contiguous sub-blocks) of length n for every $n \geq 0$. Since the famous theorem of Morse and Hedlund [23] states that a right infinite word is aperiodic if and only if it has *at least* $n+1$ distinct factors of length n for every $n \geq 0$, Sturmian words are aperiodic binary words of minimal factor complexity (see [20, Chap. 2] for more details).

A geometric interpretation of Sturmian words can be given in terms of *cutting sequences*. Consider rays of the form $y = \alpha x + \rho$ with irrational slope $\alpha > 0$ and real intercept $\rho \geq 0$. The cutting sequences of such lines are defined by labeling the intersections of $y = \alpha x + \rho$ with the (first quadrant of the) integer grid, using 0 if the grid line crossed is vertical and 1 if it is horizontal (and either 01 or 10 if the ray passes through a point of $\mathbb{Z}^2$). The sequence of labels, read from $(0, \rho)$ out, is called the cutting sequence of $y = \alpha x + \rho$ and is a Sturmian word.

Among many equivalent definitions, Sturmian words are also characterized by the *balance property*. A (finite or infinite) binary word is balanced if in every two factors of the same length, the difference between occurrences of each letter is bounded by 1. Sturmian words are precisely aperiodic balanced binary words [24].

The case of finite segments is also source of a large number of interesting correspondences between digital geometry, arithmetic and combinatorics. For a general overview, see [17]. In particular, finite factors of Sturmian words are precisely the binary balanced words. Among these, a special role is played by *Christoffel words*. The literature about Christoffel words is very rich (see for example [27,20,6]). A lower (resp., upper) Christoffel word is a binary word that encodes the digital approximation from below (resp., from above) of a Euclidean segment, where 0 (resp., 1) represents a horizontal (resp., vertical) unit segment (see Fig. 1).

For every pair of positive integers (a, b) there is one lower (resp. upper) Christoffel word having a occurrences of 0 and b occurrences of 1. These two words define the contour (lower and upper, respectively) of a region enclosing the Euclidean segment from $(0, 0)$ to (a, b) (see Fig. 2).

---

* Corresponding author.

E-mail addresses: alessandro.deluca@unina.it (A. De Luca), gabriele.fici@unipa.it (G. Fici).
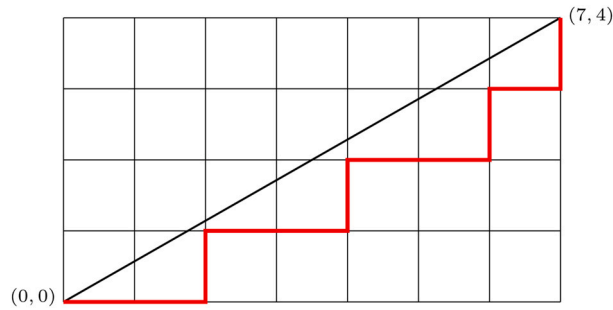
Fig. 1. The lower Christoffel word $w_{7,4} = 00100100101$. The upper Christoffel word $W_{7,4} = 10100100100$ is the reversal of $w_{7,4}$.
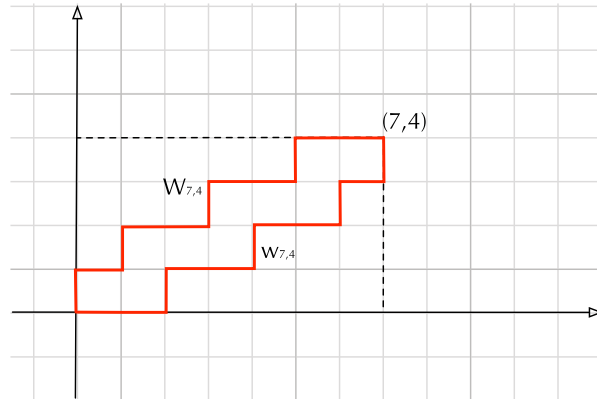

Fig. 2. The region $R(7,4)$ of the discrete plane whose contour is delimited by the lower Christoffel word $w_{7,4}$ and the upper Christoffel word $W_{7,4}$ is precisely the region of points in the grid that have Euclidean distance smaller than $\sqrt{2}$ from the Euclidean segment joining $(0,0)$ to $(7,4)$.

We give a purely combinatorial proof of the fact that *every* balanced word with a occurrences of 0 and b occurrences of 1 encodes a path that is contained in this region.

So, given positive a and b, the set of balanced words with a occurrences of 0 and b occurrences of 1 is in general a proper subset of the set of binary words encoding a path that lies at distance bounded by $\sqrt{2}$ from the Euclidean segment joining the origin to the point (a,b). A natural question that therefore arises is that of counting, for every pair of positive integers (a,b), the exact number of balanced words with a occurrences of 0 and b occurrences of 1. While a recursive formula for this problem was known [1], in this paper we present an exact closed formula.

We also study the set of minimal non-balanced words and the set of minimal almost balanced words, providing new combinatorial characterizations for these classes of words already studied in [26].

2. Preliminaries

A *word* over a finite alphabet Σ is a concatenation of letters from Σ. The *length* of a word w is denoted by $|w|$. The empty word ε has length 0. The set of all words over Σ is denoted Σ^* and is a free monoid with respect to concatenation. For a letter $x \in \Sigma$, $|w|_x$ denotes the number of occurrences of x in w. If $\Sigma = \{x_1, \ldots, x_\sigma\}$ is ordered, the vector $(|w|_{x_1}, \ldots, |w|_{x_\sigma})$ is the *Parikh vector* of w.

Let $w = uv$, with $u, v \in \Sigma^*$. We say that u is a *prefix* of w and that v is a *suffix* of w. A *factor* of w is a suffix of a suffix (or, equivalently, a suffix of a prefix). If $w = w_1 w_2 \cdots w_n$, with $w_k \in \Sigma$ for all k, we let $w[i..j]$ denote the nonempty factor $w_i w_{i+1} \cdots w_j$ of w, whenever $1 \le i \le j \le n$. A factor u of a word $w \ne u$ is a *border* of w if u is both a prefix and a suffix of w; in this case, w has *period* $|w| - |u|$. A word w is *unbordered* if its longest border is ε; i.e., if its smallest period is $|w|$. The prefix ρ_w whose length equals the smallest period of a nonempty w is called its *fractional root*. For a word w, the *n-th power* of w is the word w^n obtained by concatenating n copies of w; furthermore, we let w^ω denote the infinite periodic word $www\cdots$.

Two words w and w' are *conjugate* if $w = uv$ and $w' = vu$ for some words u and v. The conjugacy class of a word w contains $|w|$ elements if and only if w is *primitive*, i.e., $w = v^n$ for some v implies $n = 1$.

A nonempty word $w = w_1 w_2 \cdots w_n$, $w_k \in \Sigma$ for all k, is a *palindrome* if it coincides with its *reversal* $w^R = w_n w_{n-1} \cdots w_1$. The empty word is also assumed to be a palindrome.

