Minimal Non-Convex Words

Xavier Provençal[a,b,1],

[a] LIRMM - UMR 5506 CNRS, Université Montpellier II,
161 rue Ada, 34000 Montpellier, France.
[b] LAMA - UMR 5127 CNRS, Université de Savoie,
73376 Le Bourget du Lac, France.

Abstract

Using a combinatorial characterization of digital convexity based on words, one defines the language of convex words. The complement of this language forms an ideal whose minimal elements, with respect to the factorial ordering, appear to have a particular combinatorial structure very close to the Christoffel words. In this paper, those words are completely characterized as those of the form uw^kv where $k \geq 1$, $w = u \cdot v$ and u, v, w are Christoffel words. Also, by considering the most balanced among the unbalanced words, we obtain a second characterization for a special class of the minimal non-convex words that are of the form u^2v^2 corresponding to the case $k = 1$ in the previous form.

1. Introduction

In computer imagery, polyominoes, that are defined as the interior of a closed non-intersecting grid path of $\mathbb{Z}^2$, are used to represent discretized objects from the *real world*. Given such a polyomino, one may ask whether the *real* discretized object is convex or not. From this comes the notion of digital convexity. We say that a polyomino P is *digitally convex* if its convex hull contains no points of $\mathbb{Z}^2$ outside of P (see Eckhardt (2001) for a review on digital convexity).

Over the last 40 years, many characterizations of digital convexity have been stated. Recently, a new one, based completely on words combinatoric was established in Brlek et al. (2009). It uses the Freeman chain code to represent the boundary of a *hv*-convex polyomino P by a word over a four

Email address: `xavier.provencal@lirmm.fr` (Xavier Provençal)

[1] Author supported by a scholarship from FQRNT (Québec).

letter alphabet; this word is then split into four *quadrant words*, each being written over a two letter alphabet. It was shown that P is digitally convex if and only if the unique factorization as decreasing Lyndon words of each of its quadrant words is composed of only Christoffel words. Moreover, in the case of a digitally convex polyomino, its convex hull is directly given by the Lyndon factorization.

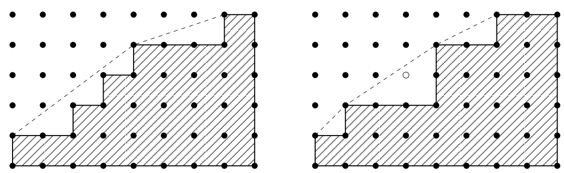

Figure 1: Left: A convex polyomino with North-West quadrant word $1001010100010 = (1, 0010101, 0001, 0)_{\mathbf{Lyn}}$. Right: a non-convex polyomino with North-West quadrant word $1010001100100 = (1, 01, 00011001, 00)_{\mathbf{Lyn}}$.

Following this characterization, we define the set of *convex words* **CV** over a two letter alphabet and show that its complement $\mathbf{NC} = \mathbf{CV}^c$ is generated by a set of minimal length words **NCM** which are related to, but are not, Christoffel words.

Also, we show that a special class of words of **NCM** are the limit case of the application of a theorem by Berstel and de Luca stating that Christoffel words are the balanced Lyndon words. It is well known that central words of Christoffel words are the limit case of the application of the Fine and Wilf theorem (see Fine and Wilf (1965)). Indeed, any non-trivial Christoffel word aub is such that u admits two periods p and q such that $|u| = p + q - 2$ and $\gcd(p, q) = 1$ but, in general, u does not admit period 1. In a similar way, we show that the *basic words* of **NCM**, which are Lyndon words but not Christoffel words, are *almost balanced* that is they admit exactly one pair of unbalanced factors. More precisely, we show that basic words of **NCM** are exactly the minimal words, with respect to the factorial order, having this property.

This paper is structured as follows. Section 2 introduces basic definitions and previous results. Section 3 defines the set of minimal non-convex words and Section 4 provides our general characterization of these words. Section 5 is dedicated to the link between minimal non-convex words and minimal almost balanced words. Section 6 generalizes to words over a four letter alphabet, that is any word coding the boundary of a polyomino. Finally Section 7 concludes briefly.

2. Preliminaries

We use the four letter alphabet $\{0, 1, \overline{0}, \overline{1}\}$ to encode polyominoes with the convention that its boundary is coded in a clockwise manner where 0 codes a step to the right, 1 codes a step up, $\overline{0}$ is for a step left and $\overline{1}$ for a step down. The word coding the boundary of a polyomino is called its *boundary word.*

Definition 1. *Given w the boundary word of a hv-convex polyomino P, a factor of ww is called a quadrant word if it is a non-expendable word written over a two letter alphabet.*

We refer the reader to Lothaire (1997) for basic notations and definitions of words. Throughout this paper, the term *Christoffel word* designates what some authors call *primitive lower Christoffel word.* We denote the set of all Christoffel words $\mathbf{C}$ and the set of Lyndon words $\mathbf{Lyn}$. Even though the concatenation of words is written in a multiplicative way, we use the notation $w = (x_1, x_2, \ldots, x_n)_\bigstar$ when w factorizes as $x_1 \cdot x_2 \cdots x_n$ according to a criterion $\bigstar$.