The following proposition, whose proof is straightforward, is well known (cf. [11,8]).

Proposition 1. *A word is a conjugate of its reversal if and only if it is a concatenation of two palindromes. Moreover, these two palindromes are uniquely determined if and only if the word is primitive.*

2.1. Christoffel words

We now define Christoffel words and their properties. We point the reader to the classical references [4,28,20,2] for more details on Christoffel words.

In what follows, we fix the alphabet $\{0,1\}$ and represent a word over $\{0,1\}$ as a path in $\mathbb{N} \times \mathbb{N}$ starting at $(0,0)$, where 0 (resp., 1) encodes a horizontal (resp., vertical) unit segment. For every pair of nonnegative integers (a,b) (not both 0), every word w that encodes a path from $(0,0)$ to (a,b) must have exactly a zeroes and b ones, i.e., it has Parikh vector $(a,b) = (|w|_0, |w|_1)$. We also say that $b = |w|_1$ is the *height* of w. We define the *slope* of a word w with Parikh vector (a,b) as the rational number b/a if $a \neq 0$, or ∞ otherwise. We remark here that some authors (see, e.g., [20]) call slope the ratio $b/(a+b)$ between the height and the length of the word; in this paper we will refer to this quantity as the *mechanical slope*.

Definition 2. Given a pair of nonnegative integers (a,b) (not both 0), the *lower (resp., upper) Christoffel word* $w_{a,b}$ (resp., $W_{a,b}$) is the word encoding the path from $(0,0)$ to (a,b) that is closest from below (resp., from above) to the Euclidean segment, without crossing it.

In other words, $w_{a,b}$ (resp. $W_{a,b}$) is the digital approximation from below (resp., from above) of the Euclidean segment joining $(0,0)$ to (a,b). For example, the lower Christoffel word $w_{7,4}$ is the word 00100100101 (see Fig. 1).

By construction, we have that the upper Christoffel word $W_{a,b}$ is the reversal of the lower Christoffel word $w_{a,b}$ (and the two words are conjugates, see below).

Notice that by definition we allow a or b (but not both) to be 0, that is, we have $w_{n,0} = W_{n,0} = 0^n$ and $w_{0,n} = W_{0,n} = 1^n$; the words of length 1 are both lower and upper Christoffel words.

If a and b are coprime, the Christoffel words $w_{a,b}$ and $W_{a,b}$ do not intersect the Euclidean segment joining $(0,0)$ to (a,b) (other than at the end points) and are primitive words. If instead $a = n\alpha$ and $b = n\beta$ for some $n > 1$, then $w_{a,b} = (w_{\alpha,\beta})^n$ (resp., $W_{a,b} = (W_{\alpha,\beta})^n$). Hence, $w_{a,b}$ (resp., $W_{a,b}$) is primitive if and only if a and b are coprime.

Therefore, for every $n > 1$, there are $2\phi(n)$ primitive Christoffel words of length n (in particular, there are $\phi(n)$ primitive lower Christoffel words and $\phi(n)$ primitive upper Christoffel words), where ϕ is the Euler totient function.

Remark 3. From the definition, it follows that the region $R(a,b)$ of the discrete plane whose contour is delimited by the lower and the upper Christoffel word with Parikh vector (a,b) is precisely the region of points in the grid that have Euclidean distance smaller than $\sqrt{2}$ from the Euclidean segment joining $(0,0)$ to (a,b), see Fig. 2. Notice that by definition, this region does not contain internal points with integer coordinates.

Consider the sequence $(k \cdot b \bmod (a+b))$, for $k = 1, \ldots, a+b$. Each subsequent number in the sequence is obtained either by adding b (supposing to start from 0) or by subtracting a. Writing 0 whenever we add b and 1 when we subtract a we get the lower Christoffel word $w_{a,b}$ (as is well known and easy to see from the definition). If we divide each term in the sequence by a, we get the sequence of vertical distances between the endpoints of paths encoded by prefixes of $w_{a,b}$ and the Euclidean segment joining $(0,0)$ to (a,b); if instead we divide by $\sqrt{a^2 + b^2}$ (the length of the Euclidean segment) we get the sequence of Euclidean distances. In particular, the point with integer coordinates that is closest to the Euclidean segment (without lying on the segment itself) is at distance $1/\sqrt{a^2 + b^2}$, whereas the point on the path representing $w_{a,b}$ that is *farthest* from the segment is always at distance $(a+b-1)/\sqrt{a^2 + b^2}$. These two points play a crucial role in the combinatorics of Christoffel words (see Sec. 2.2).

For example, for the path $w_{7,4}$ depicted in Fig. 1, the sequence $(4k \bmod 11)$ is

$$4, 8, 1, 5, 9, 2, 6, 10, 3, 7, 0.$$

Taking for instance $k = 5$, we get that the point $(4,1)$, which is the endpoint of the path corresponding to the prefix 00100 of $w_{7,4}$, has vertical distance $9/7 \approx 1.286$ and Euclidean distance $9/\sqrt{65} \approx 1.116$ from the Euclidean segment joining $(0,0)$ and $(7,4)$.

We now give a fundamental definition for the combinatorics of Christoffel words.

Definition 4. A *central word* is a word that has coprime periods p and q and length $p + q - 2$.

It is well known that central words are binary palindromes [14,20]. The following theorem gives a structural characterization (see [12]).

Theorem 5 (*Combinatorial Structure of Central Words*). *A word w is central if and only if it is a power of a letter or there exist palindromes P and Q such that $w = P01Q = Q10P$. Moreover,*

- *P and Q are central words;*
- *$|P| + 2$ and $|Q| + 2$ are coprime and w has periods $|P| + 2$ and $|Q| + 2$;*
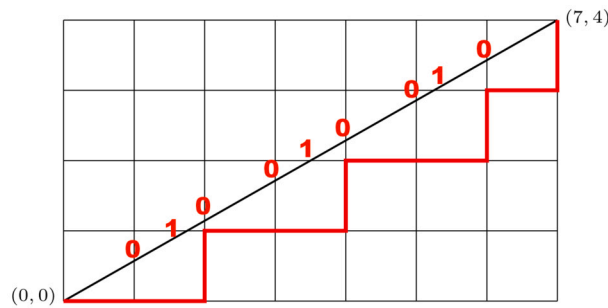- *if $|P| < |Q|$, Q is the longest palindromic (proper) suffix of w.*

Fig. 3. The central word $C = 010010010$ is the central factor of the primitive lower Christoffel word $w_{7,4} = 0C1$. It encodes the intersections of the Euclidean segment joining $(0,0)$ and $(7,4)$ with the grid (0 for a vertical intersection and 1 for a horizontal intersection).