Notation 2. *Given a word w, its unique factorization as decreasing Lyndon words $w = l_1 l_2 \cdots l_m$ is noted $w = (l_1, l_2, \ldots, l_m)_{\mathbf{Lyn}}$.*

We recall the combinatorial characterization of digital convexity stated in Brlek et al. (2009):

Theorem 3 (Brlek et al. (2009), Proposition 7). *A polyomino is digitally convex if and only if each one of its quadrant words w is such that $w = (l_1, l_2, \ldots, l_m)_{\mathbf{Lyn}}$ with $l_i \in \mathbf{C}$ for all $1 \leq i \leq m$.*

From this, we define $\mathbf{CV} = \{(l_1, l_2, \ldots, l_m)_{\mathbf{Lyn}} \,|\, l_i \in \mathbf{C} \text{ for all } 1 \leq i \leq m\}$. For technical reasons, we assume that $\varepsilon \in \mathbf{CV}$.

2.1. Some Properties of Christoffel words

Since Borel and Laubie reinvestigated the Christoffel words (see Borel and Laubie (1993) and Christoffel (1875)) their impressive combinatorial structure has been studied by many authors. We present here only a few of these properties; we refer the reader to Berstel et al. (2009) for a comprehensive self-contained survey on Christoffel words. We consider Christoffel words over the two letter alphabet $\mathcal{A} = \{a, b\}$ with $a < b$, and this order is extended to words using lexicographic order.

Over $\mathcal{A}$, a word w is said to be *balanced* if for all $u, v \in \text{Factor}(w)$:

$$|u| = |v| \implies \delta(u, v) \leq 1,$$

where $\delta(u, v) = ||u|_a - |v|_a|$. Note that since we only consider a two letter alphabet, when u and v have same length, we also have $\delta(u, v) = ||u|_b - |v|_b|$. A pair u, v such that $\delta(u, v) > 1$ is called *unbalanced*. See Vuillon (2003) for a survey on balanced words and their generalizations.

Christoffel words are strongly related to the balance property. Indeed, the following results provide two different characterizations of Christoffel words using this notion.

Theorem 4 (Berstel and de Luca (1997), Theorem 3.2). *The set of Christoffel words is exactly the set of balanced Lyndon words.*

Theorem 5 (de Luca and Mignosi (1994), Corollary 1). *Given a word* $u \in \mathcal{A}^*$*, the words* aua, aub, bua *and* bub *are balanced if and only if* $aub \in \mathbf{C}$*.*

Consider the functions $P, S : \mathcal{A}^+ \to \mathbf{C}$ defined as follows: $P(w)$ (resp. $S(w)$) is the longest proper prefix (resp. proper suffix) of w that is a Christoffel word. These functions are well defined since $\{a, b\} \in \mathbf{C}$. Also, given $x, y \in \mathcal{A}^*$, define $x \oplus y = |x|_a|y|_b - |x|_b|y|_a$,

Theorem 6 (Borel and Laubie (1993)). *Given* $x, y \in \mathbf{C}$*, the word* $w = xy$ *is a Christoffel word if and only if* $x \oplus y = 1$*. In such case,* $x = P(w)$*,* $y = S(w)$*,* $u < w < v$ *and* $w = x \cdot y$ *is the only factorization of* w *as two Christoffel words.*

This is called the *standard decomposition* of w and is denoted $w = (x, y)_{\mathbf{C}}$. Theorem 7 gives another well known characterization of Christoffel words. First, define the functions G and D as follows:

$$\begin{aligned} G, D : \mathbf{C} \setminus \{a, b\} &\longrightarrow \mathbf{C} \\ G(u, v)_{\mathbf{C}} &= (u, uv)_{\mathbf{C}}, \\ D(u, v)_{\mathbf{C}} &= (uv, v)_{\mathbf{C}}. \end{aligned}$$

Theorem 7 (Borel and Laubie (1993); Berstel and de Luca (1997)). *A word* w *is a non-trivial Christoffel word if and only if there exists a unique sequence* $H_1, H_2, \ldots, H_k \in \{G, D\}$ *such that*

$$w = H_1 \circ H_2 \circ \cdots \circ H_k(a, b)_{\mathbf{C}}.$$

Given a Christoffel word $w = (u, v)_{\mathbf{C}}$, this last result imposes a strict structure on w in terms of the standard factorizations of u and v.

We conclude this section by recalling a nice combinatorial property of Christoffel words. Any non-trivial Christoffel word w may be written as $w = aw'b$, in such case the word w' is called a *central word*. Using the notation $\overline{}$ for the involution that maps a on b and b on a, the set of Christoffel word is closed by the application of $\overline{}$ to the central words.

Property 8 (de Luca and Mignosi (1994)). *Given any word $u \in \mathcal{A}^*$, $aub \in CP$ if and only if $a\overline{u}b \in \mathbf{C}$.*

This property is a direct consequence of Theorem 5 since $\delta(u, v) = \delta(\overline{u}, \overline{v})$ for all $u, v \in \mathcal{A}^*$ having same length.

3. Convex and non-convex words

Proposition 9 (Reutenauer (2008)). *The language* **CV** *is factorial.*

Proof. Let $w = xyz \in \mathbf{CV}$, we show that $y \in \mathbf{CV}$. Consider the Lyndon factorization $w = (l_1, l_2, \ldots, l_{n_w})_{\mathbf{Lyn}}$. Since w is a convex word, each l_i is a Christoffel word and by Theorem 4 it is balanced. There are two cases to consider.

– There exist $1 \leq k \leq n_w$ such that y is a factor of l_k. In this case, each factor of the Lyndon factorization of y is also a factor of l_k. Since the balance property is factorial, the Lyndon factorization of y contains only Christoffel words and $y \in \mathbf{CV}$.

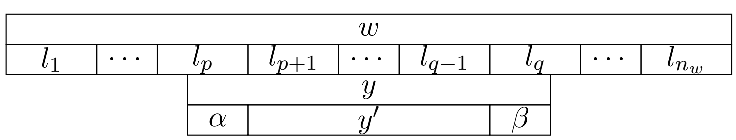

Figure 2: Illustration of the second case.

– There exist $p < q$ such that $y = \alpha l_{p+1} l_{p+2} \cdots l_{q-1} \beta$ where $\alpha \in \text{Suffix}(l_p)$ and $\beta \in \text{Prefix}(l_q)$, as illustrated in Figure 2. In such case, let $\alpha = (a_1, a_2, \ldots, a_{n_\alpha})_{\mathbf{Lyn}}$ and $\beta = (b_1, b_2, \ldots, b_{n_\beta})_{\mathbf{Lyn}}$. Since each word l_i is balanced, we have that each a_i and each b_i is a Christoffel words. Now, by construction we have

$$a_1 \geq a_2 \geq \cdots \geq a_{n_\alpha} \geq l_p \geq l_{p+1} \geq \cdots \geq l_{q-1} \geq l_q \geq b_1 \geq b_2 \geq \cdots \geq b_{n_\beta}.$$

We conclude that the unique factorization of y as decreasing Lyndon words is $y = (a_1, \ldots, a_{n_\alpha}, l_{p+1}, \ldots, l_{q-1}, b_1, \ldots, b_{n_\beta})_{\mathbf{Lyn}} \in \mathbf{CV}$.

□

From a geometrical point of view, Proposition 9 simply expresses the obvious fact that each part of the border of a convex shape is convex. On the other hand it implies that the language of *non-convex words* $\mathbf{NC} = \mathbf{CV}^c$ is an ideal of the monoid $\mathcal{A}^*$. A natural question is to identify the generators of this ideal. These generators are minimum non-convex words with respect to the factorial order, more precisely the set

$$\mathbf{NCM} = \{w \in \mathbf{NC} \,|\, \forall x \in \text{Factor}(w), u \neq w \implies u \in \mathbf{CV}\}.$$

By definition, a word $w \in \mathbf{NCM}$ cannot admit any other word of $\mathbf{NCM}$ as a proper factor. Moreover, note that on a two letter alphabet, all words of length smaller of equal to 3 are in $\mathbf{CV}$. Clearly, $\mathbf{NC}$ is an ideal of $\mathcal{A}^*$ generated by the set $\mathbf{NCM}$.

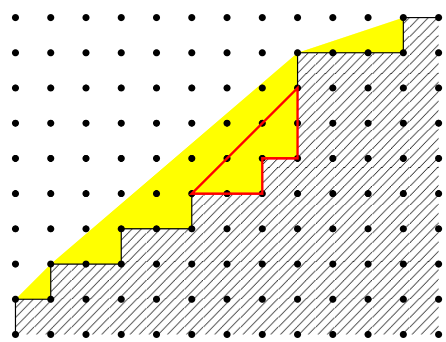

Figure 3: The North-West part of a polyomino with its convex hull in yellow, showing that it is not digitally convex. The part the boundary corresponding to the word $aababb \in$ $\mathbf{NCM}$ is highlight in red.

n	$\mathbf{NCM} \cap \mathcal{A}^n$
4	$\{aabb\}$
6	$\{aaabab, aababb, ababbb\}$
8	$\{aaaabaab, aabababb, abbabbbb\}$
9	$\{aaabaabab, ababbabbb\}$
10	$\{aaaaabaaab, aabaababab, aababababb, abababbabb, abbbabbbbb\}$

Table 1: Elements of $\mathbf{NCM}$ of length up to 10.

4. Characterization of NCM

This first result about the set **NCM** is a consequence of Theorem 3.

Lemma 10 (Reutenauer (2008)). NCM $\subset$ **Lyn**.

Proof. We argue by contradiction. Let $w \in$ **NCM** and suppose that $w \notin$ **Lyn**. Consider $(l_1, l_2, \ldots, l_m)_{\mathbf{Lyn}} = w$ with $m > 1$. Since w is not convex, by Theorem 3 there exists $1 \leq i \leq m$ such that $l_i \notin \mathbf{C}$. This word $l_i \notin \mathbf{CV}$ contradicting the factorial minimality of w. □

We can now establish a complete characterization of the set **NCM**.