Example 6. The word $w = 010010$ is a central word, since it has periods 3 and 5 and length $6 = 3 + 5 - 2$. We have $010010 = 010 \cdot 01 \cdot 0 = 0 \cdot 10 \cdot 010$, so that $P = 010$ and $Q = 0$.

Proposition 7 ([25]). *A word C is central if and only if the words $0C1$ and $1C0$ are conjugates.*

Proposition 8 ([3]). *A word is a primitive lower (resp., upper) Christoffel word if and only if it has length 1 or it is of the form $0C1$ (resp., $1C0$) where C is a central word.*

Example 9. Let $a = 7$ and $b = 4$. We have $w_{7,4} = 00100100101 = 0 \cdot 010010010 \cdot 1$, where $C = 010010010$ is a central word since it has periods 3 and 8 and length $9 = 3 + 8 - 2$; see Fig. 3.

So, central words are the "central" factors of primitive Christoffel words of length ≥ 2. A geometric interpretation of the central word C is the following: it encodes the intersections of the Euclidean segment joining $(0,0)$ to (a,b) (0 for a vertical intersection and 1 for a horizontal intersection). In other words, the word C is the cutting sequence of the Euclidean segment joining $(0,0)$ to (a,b); see again Fig. 3.

Proposition 10 ([5]). *Let $w_{a,b} = 0C1$ be a primitive lower Christoffel word. The central word C has periods a' and b', the multiplicative inverses of a and b modulo $a + b$, respectively.*

Definition 11. A word over $\{0,1\}$ is *balanced* if the number of 0s (or, equivalently, 1s) in every two factors of the same length differs by at most 1.

We let *Bal* denote the set of balanced words over the alphabet $\{0,1\}$.

Remark 12. Balance is a *factorial property*, i.e., every factor of a balanced word is itself balanced.

Binary balanced words are precisely the finite factors of Sturmian words.

From the definition of balanced word, it immediately follows that if w is a binary balanced word, then for every $1 \leq k \leq |w|$, the factors of w of length k can have at most two different Parikh vectors: $(i, k - i)$, $(i - 1, k - i + 1)$, for some $0 \leq i \leq k$. According to [29], we call the factors of the first kind *light factors* and those of the second kind *heavy factors*. By convention, if all factors of length k have the same Parikh vector, we consider them light.

For example, the word 1011010 is balanced; its factors of length 3 are: 010, 011, 101, 110. The first one is the only light factor of length 3, the other ones are heavy.

Notice that in a primitive lower Christoffel word $w_{a,b} = 0C1$, all prefixes are light factors and all proper suffixes are heavy. Indeed, if $0C1$ is a primitive lower Christoffel word then since C is a palindrome, for every prefix $0v$ of $w_{a,b}$ we have that the suffix of the same length is $v^R 1$, and clearly v and v^R have the same Parikh vector. The situation is of course symmetric for primitive upper Christoffel words.

Theorem 13 ([3]). *Let w be a word over $\{0,1\}$. Then w is a primitive (lower or upper) Christoffel word if and only if it is balanced and unbordered.*

In particular, the set of primitive lower Christoffel words is precisely the set of balanced *Lyndon words* over the alphabet $\{0,1\}$ (for the order $0 < 1$). A Lyndon word is a primitive word that is lexicographically smaller than all its conjugates (or, equivalently, lexicographically smaller than all its proper suffixes).
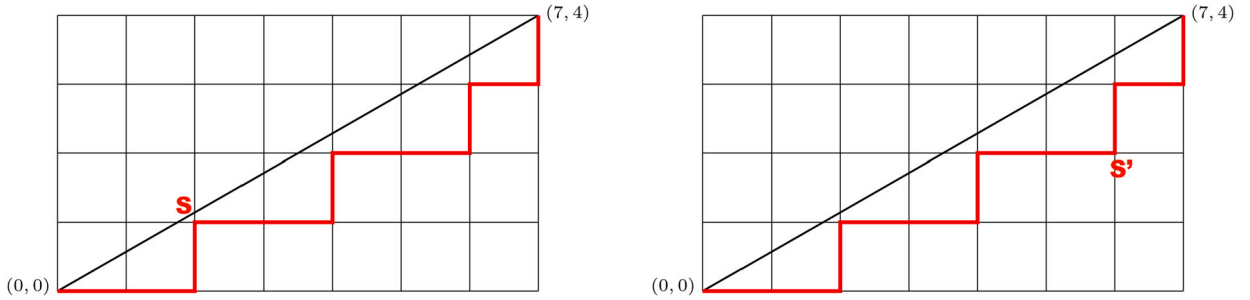
Fig. 4. The standard factorization $0Q1 \cdot 0P1 = 001 \cdot 00100101$ (left) and the palindromic factorization $0P0 \cdot 1Q1 = 00100100 \cdot 101$ (right) of the lower Christoffel word $w_{7,4}$. The point S determined by the standard factorization is the closest to the Euclidean segment, while the point S' determined by the palindromic factorization is the farthest.

Proposition 14 ([6]). *For every coprime a, b, the primitive lower Christoffel word $w_{a,b}$ is the greatest (for the lexicographic order) Lyndon word among all Lyndon words having Parikh vector (a, b).*

2.2. Factorizations of Christoffel words

A basic result in the theory of Lyndon words is that every Lyndon word w of length $|w| \geq 2$ has a *standard factorization* $w = uv$, where v is the lexicographically least proper suffix of w (or, equivalently, the longest proper suffix of w that is a Lyndon word), see [19]. Since primitive lower Christoffel words are Lyndon words, every primitive lower Christoffel word of length $|w| \geq 2$ has a standard factorization.

On the other hand, by Proposition 7, a primitive lower Christoffel word is a conjugate of its reversal (the corresponding upper Christoffel word); hence by Proposition 1, every primitive lower Christoffel word of length $|w| \geq 2$ has a unique *palindromic factorization*.

Let $w_{a,b} = 0C1$ be a primitive lower Christoffel word, so that a and b are coprime integers. If the central word C is not a power of a single letter, then by Theorem 5 there exist central words P and Q such that $C = P01Q = Q10P$ so that $w_{a,b} = 0C1 = 0P0 \cdot 1Q1 = 0Q1 \cdot 0P1$.

Hence, we have the following factorizations:

1. $0C1 = 0P0 \cdot 1Q1$ (palindromic factorization);
2. $0C1 = 0Q1 \cdot 0P1$ (standard factorization).

If instead $C = 0^n$ (the case $C = 1^n$ is analogous) we have:

1. $0C1 = 0^{n+1} \cdot 1$ (palindromic factorization);
2. $0C1 = 0 \cdot 0^n 1$ (standard factorization).

Actually, for every lower Christoffel word $w_{a,b}$ (not necessarily primitive), the following holds: Let P, P' be two palindromes such that $w_{a,b} = PP'$. Then $P'P = w_{a,b}^R = W_{a,b}$.

Remark 15. Notice that the lengths of the two factors in both the palindromic and the standard factorization of the primitive lower Christoffel word $w_{a,b} = 0C1$ are precisely the two coprime periods of the central word C whose sum is $a + b$, i.e., by Proposition 10, the multiplicative inverses of a and b modulo $a + b$.

From the geometric point of view, the standard factorization divides $w_{a,b}$ in two shorter Christoffel words, and it determines the point S of the encoded path that is *closest* to the Euclidean segment from $(0, 0)$ to (a, b); the palindromic factorization, instead, divides $w_{a,b}$ in two palindromes and determines the point S' that is *farthest* from the Euclidean segment (see [7,30]). An example is given in Fig. 4.

Remark 16. In the lower Christoffel word $(0C1)^2 = 0C1 \cdot 0C1$ we have precisely one occurrence as a factor of the upper primitive Christoffel word $1C0$. Using the palindromic factorization, we can write

$$(0C1)^2 = 0P0 \cdot 1Q10P0 \cdot 1Q1 = 0P0 \cdot 1C0 \cdot 1Q1$$

(the case when C is a power of a letter is analogous); so, the occurrence of $1C0$ in $(0C1)^2$ begins precisely in the point S' (see [7]).

Table 1

The Christoffel matrix $\mathcal{A}_{7,4}$.

$$\mathcal{A}_{7,4} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

2.3. Digital approximations of a given Euclidean segment

Fix a point (a, b) in the grid (a and b nonnegative integers, not both 0) and consider all the binary words that encode a path from $(0,0)$ to (a,b), i.e., all the words that have Parikh vector (a,b). We aim at giving a combinatorial interpretation of those paths that better approximate the Euclidean segment from $(0,0)$ to (a,b) according to some definition of digital approximation. Among these, there are the paths encoded by the conjugates of the lower and of the upper Christoffel word, as we are showing below.

Let $a, b > 0$ and $n = a + b$. The *Christoffel matrix* $\mathcal{A}_{a,b}$ is the $n \times n$ matrix in which the first column is a block of a 0's followed by a block of b 1's, and every subsequent column is obtained by shifting up the block of 1's by b positions, modulo n, see [16,7]. An example is given in Table 1.

The first row of the Christoffel matrix $\mathcal{A}_{a,b}$ is the lower Christoffel word $w_{a,b}$. Every subsequent row is obtained from the previous one by swapping a 01 factor with 10. The last row is the upper Christoffel word $W_{a,b}$. Actually, the rows are precisely the conjugates of $w_{a,b}$ and appear in lexicographic order. The rows are all distinct if and only if $w_{a,b}$ is primitive, i.e., a and b are coprime.

Definition 17. A word is *circularly balanced* if all its conjugates (including the word itself) are balanced.

Proposition 18 (*see for instance [13]*). *A word is circularly balanced if and only if it is a conjugate of a Christoffel word, i.e., if and only if it is a row of a Christoffel matrix.*

So, the conjugates of the lower and upper Christoffel word with Parikh vector (a,b), i.e., the circularly balanced words with Parikh vector (a,b), are contained in the region $R(a,b)$ of the discrete plane whose contour is delimited by the lower and the upper Christoffel word with Parikh vector (a,b) (Fig. 2).

From a geometric point of view, however, *every* balanced word with Parikh vector (a,b) is a digital approximation of the Euclidean segment from $(0,0)$ to (a,b). Indeed, all these words encode paths that are contained in the region $R(a,b)$. A combinatorial proof of this fact is given in the next Proposition 20.

It is worth noticing that, in general, there also exist non-balanced words with Parikh vector (a,b) whose corresponding paths lie in the region $R(a,b)$. As an example, the word 01001001100 is contained in the region $R(7,4)$ but is not balanced.

Example 19. Out of the $\binom{7+4}{4} = 330$ binary words with Parikh vector $(7,4)$, only 112 of them encode paths that lie no more than $\sqrt{2}$ away from the Euclidean segment joining $(0,0)$ to $(7,4)$, i.e., are contained in the region $R(7,4)$ delimited by the lower and the upper Christoffel words of Parikh vector $(7,4)$ (see Fig. 2). In particular, all 19 balanced words with Parikh vector $(7,4)$ are among them (see Example 31).

Proposition 20. *Let $a, b > 0$ and $n = a + b$. For every balanced word u with Parikh vector (a,b), and for every $1 < k < n$, the prefix of length k of u has the same Parikh vector as the prefix of length k of either $w_{a,b}$ or $W_{a,b}$.*

Proof. By contradiction, let k be the least integer such that there exist a balanced word x with Parikh vector (a,b) whose prefix u_1 of length k satisfies either $|u_1|_1 < h$ or $|u_1|_1 > h + 1$, where $h = |w_{a,b}[1..k]|_1$ (and $h + 1 = |W_{a,b}[1..k]|_1$). Without loss of generality, assume the former case holds.

Let $n = dk + r$ for some integers d, r with $d \geq 1$ and $0 \leq r < k$, so that we may write $x = u_1 \cdots u_d v$ for some words such that $|v| = r$ and $|u_i| = k$ for $i = 1, \ldots, d$. Since x is balanced, each of its factors of length k has at most h ones, whereas all factors of the same length in $w_{a,b}$ have at *least* h ones. Together with $|x|_1 = q$ and $|u_1|_1 < h$, this yields $|u_1 \cdots u_d|_1 < |w_{a,b}[1..dk]|_1$ and hence $|v|_1 > |w_{a,b}[dk+1..n]|_1$. In particular, this implies $r > 0$.

Therefore, v^R is a prefix of the balanced word x^R that has more ones than $(w_{a,b}[dk+1..n])^R = (W_{a,b}[1..r])$. As $r < k$, this contradicts our choice for k. $\quad\square$

From Proposition 20, we can derive the following characterization of lower Christoffel words, which is somehow dual to that presented in Proposition 14.

Corollary 21. *For every pair (a, b), the lower Christoffel word $w_{a,b}$ is the smallest (in the lexicographic order) balanced word having Parikh vector (a, b).*

From Corollary 21 and Proposition 14, it follows that if a and b are coprime integers, all Lyndon words with Parikh vector (a, b) are lexicographically smaller than all balanced words with the same Parikh vector, with the exception of the primitive lower Christoffel word, which is the only word in the intersection of the two sets.

3. Counting balanced words with a given Parikh vector

From the results presented in the previous section, a natural question arises:

Problem 22. Given two positive integers a and b, how many balanced words are there with Parikh vector (a, b)?

A recursive formula for the number $Bal(a, b)$ of balanced words with Parikh vector (a, b) was given in [1]. Here, we give an explicit closed formula.

Notice that the number of binary balanced words of *length $n = a + b$* is known to be

$$1 + \sum_{i=1}^{n} (n - i + 1)\phi(i), \tag{1}$$

where ϕ is the Euler totient function, i.e., $\phi(n)$ is the number of positive integers smaller than or equal to n and coprime with n (cf. [21,18]). But this seems to be of little help for obtaining a bivariate formula depending on specific values of a and b.