Theorem 11. $\mathbf{NCM} = \{uw^kv \mid (u,v)_{\mathbf{C}} = w \in \mathbf{C}$ *and* $k \geq 1\}$.

In order to prove this Theorem, we need to introduce some combinatorial tools. The complete proof is given in Section 4.2

4.1. Left and right factorizations

In order to analyse the inner structure of a Christoffel word w, we introduce the two following factorizations, each being obtained by iteration of the standard factorization of a Christoffel word. The *right factorization* of w recursively decomposes its suffixes while the *left factorization* does the same for prefixes.

Definition 12. *Given* $w \in \mathbf{C} \setminus \{a, b\}$, *let* m *be the smallest integer such that* $S^{m+1}(w) = b$ *and for* $0 \leq k \leq m$, *let* $(u_k, v_k) = S^k(w)$. *The factorization* $w = u_0 \cdot u_1 \cdots u_m \cdot b$ *is called the* right factorization *of* w *and is denoted* $w = (u_0, u_1, \ldots, u_{m_1}, b)_R$

Definition 13. *Given* $w \in \mathbf{C} \setminus \{a, b\}$, *let* m' *be the smallest integer such that* $P^{m'+1}(w) = a$, *and for* $0 \leq k \leq m'$, *let* $(u_k, v_k) = P^{m'-k}(w)$. *The factorization* $w = a \cdot v_0 \cdot v_1 \cdots v_{m'}$ *is called the* left factorization *of* w *and is denoted* $w = (a, v_0, v_1, \ldots, v_{m'})_L$

By abuse of notation, we write $a = (a)_L$ and $b = (b)_R$. Table 2 shows how left and right factorizations of the word $w = aabaababaabab$ are obtained. In this example, both factorizations have the same length while it is not the case in general; for example: $aaaab = (a, aaab)_L = (a, a, a, a, b)_R$. Note that given a Christoffel word, both these factorization exist and are unique.

Property 14. *Let* $w \in \mathbf{C}$ *such that* $w = (u, v)$. *If for some* $x, y \in \mathbf{C}$:

Left factorization

<table>
<tr><td colspan="4">$(a\,a\,b\,a\,a\,b\,a\,b \quad , \; a\,a\,b\,a\,b)$</td></tr>
<tr><td colspan="3">u_2</td><td>v_2</td></tr>
<tr><td colspan="3">$(a\,a\,b \; , \; a\,a\,b\,a\,b)$</td><td></td></tr>
<tr><td colspan="2">u_1</td><td>v_1</td><td></td></tr>
<tr><td colspan="2">$(a, \; a\,b)$</td><td></td><td></td></tr>
<tr><td>a</td><td>v_0</td><td></td><td></td></tr>
</table>

Right factorization

<table>
<tr><td colspan="4">$(a\,a\,b\,a\,a\,b\,a\,b, \quad a\,a\,b\,a\,b)$</td></tr>
<tr><td>u_0</td><td colspan="3">v_0</td></tr>
<tr><td></td><td colspan="3">$(a\,a\,b, \quad a\,b)$</td></tr>
<tr><td></td><td>u_1</td><td colspan="2">v_1</td></tr>
<tr><td></td><td></td><td colspan="2">$(a\,, \; b)$</td></tr>
<tr><td></td><td></td><td>u_2</td><td>b</td></tr>
</table>

Table 2: The left factorization $(a, ab, aabab, aabab)_L$ and the right factorization $(aabaabab, aab, a, b)_R$ of w.

(i) $u = (x, y)_{\mathbf{C}}$ *then* $w = (xy, (xy)^k y)_{\mathbf{C}}$ *for some* $k \geq 0$ *and the inequality* $x < u < w < v \leq y$ *holds.*

(ii) $v = (x, y)_{\mathbf{C}}$ *then* $w = (x(xy)^k, xy)_{\mathbf{C}}$ *for some* $k \geq 0$ *and the inequality* $x \leq u < w < v < y$ *holds.*

Proof. Let $w = (u, v)_{\mathbf{C}} = ((x, y)_{\mathbf{C}}, v)_{\mathbf{C}}$ as in (i). By Theorem 7 there exist $H_1, H_2, \ldots, H_k \in \{G, D\}$ such that $w = H_1 \circ H_2 \circ \cdots \circ H_m(a, b)$. Since $u \neq a$, there exists i such that $H_i = D$. Let $k \geq 0$ be such that $w = G^k \circ D \circ H_{l+2} \circ \cdots \circ H_m(a, b)$ and let $(x, y)_{\mathbf{C}} = H_{k+2} \circ \cdots \circ H_m$, then

$$(u, v)_{\mathbf{C}} = G^k \circ D(x, y)_{\mathbf{C}} = (xy, (xy)^k y)_{\mathbf{C}}.$$

$x < u < w$ since x (resp. u) is a proper prefix of u (resp. w). On the other hand, $w < v$ since v is proper suffix of the Lyndon word w. Finally, it is clear that $v < y$ if $k \geq 1$ and $y = v$ if $k = 0$. One shows (ii) in a similar way. □

Since all Christoffel words are Lyndon words, the previous result implies a direct link between the standard factorization of a Christoffel word and the Lyndon factorization of its central word.