For simplicity of exposition, we will make use of the infinite periodic word $w_{a,b}^{\omega}$. Notice that the infinite periodic word $w_{a,b}^{\omega}$ (resp. $W_{a,b}^{\omega}$) is the digital approximation from below (resp., from above) of the ray $y = \frac{b}{a} x$.

Remark 23. From the definition of Christoffel word, it follows that for all lengths $k \geq 0$, the height of the prefix of length k of $w_{a,b}^{\omega}$ (resp. of $W_{a,b}^{\omega}$) is $\left\lfloor k \frac{b}{a+b} \right\rfloor$ (resp. $\left\lceil k \frac{b}{a+b} \right\rceil$).

Lemma 24. *Let u be a balanced word of Parikh vector (a, b), and let (α, β) be the Parikh vector of its fractional root ρ_u. Then, either:*

- $\frac{b-1}{a+1} < \frac{\beta}{\alpha} < \frac{b}{a}$, *in which case u is a heavy factor of $w_{\alpha,\beta}^{\omega}$; or*
- $\frac{b}{a} \leq \frac{\beta}{\alpha} < \frac{b+1}{a-1}$, *in which case u is a light factor of $w_{\alpha,\beta}^{\omega}$.*

Proof. It is known (see [13]) that ρ_u is a conjugate of the primitive lower Christoffel word $w_{\alpha,\beta}$, so that u is a factor of length $a + b$ of $w_{\alpha,\beta}^{\omega}$. If $(\alpha, \beta) = (a, b)$, then u is unbordered, i.e., $u = w_{\alpha,\beta}$ or $u = W_{\alpha,\beta}$, hence u is a light factor by definition. Suppose then $(\alpha, \beta) \neq (a, b)$, i.e., $\beta/\alpha \neq b/a$. If u is a heavy factor (resp. a light factor) of $w_{\alpha,\beta}^{\omega}$, then by Proposition 20 it has the same Parikh vector as the prefix of length $a + b$ of $W_{\alpha,\beta}^{\omega}$ (resp., of $w_{\alpha,\beta}^{\omega}$); therefore, by Remark 23, we have

$$\left\lceil \frac{\beta}{\alpha + \beta}(a + b) \right\rceil = b \quad \left(\text{resp.,} \quad \left\lfloor \frac{\beta}{\alpha + \beta}(a + b) \right\rfloor = b \right). \tag{2}$$

Solving for β (resp., for α), eq. (2) is equivalent to $\frac{b-1}{a+1}\alpha < \beta < \frac{b}{a}\alpha$ (resp., to $\frac{a-1}{b+1}\beta < \alpha < \frac{a}{b}\beta$), whence the statement. $\square$

Conversely, all the heavy (resp. light) factors of length $a + b$ and minimal period $\alpha + \beta$ of the words $w_{\alpha,\beta}^{\omega}$ such that $\frac{b-1}{a+1} < \frac{\beta}{\alpha} < \frac{b}{a}$ (resp. $\frac{b}{a} < \frac{\beta}{\alpha} < \frac{b+1}{a-1}$) have Parikh vector (a, b).

The previous result allows us to give the following formula:

Theorem 25. *Let $H_{\alpha,\beta}(n)$ (resp. $L_{\alpha,\beta}(n)$) be the cardinality of the set of heavy (resp. light) factors of length n of $w_{\alpha,\beta}^{\omega}$ having minimal period $\alpha + \beta$, for positive α and β. The number of balanced words with Parikh vector (a, b) is given by*

$$Bal(a, b) = \sum_{\alpha=1}^{a} \sum_{\frac{b-1}{a+1}\alpha < \beta < \frac{b}{a}\alpha} H_{\alpha,\beta}(a + b) + \sum_{\beta=1}^{b} \sum_{\frac{a-1}{b+1}\beta < \alpha \leq \frac{a}{b}\beta} L_{\alpha,\beta}(a + b)$$

if $ab \neq 0$, and $Bal(a, b) = 1$ otherwise.

In this way, we reduced the problem of computing $Bal(a, b)$ to that of computing the quantities $H_{\alpha,\beta}(n)$ and $L_{\alpha,\beta}(n)$. We will now show arithmetic formulas for computing these quantities.

Proposition 26. *Let $N_{\alpha,\beta}(n)$ be the cardinality of the set of factors of length n of $w_{\alpha,\beta}^{\omega}$ having minimal period $\alpha + \beta$, for positive α and β. Then*

$$N_{\alpha,\beta}(n) = \begin{cases} 0 & \text{if } \gcd(\alpha,\beta) > 1 \text{ or } n < \alpha + \beta \\ 2(n - \alpha - \beta + 1) & \text{if } n < \alpha + \beta + \min\{\alpha',\beta'\}, \\ n - \max\{\alpha',\beta'\} + 1 & \text{if } \alpha + \beta + \min\{\alpha',\beta'\} \leq n < \alpha + \beta + \max\{\alpha',\beta'\}, \\ \alpha + \beta & \text{otherwise,} \end{cases}$$

where α' (resp., β') is the multiplicative inverse of α (resp., β) modulo $\alpha + \beta$.

Proof. The first case is trivial, since all factors of $w_{\alpha,\beta}^{\omega}$ have the period $(\alpha + \beta)/\gcd(\alpha,\beta)$. If α and β are coprime, then by Remark 15 the Christoffel word $w_{\alpha,\beta}$ can be factored as UV, where U and V are palindromes of length α' and β' respectively. Therefore, we need to count factors of length n in $(UV)^{\omega}$ that contain an occurrence of UV or of $VU = W_{\alpha,\beta}$. It is easy to see that if $r := n - \alpha - \beta < \min\{\alpha',\beta'\} = \min\{|U|,|V|\}$, such factors are precisely all factors of length $n = \alpha' + \beta' + r$ contained in the words $(V[1..r])^R UV(U[1..r])$ and $(U[1..r])^R VU(V[1..r])$. This gives $N_{\alpha,\beta}(n) = 2(r+1)$ as desired.

If $\alpha' = |U| \leq r < |V| = \beta'$ (resp., $\beta' \leq r < \alpha'$), then a factor of $(UV)^{\omega}$ of length n contains an occurrence of UV or VU if and only if it occurs as a factor in $(V[1..r])^R UVU(V[1..r])$ (resp., in $(U[1..r])^R VUV(U[1..r])$); thus, $N_{\alpha,\beta}(n)$ equals $|U| + r + 1 = n - \beta' + 1$ (resp., $|V| + r + 1 = n - \alpha' + 1$).

Finally, if $n \geq \alpha + \beta + \max\{\alpha',\beta'\}$, then *all* $\alpha + \beta$ distinct factors of length n in $(UV)^{\omega}$ contain at least an occurrence of UV or VU. $\square$

Example 27. Let $\alpha = 7$, $\beta = 3$, so that $\alpha' = 3$ and $\beta' = 7$, and $n = 11$. The $10 = \alpha + \beta$ factors of length 11 of $w_{7,3}^{\omega}$ are: 00010010010, 00100010010, 00100100010, 00100100100, 01000100100, 01001000100, 01001001000, 10001001001, 10010001001, 10010010001. Among them, those of minimal period 10 are the following four ($4 = 2(10 - \alpha - \beta + 1)$): 00010010010, 01001001000, 10001001001, 10010010001.

Let $\alpha = 5$, $\beta = 4$, so that $\alpha' = 2$ and $\beta' = 7$, and $n = 11$. The $9 = \alpha + \beta$ factors of length 11 of $w_{5,4}^{\omega}$ are: 00101010100, 01001010101, 01010010101, 01010100101, 01010101001, 10010101010, 10100101010, 10101001010, 10101010010. Among them, those of minimal period 9 are the following five ($5 = 10 - \max(\alpha' + \beta') + 1$): 00101010100, 01001010101, 01010101001, 10010101010, 10101010010.

Next, we need to distinguish by height, i.e., between light and heavy factors. First of all, notice that for every $n \geq 0$, $w_{\alpha,\beta}^{\omega}$ has exactly $\alpha + \beta$ distinct factors of length $\alpha + \beta + n - 1$. Moreover, we have the following

Proposition 28. *Let α, β be coprime positive integers, and u be any factor of $w_{\alpha,\beta}^{\omega}$ of length $\alpha + \beta + n - 1$ for some $n \geq 0$. The number of occurrences of heavy (resp. light) factors of length n in u is $n\beta \bmod (\alpha + \beta)$ (resp. $n\alpha \bmod (\alpha + \beta)$).*

Proof. We give the proof for heavy factors only, as the one for light factors will follow from it. First notice that u has exactly $\alpha + \beta$ occurrences of factors of length n. We prove our claim by induction on n. The case $n = 0$ is trivial, as the empty word is a light factor by our definition. Let then $n \geq 1$, and assume the statement holds for $n - 1$, so that $u = u'x$ for some $x \in \{0,1\}$ and u' has $(n-1)\beta \bmod (\alpha + \beta)$ occurrences of heavy factors of length $n - 1$.

If the nth letter of $w_{\alpha,\beta}^{\omega}$ is 0, then each occurrence of a heavy factor of length $n - 1$ in u' is necessarily followed by 0, by the balance condition. It follows that the occurrences of heavy factors of length $n - 1$ are obtained by extending heavy factors with a 0 or light factors with a 1. Since the suffix of u of length $\alpha + \beta$ contains β ones, we get

$$(n-1)\beta \bmod (\alpha + \beta) + \beta = n\beta \bmod (\alpha + \beta) \tag{3}$$

heavy factors of length n, where equality in (3) is due to fact that the total number of occurrences of factors of length n is $\alpha + \beta$.

If the nth letter of $w_{\alpha,\beta}$ is 1 instead, then each occurrence of a light factor of length $n - 1$ in u' must be followed by 1, so that heavy factors of length n are only obtained by extending heavy factors with a 1. Since the number of zeroes in $u[n..\alpha + \beta + n - 1]$ is α, we obtain $(n-1)\beta \bmod (\alpha + \beta) - \alpha$ occurrences of heavy factors, once again yielding our claimed value. $\square$

Proposition 29. *Let $H_{\alpha,\beta}(n)$ be the cardinality of the set of heavy factors of length n of $w_{\alpha,\beta}^{\omega}$ having minimal period $\alpha + \beta$, for positive α and β. Then*

$$H_{\alpha,\beta}(n) = \begin{cases} 0 & \text{if } N_{\alpha,\beta}(n) = 0, \\[2mm] 2\left(\displaystyle\sum_{i=0}^{n-\alpha-\beta} \lceil \sigma(n-i) \rceil + \lfloor \sigma i \rfloor\right) - \lfloor \sigma n \rfloor N_{\alpha,\beta}(n) & \text{if } n < \alpha + \beta + \min\{\alpha', \beta'\}, \\[2mm] \left(\displaystyle\sum_{i=0}^{n-\alpha'} \lfloor \sigma(n-i) \rfloor + \lceil \sigma i \rceil\right) - \lfloor \sigma n \rfloor N_{\alpha,\beta}(n) & \text{if } \alpha + \beta + \beta' \le n < \alpha + \beta + \alpha', \\[2mm] \left(\displaystyle\sum_{i=0}^{n-\beta'} \lceil \sigma(n-i) \rceil + \lfloor \sigma i \rfloor\right) - \lfloor \sigma n \rfloor N_{\alpha,\beta}(n) & \text{if } \alpha + \beta + \alpha' \le n < \alpha + \beta + \beta', \\[2mm] n\beta \bmod (\alpha + \beta) & \text{otherwise,} \end{cases}$$

where α' (resp., β') is the multiplicative inverse of α (resp., β) modulo $\alpha + \beta$, and $\sigma = \beta/(\alpha + \beta)$.

Proof. The first case is clear, as $H_{\alpha,\beta}(n) \le N_{\alpha,\beta}(n)$ in general. If $\gcd(\alpha, \beta) = 1$, then as in the proof of Proposition 26 we set $r = n - \alpha - \beta = n - \alpha' - \beta'$ for simplicity, and write $w_{\alpha,\beta} = UV$ for palindromes U, V such that $|U| = \alpha'$ and $|V| = \beta'$.

If $0 < r < \max\{\alpha', \beta'\}$, we can obtain $H_{\alpha,\beta}(n)$ by calculating the sum of all heights of factors of length n and minimal period $\alpha + \beta$, and then subtracting the sum of heights of $N_{\alpha,\beta}(n)$ *light* factors, i.e., $\lfloor \sigma n \rfloor N_{\alpha,\beta}(n)$.

For the case $r < \min\{\alpha', \beta'\}$, by Proposition 26 it suffices to observe that every factor of length n in $(V[1..r])^R UV(U[1..r])$ (and similarly in its reverse) can be split in two parts, the first one ending after the occurrence of UV. In other words, for $i = 0, \ldots, r$ the two parts are $(VU(V[1..r-i]))^R$ and $U[1..i]$, respectively of heights $\lceil \sigma(n-i) \rceil$ and $\lfloor \sigma i \rfloor$.

The cases $\beta' \le r < \alpha'$ and $\alpha' \le r < \beta'$ are similarly dealt with, by considering the words $(V[1..r])^R UVU(V[1..r])$ and $(U[1..r])^R VUV(U[1..r])$ found in the proof of Proposition 26.