Corollary 15. *Given* $w = aub \in \mathbf{C}$ *with left and right factorizations*

$$w = aub = (a, v_0, v_1, \ldots, v_m)_L = (u_0, u_1, \ldots, u_{m'}, b)_R,$$

the words ub *and* au *factorize as follows*

$$ub = (v_0, v_1, \ldots, v_m)_{\mathbf{Lyn}} \text{ and } au = (u_0, u_1, \ldots, u_{m'})_{\mathbf{Lyn}}.$$

4.2. Proof of Theorem 11

In order to prove Theorem 11 we show the equality by inclusion on both sides.

Lemma 16. $\{uw^kv \,|\, (u,v)_{\mathbf{C}} = w \in \mathbf{C}$ *and* $k \geq 1\} \subseteq \mathbf{NCM}$.

Proof. Given a non-trivial Christoffel word $w = (u,v)_{\mathbf{C}}$ and an integer $k \geq 1$, we show that $z = uw^kv \in \mathbf{NCM}$. First, we show that $z \in \mathbf{Lyn} \setminus \mathbf{C}$. Let $p = |u|_b, q = |u|_a, r = |v|_b$ and $s = |v|_a$. Since the concatenation of u and v is a Christoffel word, we have that $u \oplus v = 1$. One checks that $uw^l \oplus w = 1$ for $0 \leq l \leq k$ so that $uw^k \in \mathbf{C}$. On the other hand $uw^k \oplus v = k+1 \geq 2$, implying $z \notin \mathbf{C}$. On the other hand, $z \in \mathbf{Lyn}$ since it is the concatenation of increasing Lyndon words.

Let $z = az'b$. Since $\mathbf{CV}$ is factorial, all that remains to show is that $az', z'b \in \mathbf{CV}$. Consider the right factorization of $v = (u_0, u_1, \ldots, u_m, b)_L$. In this factorization, each factor u_i is a Christoffel word and $w = (u,v)_{\mathbf{C}}$, $v = (u_0, v_0)_{\mathbf{C}}$, $v_0 = (u_1, v_1)_{\mathbf{C}}$, and so on. By Proposition 14, the following inequalities hold:

$$u_m \leq u_{m-1} \leq \cdots \leq u_0 \leq u \leq uw^k.$$

Thus, $az' = (uw^k, u_0, u_1, \ldots, u_m)_{\mathbf{Lyn}}$ and all these factors are Christoffel words, so $az' \in \mathbf{CV}$. One can checks that $z'b \in \mathbf{CV}$ in a similar way. □

Lemma 17. $\mathbf{NCM} \subseteq \{uw^kv \,|\, (u,v)_{\mathbf{C}} = w \in \mathbf{C}$ *and* $k \geq 1\}$

Proof. Let $z \in NCM$. Since z is a Lyndon word of length at least 4, there exists $z' \in \mathcal{A}^+$ such that $z = az'b$. Consider the Lyndon factorization of $z' = (l_1, l_2, \ldots, l_m)_{\mathbf{Lyn}}$. Since z' is a convex word, all of those l_i are Christoffel words.

For practical reasons, define $l_0 = a$ and $l_{m+1} = b$. Let $\alpha = al_1l_2\cdots l_p$ where $p = \max\{0 \leq i \leq m \,|\, al_1l_2\cdots l_i < l_{i+1}\} + 1$. Similarly, let $\beta = l_ql_{q+1}\ldots l_m$ where $q = \min\{1 \leq j \leq m+1 \,|\, l_{j-1} < l_jl_{j+1}\cdots l_mb\} - 1$. This construction is illustrated in Figure 4. Note that by definition of the Lyndon factorization, we have that $l_i \geq l_{i+1}$ for all $i \in \{1, 2, \ldots, m-1\}$. More precisely the minimality of p and the maximality of q imply that

$$l_p > l_{p+1} \text{ and } l_{q-1} > l_q. \tag{1}$$

From the above construction, we have $az' = (\alpha, l_{p+1}, \ldots, l_m)_{\mathbf{Lyn}}$ and $z'b = (a, l_1, l_2, \ldots, l_{q-1}, \beta)_{\mathbf{Lyn}}$. On the other hand, both words az' and $z'b$ are convex, implying that $\alpha, \beta \in \mathbf{C}$. So from Corollary 15,

$$\alpha = (a, l_1, l_2, \ldots, l_p)_L \text{ and } \beta = (l_q, l_{q+1}, \ldots, l_m)_R.$$

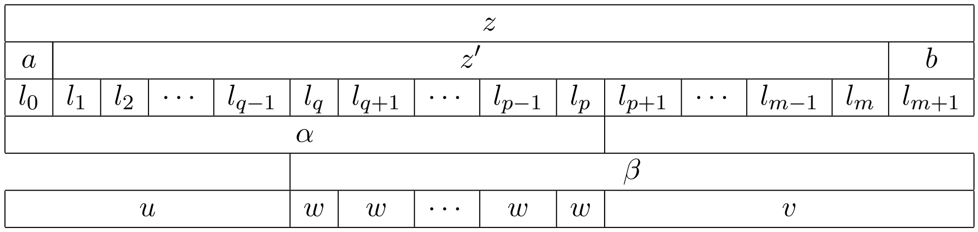

Figure 4: Factorization of z.

At this point, we claim that there exists $u, v, w \in \mathbf{C}$ such that $u = al_1l_2\cdots l_{q-1}$, $v = l_{p+1}l_{p+2}\cdots l_m b$ and $w = l_q = l_p = (u, v)_{\mathbf{C}}$ so that $z = uw^k v$ where $k = p - q + 1 \geq 1$. In order to prove this claim we proceed in three steps:

(*i*) $p \geq q$.

(*ii*) $u = al_1l_2\cdots l_{q-1} \in \mathbf{C}$ and $v = l_{p+1}l_{p+2}\cdots l_m b \in \mathbf{C}$.

(*iii*) $w = l_q = l_{q+1} = \cdots = l_p = uv$.

(*i*): We proceed by contradiction. Suppose $p < q$. In this case, one has that $z = (\alpha, l_{p+1}, \ldots, l_{q-1}, \beta)_{\mathbf{Lyn}}$ which contradicts the uniqueness of the Lyndon factorization since z is a Lyndon word.

(*ii*): By construction of the left factorization of α, we have $l_{i+1} = (al_1l_2\cdots l_{i-1}, l_i)_{\mathbf{C}}$ for all $i \in \{1, 2, \ldots, p-1\}$, so $u \in \mathbf{C}$. Similarly, the construction of the right factorization of β imply $l_{i-1} = (l_i, l_il_{i+1}\cdots l_m b)_{\mathbf{C}}$ for all $i \in \{q+1, q+2, \ldots, m\}$ so $v \in \mathbf{C}$.

(*iii*): Consider $(a, l_1, l_2, \ldots, l_p)_L$, the left factorization of α, Property 14 implies that for any $i \in \{1, 2, \ldots, p-1\}$, let $x = al_1l_2\cdots l_{i-1}$,

$$l_{i+1} = (xl_i)^k l_i \text{ with } k \geq 0 \implies \big(l_i > l_{i+1} \implies |l_i| < |l_{i+1}|\big). \tag{2}$$

Similarly, when considering the right factorization of β, Property 14 implies that for any $i \in \{q, q+2, \ldots, m-1\}$, letting $y = l_{i+2}\cdots l_m b$,

$$l_i = l_{i+1}(l_{i+1}y)^k \text{ with } k \geq 0 \implies \big(l_i > l_{i+1} \implies |l_i| > |l_{i+1}|\big). \tag{3}$$

Now consider any $i \in \{q, q+1, \ldots, p-1\}$. Clearly Equations (2) and (3) force that $l_i = l_{i+1}$, so $l_q = l_{q+1} = \cdots = l_p$. Finally, using Equations (1) and (2) one concludes that $P(l_q) = u$ and similarly, Equations (1) and (3) imply $S(l_p) = v$. □

5. Almost balanced words

Using the characterization obtained from theorem 11 we show a string link between minimal non-convex words and a new class of minimal words called *almost balanced*. We defined the set of almost balanced words as

$$\mathbf{AB} = \{w \in \mathcal{A}^* \mid \exists! \, \{u,v\} \in \text{Factor}(w) \text{ such that } |u| = |v| \text{ and } \delta(u,v) > 1\}$$

As in the case of non-convex words, among those words we focus our attention on the set minimal ones with respect to the factorial order and define the set of *minimal almost balanced* words:

$$\mathbf{ABM} = \{w \in \mathbf{AB} \mid \forall u \in \text{Factor}(w), u \neq w \implies u \notin \mathbf{AB}\}.$$

By analogy to the characterization of the words of **NCM** given in the previous section, we consider the words of the form $z = u^2v^2$ where $u, v, uv \in \mathbf{C}$. These words correspond to the case $z = uw^kz$ with $w = (u,v)_{\mathbf{C}}$ and $k = 1$, we call those words the *basic words of* **NCM**.

Theorem 18. $\mathbf{ABM} = \left\{z, \overline{z} \in \mathcal{A}^+ \mid z = u^2v^2 \textit{ where } u, v, uv \in \mathbf{C}\right\}$.

In order to show this, we need some extra results abouts balanced words. The proof of Theorem 18 appears in section 5.1

We may refine the balance property in order to consider only factor of a given length. Clearly all words are balanced when considering only factors of length one. In particular, every non-balance words have a specific maximum length of factors until which it is balanced.

Theorem 19 (Coven and Hedlund (1973), Lemma 3.06). *Given a word $w \in \mathcal{A}^*$ such that for some $n \geq 2$, for all $u, v \in Factor(w)$*

$$|u| = |v| < n \implies \delta(u,v) \leq 1,$$

but there exist $u, v \in Factor(w)$ such that $|u| = |v| = n$ and $\delta(u,v) > 1$. Then there exist a palindrome p of length $n-2$ such that $apa, bpb \in Factor(w)$.

This result allow to establish a general form for the words of **ABM**.