Finally, in the case where $r \ge \max\{\alpha', \beta'\}$, all $N_{\alpha,\beta}(n) = \alpha + \beta$ factors of length n have minimal period $\alpha + \beta$ and occur exactly once in each factor of $w_{\alpha,\beta}^\omega$ of length $\alpha + \beta + n - 1$. The assertion then follows by Proposition 28. $\square$

Example 30. Let $\alpha = 7$, $\beta = 3$, so that $\alpha' = 3$ and $\beta' = 7$, and $n = 11$. Among the four factors of length 11 and minimal period $10 = \alpha + \beta$ of $w_{7,3}^\omega$, namely 00010010010, 01001001000, 10001001001, 10010010001 (see Example 27), the heavy factors are the last two of the list, while the first two are light.

Let $\alpha = 5$, $\beta = 4$, so that $\alpha' = 2$ and $\beta' = 7$, and $n = 11$. Among the five factors of length 11 and minimal period $9 = \alpha + \beta$ of $w_{5,4}^\omega$, namely 00101010100, 01001010101, 01010101001, 10010101010, 10101010010 (see Example 27), the heavy factors are the last four of the list, while the first one is the only light factor.

Since

$$L_{\alpha,\beta}(n) = N_{\alpha,\beta}(n) - H_{\alpha,\beta}(n),$$

we have all the ingredients to count the number of balanced words with a given Parikh vector.

Example 31. Let $a = 7$, $b = 4$. There are 19 balanced words with Parikh vector $(7, 4)$, namely: 00100100101, 00100101001, 00101001001, 00101001010, 00101010010, 00101010100, 01001001001, 01001001010, 01001010010, 01001010100, 01010010010, 01010010100, 10001001001, 10010001001, 10010010001, 10010010010, 10010010100, 10010100100, 10100100100. The pairs (α, β) with $1 \le \alpha$, $1 \le \beta$ and $\frac{b-1}{a+1} < \frac{\beta}{\alpha} < \frac{b}{a}$ are: $(2,1)$, $(5,2)$, and $(7,3)$; those with $1 \le \alpha$, $1 \le \beta$ and $\frac{b}{a} \le \frac{\beta}{\alpha} < \frac{b+1}{a-1}$ are: $(3,2)$, $(4,3)$, $(5,3)$, $(5,4)$ and $(7,4)$. Theorem 25 gives

$$\begin{aligned} Bal(7,4) &= H_{2,1}(11) + H_{5,2}(11) + H_{7,3}(11) \\ &\quad + L_{3,2}(11) + L_{4,3}(11) + L_{5,3}(11) + L_{5,4}(11) + L_{7,4}(11) \\ &= 2 + 1 + 2 + 3 + 2 + 6 + 1 + 2 = 19. \end{aligned}$$

In Fig. 5 we plotted the number of binary balanced words with Parikh vector (a, b) for a and b ranging from 1 to 36.

4. Minimal forbidden balanced words and minimal almost balanced words

In this section, we give a description of those words that are not balanced but such that all of their proper factors are balanced. Geometrically, they represent those paths that have a minimal difference with the paths that are encoded by balanced words.

Definition 32. A balanced word v over $\{0, 1\}$ is *right special* (resp., *left special*) if both $v0$ and $v1$ are balanced (resp., if both $0v$ and $1v$ are balanced). It is *bispecial* if it is both left and right special. A bispecial balanced word v is *strictly bispecial* if all of $0v1$, $1v0$, $0v0$, $1v1$ are balanced (see [3]).
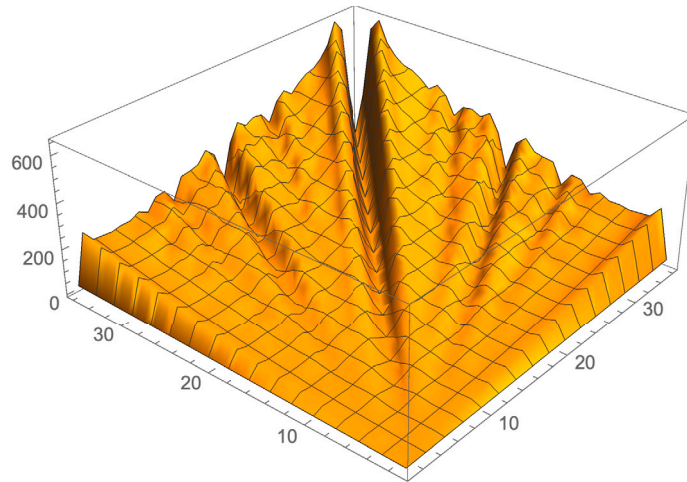
Fig. 5. A three-dimensional plot showing the number of binary balanced words with Parikh vector (a, b) for a and b ranging from 1 to 36.

Strictly bispecial balanced words are precisely central words [12]. Hence, left special (resp. right special) balanced words are precisely the prefixes (resp. suffixes) of central words.

Theorem 33 *([15])*. *A balanced word v is bispecial if and only if $0v1$ is a (not necessarily primitive) lower Christoffel word.*

In particular, if (and only if) $0v1$ is a *primitive* lower Christoffel word (i.e., v is a palindrome, hence a central word) the word v is a strictly bispecial balanced word.

Example 34. Let $0v1$ be the Christoffel word $0 \cdot 0100 \cdot 1$. The word 0100 is bispecial but not strictly bispecial. Indeed, $0 \cdot 0100 \cdot 1$, $0 \cdot 0100 \cdot 0$ and $1 \cdot 0100 \cdot 1$ are balanced, but $1 \cdot 0100 \cdot 0$ is not.

Definition 35. Let L be a factorial language, i.e., a (finite or infinite) set of words closed under taking factors. We say that a word w is a *minimal forbidden word* for L if w does not belong to L but all proper factors of w do.

Let $\mathcal{MF}(L)$ denote the set of minimal forbidden words of the factorial language L. A word $w = xvy$, x, y letters, belongs to $\mathcal{MF}(L)$ if and only if

1. $xvy \notin L$;
2. $xv, vy \in L$.

Theorem 36 *([22])*. *There is a bijection between factorial languages and their sets of minimal forbidden words.*

As a consequence, $\mathcal{MF}(L)$ uniquely determines L.

The next theorem gives a characterization of the minimal forbidden words for the language *Bal* of binary balanced words.

Theorem 37 *([15])*. $\mathcal{MF}(Bal) = \{ yvx \mid \{x, y\} = \{0, 1\}, \ xvy \text{ is a non-primitive Christoffel word} \}$.

Example 38. The word $0 \cdot 0010 \cdot 1$ is not balanced, but all its proper factors are. Indeed, $1 \cdot 0010 \cdot 0$ is the square of the primitive upper Christoffel word 100.

The word $0 \cdot 0010010 \cdot 1$ is not balanced, but all its proper factors are. Indeed, $1 \cdot 0010010 \cdot 0$ is the cube of the primitive upper Christoffel word 100.