Lemma 20. *Given a word $w \in ABM$, there exist a palindrome p such that $w \in \{apabpb, bpbapa\}$.*

Proof. Let u, v be the unique pair of factors of w such that $|u| = |v|$ and $\delta(u, v) > 1$. By Theorem 19 there exist a palindrome p such that $u = apa$ and $v = bpb$ and so $\delta(u, v) = 2$. We show that both factors u and v must be consecutive in w, with no overlap. Without loss of generality, we assume that the factor u occurs before v in w.

– By contradiction, suppose there exist some non-empty word x such that uxv is a factor of w. In such case, $|ux| = |xv|$ and $\delta(ux, xv) = \delta(u, v) = 2$ but u, v is suppose to be the unique unbalanced pair. Contradiction.

– Again by contradiction, suppose the two factors u and v overlap in w, that is there exist a factor x of w such that $u \in \text{Prefix}(x)$ and $v \in \text{Suffix}(x)$ but $|x| < |u| + |v|$. Since the last letter of u is different from the first letter of v, there must be an overlap between the two occurrences of p. Call α this overlap, as shown in Figure 5, both word $a\alpha a$ and $b\alpha b$ appear in x.

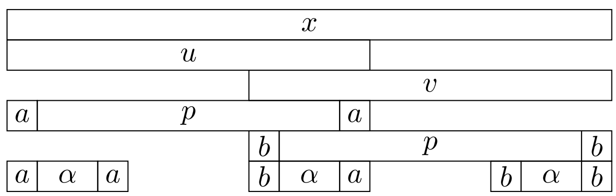

Figure 5: The overlap α of the two occurrences of p appears in the prefix αa and in the suffix $b\alpha$.

Again, we obtain a second pair of unbalanced factors: $|a\alpha a| = |b\alpha b|$ and $\delta(a\alpha a, b\alpha b) = 2$. Contradiction.

Therefore, w admits a factor of the form $apabpb$. Finally, since the words of **ABM** are minimal with respect to the factorial order, it must be that $w = apabpb$. □

We may now proceed with the proof of Theorem 18.

5.1. Proof of Theorem 18

Using the previous Theorem, we begin by providing a general form, closely related to Christoffel words, for the words of **ABM**. Then, equivalence between this general form and the one claimed in Theorem 18 is explicitly given by Lemma 22.

Lemma 21. $\mathbf{ABM} = \{apabpb, bpbapa \in \mathcal{A}^+ \mid apb \in \mathbf{C}\}$.

Proof. We start by showing the inclusion from left to right. Let $z \in \mathbf{ABM}$, by Lemma 20, there exist a palindrome p such that $z = apabpb$ (the case $w = bpbaba$ is similar). It remains to see that $apb \in \mathbf{C}$.

By Theorem 5 it suffices to see that apa, apb, bpa and bpb are all balanced. Using the fact that p is a palindrome and that both words $apabp$ and $pabpb$ are balanced, we have that for all $u, v \in \text{Factor}\{ap, pa, bp, pb\}$, $|u| = |v| \implies \delta(u, v) \leq 1$. Obviously, the four words apa, apb, bpa, bpb are balanced. This shows the first inclusion.

In order to show the inclusion from right to left, without loss of generality, let $z = bpbapa$ where $apb \in C$. Since $\delta(apa, bpb) = 2$ it remains to see that no other pair of factors of z are unbalanced. Let $u, v \in \text{Factor}(z)$ be such that $|u| = |v|$, there are two cases to consider:

- If $|u| = |v| < |p| + 2$, consider the standard factorization of $apb = (x, y)_{\mathbf{C}}$. By Theorem 7 the word $xapbapb$ is a Christoffel word and by Theorem 4 it is balanced. Also, by Theorem 5 apa and bpb are balanced words. Since $|u|$ and $|v|$ are smaller then $|p| + 2$ then both words u and v are factors of at least one of the words apa, bpb or $apbapb$, so $\delta(u, v) \leq 1$.

- If $|u| = |v| \geq |p| + 2$, then without loss of generality, assume that the factor u occurs before v in w. Let α, A, B, C, D be such that $u = B\alpha$, $v = \alpha C$ and $z = AB\alpha CD$, as shown in Figure 6.

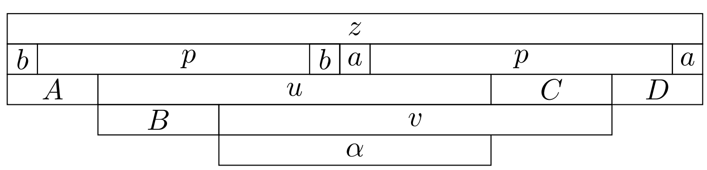

Figure 6: In the case where $|u| = |v| \geq |p| + 2$ an overlap α may occur.

In such case, we have $\delta(u, v) = \delta(B\alpha, \alpha C) = \delta(B, C)$. One easily checks that either $|\alpha| \geq 1$ and $|B| = |C| < |p| + 2$ implying that $\delta(u, v) \leq 1$, either $|\alpha| = 0$ and the only unbalanced pair of factors in z if $u = bpb$ and $v = apa$. This concludes the proof.