Corollary 39 *([15])*. *For every $n > 0$, there are exactly $n - \phi(n) - 1$ words of length n in $\mathcal{MF}(Bal)$ that start with 0, and they are all Lyndon words. Here ϕ is the Euler totient function.*

Recall that a word w is not balanced if and only if there exists a palindrome v such that $0v0$ and $1v1$ are both factors of w [20]. We call such a pair of factors an *imbalance pair*. In 2011, Provençal [26] studied the language of *minimal almost balanced words*, *MAB*s, i.e., minimal (for the factorial order) words with the property that there exists a unique imbalance pair.

Example 40. The word 000101 is almost balanced with unique imbalance pair 000, 101, but all its proper factors are balanced. Hence it is a *MAB*.

On the contrary, the word 000100101 is not a *MAB*, since 000, 101 and 000100, 100101 are distinct imbalance pairs.

Theorem 41 ([26]). $MAB = \{u^2 v^2, (u^2 v^2)^R \mid uv$ *is the standard factorization of a primitive lower Christoffel word*$\}$.

We now give an alternative description:

Theorem 42. $MAB = \{ywx \mid \{x, y\} = \{0, 1\},$ xwy *is the square of a primitive Christoffel word*$\}$. *Therefore,* $MAB \subseteq \mathcal{MF}(Bal)$.

Proof. Let $uv = 0C1$ be the standard factorization of a lower primitive Christoffel word. Let $u = 0Q1$, $v = 0P1$ for P, Q central words, so that $uv = 0Q10P1 = 0P01Q1$. Then $u^2 v^2 = 0Q10Q10P10P1 = 0Q10P01Q10P1$, hence $1Q10P01Q10P0 = (1C0)^2$. The other cases are similar. □

If $w = xCyxCy$ is a square of a primitive Christoffel word, then the corresponding MAB is $yCyxCx$. A word of this form is in fact a trapezoidal non-balanced word. These words have been studied by D'Alessandro [10] and by the authors with Michelangelo Bucci [9], where words of the form $CyxC$ are called semicentral words.

CRediT authorship contribution statement

Alessandro De Luca: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Gabriele Fici:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Data availability

No data was used for the research described in the article.

References

[1] N. Bedaride, E. Domenjoud, D. Jamet, J.-L. Rémy, On the number of balanced words of given length and height over a two letter alphabet, Discret. Math. Theor. Comput. Sci. 12 (3) (2010) 41–62.

[2] J. Berstel, Sturmian and Episturmian words (a survey of some recent results), in: CAI 2007, in: Lecture Notes in Comput. Sci., vol. 4728, Springer, 2007, pp. 23–47.

[3] J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theor. Comput. Sci. 178 (1–2) (1997) 171–203.

[4] J. Berstel, A. Lauve, C. Reutenauer, F. Saliola, Combinatorics on Words: Christoffel Words and Repetition in Words, CRM Monograph Series, vol. 27, American Mathematical Society, 2008.

[5] V. Berthé, A. de Luca, C. Reutenauer, On an involution of Christoffel words and Sturmian morphisms, Eur. J. Comb. 29 (2) (2008) 535–553.

[6] J.-P. Borel, F. Laubie, Quelques mots sur la droite projective réelle, J. Théor. Nr. Bordx. 5 (1) (1993) 23–51.

[7] J.-P. Borel, C. Reutenauer, On Christoffel classes, RAIRO Theor. Inform. Appl. 40 (1) (2006) 15–27.

[8] S. Brlek, S. Hamel, M. Nivat, C. Reutenauer, On the palindromic complexity of infinite words, Int. J. Found. Comput. Sci. 15 (2) (2004) 293–306.

[9] M. Bucci, A. De Luca, G. Fici, Enumeration and structure of trapezoidal words, Theor. Comput. Sci. 468 (2013) 12–22.

[10] F. D'Alessandro, A combinatorial problem on trapezoidal words, Theor. Comput. Sci. 273 (1–2) (2002) 11–33.

[11] A. de Luca, On some combinatorial problems in free monoids, Discrete Math. 38 (2) (1982) 207–225.

[12] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics, Theor. Comput. Sci. 183 (1) (1997) 45–82.

[13] A. de Luca, A. De Luca, Some characterizations of finite Sturmian words, Theor. Comput. Sci. 356 (1–2) (2006) 118–125.

[14] A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, Theor. Comput. Sci. 136 (2) (1994) 361–385.

[15] G. Fici, On the structure of bispecial Sturmian words, J. Comput. Syst. Sci. 80 (4) (2014) 711–719.

[16] O. Jenkinson, L.Q. Zamboni, Characterisations of balanced words via orderings, Theor. Comput. Sci. 310 (1–3) (2004) 247–271.

[17] R. Klette, A. Rosenfeld, Digital straightness—a review, Discrete Appl. Math. 139 (2004) 197–230.

[18] E.P. Lipatov, A classification of binary collections and properties of homogeneity classes (in Russian), Probl. Kibern. 39 (1982) 67–84.

[19] M. Lothaire, Combinatorics on Words, Cambridge Mathematical Library, Cambridge Univ. Press, 1997.

[20] M. Lothaire, Algebraic Combinatorics on Words, Encyclopedia of Mathematics and Its Applications., Cambridge Univ. Press, 2002.

[21] F. Mignosi, On the number of factors of Sturmian words, Theor. Comput. Sci. 82 (1991) 71–84.

[22] F. Mignosi, A. Restivo, M. Sciortino, Forbidden factors in finite and infinite words, in: J. Karhumäki, H.A. Maurer, G. Paun, G. Rozenberg (Eds.), Jewels are Forever, Contributions on Theoretical Computer Science in Honor of Arto Salomaa, Springer, 1999, pp. 339–350.

[23] M. Morse, G.A. Hedlund, Symbolic dynamics, Am. J. Math. 60 (1938) 815–866.
[24] M. Morse, G.A. Hedlund, Symbolic dynamics II: Sturmian trajectories, Am. J. Math. 62 (1940) 1–42.
[25] G. Pirillo, A curious characteristic property of standard Sturmian words, in: H. Crapo, D. Senato (Eds.), Algebraic Combinatorics and Computer Science: A Tribute to Gian-Carlo Rota, Springer, Milan, Milano, 2001, pp. 541–546.
[26] X. Provençal, Minimal non-convex words, Theor. Comput. Sci. 412 (27) (2011) 3002–3009.
[27] C. Reutenauer, Studies on finite Sturmian words, Theor. Comput. Sci. 591 (2015) 106–133.
[28] C. Reutenauer, From Christoffel Words to Markoff Numbers, Oxford University Press, 2018.
[29] M. Rigo, Formal Languages, Automata and Numeration Systems 1: Introduction to Combinatorics on Words, Wiley, 2014.
[30] L. Tarsissi, Balance properties on Christoffel words and applications, PhD thesis, Université Grenoble Alpes, 2017.