□

Finally, this last lemma only consider words of the form $z = apabpb$ since in the other case, let $z = bpbapa$ with $apb \in \mathbf{C}$ it suffices to consider $\overline{z} = a\overline{p}ab\overline{p}b$ and by Property 8, $a\overline{p}b \in \mathbf{C}$.

Lemma 22. *Given a non-trivial Christoffel word* $w = apb = (u, v)_{\mathbf{C}}$*, the following equality hold:* $apabpb = u^2v^2$.

Proof. First, let us consider the case where one of the words u or v is a trivial Christoffel word.

– If $u = a$ then $v = a^k b$ for some $k \geq 0$ and $u^2v^2 = apabpb$ where $p = a^k$.
– If $v = b$ then $u = ab^k$ for some $k \geq 0$ and $u^2v^2 = apabpb$ where $p = b^k$.

Now, if both u and v are non-trivial, we consider the central words of these Christoffel words. Let u', v' be such that $u = au'b$ and $v = av'b$. Since all three words u', v', p are palindromes, we have

$$p = u'10v' = v'01u'$$

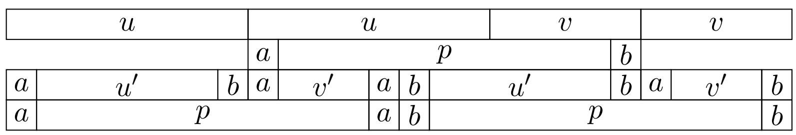

Figure 7: The palindromic structure of the centrals words ensures that $u^2v^2 = apabpb$.

As illustrated in Figure 7, the equality $u^2v^2 = apabpb$ hold.

□

6. NCM over general polyominoes

In order to extend **NCM** to words over the four letter alphabet $\{0, 1, \overline{0}, \overline{1}\}$, it suffices to notice that since a contour word cannot admit any factor of the set $\{0\overline{0}, \overline{0}0, 1\overline{1}, \overline{1}1\}$, any factor of a contour word that is written over a three letter alphabet *must* admit a sub-factor of the form $ab^k\overline{a}$ where $\{a, b\} \in \{\{0, 1\}, \{0, \overline{1}\}, \{\overline{0}, 1\}, \{\overline{0}, \overline{1}\}\}$. Since we assumed that the boundary word has been coded in a clockwise manner, the only non-convex words over more than two letters that do not admit any other non-convex word as a factor are of the form:

$$ab^k\overline{a} \text{ where } (a, b) \in \{(0, 1), (\overline{1}, 0), (\overline{0}, \overline{1}), (1, \overline{0})\} \text{ and } k \geq 1.$$

7. Conclusion

In this paper, we have provided a general form for the generators of the monoid of non-convex words that is uw^kv where $w = (u, v)_{\mathbf{C}}$ and $k \geq 1$.

Moreover, in the basic case, that is $k = 1$, we showed that those words are exactly the minimal words, with respect to the factorial order, among the *almost balanced words* that are the words admitting exactly one pair of unbalanced factors. More generally, a word $z = uw^kv$ where $w = (u, v)_{\mathbf{C}}$ admits exactly k pairs of unbalanced factors and in particular if $k = 0$ then $z = w \in \mathbf{C}$ and z is a balanced.

Finally, the above characterization of non-convexity highlights an important difference between Euclidean and discrete geometry. While Tietze's theorem (see Tietze (1929)) proves that in $\mathbb{R}^d$ convexity is a local property, the fact that **NCM** contains arbitrarily long words shows that it is not the case in discrete geometry. If one looks at a polyomino using only a finite window, it may always *seem* convex even if it is not.

Acknowledgments

The author is grateful to Christophe Reutenauer for the suggestion of this research and to Nicholas Chamandy for his useful comments.

U. Eckhardt, Digital lines and digital convexity, Digital and image geometry: advanced lectures (2001) 209–228.

S. Brlek, J.-O. Lachaud, X. Provençal, C. Reutenauer, Lyndon + Christoffel = digitally convex, Pattern Recognition 42 (10) (2009) 2239 – 2246, ISSN 0031-3203, doi:DOI: 10.1016/j.patcog.2008.11.010, selected papers from the 14th IAPR International Conference on Discrete Geometry for Computer Imagery 2008.

N. J. Fine, H. S. Wilf, Uniqueness theorems for periodic functions, Proc. Amer. Math. Soc. 16 (1965) 109–114, ISSN 0002-9939.

M. Lothaire, Combinatorics on words, Cambridge Mathematical Library, Cambridge University Press, Cambridge, ISBN 0-521-59924-5, 1997.

J.-P. Borel, F. Laubie, Quelques mots sur la droite projective réelle, J. Théor. Nombres Bordeaux 5 (1) (1993) 23–51, ISSN 1246-7405.

E. B. Christoffel, Observatio Arithmetica, Annali di Mathematica 6 (1875) 145–152.

J. Berstel, A. Lauve, C. Reutenauer, F. V. Saliola, Combinatorics on words, vol. 27 of *CRM Monograph Series*, American Mathematical Society, Providence, RI, ISBN 978-0-8218-4480-9, christoffel words and repetitions in words, 2009.

L. Vuillon, Balanced words, Bull. Belg. Math. Soc. Simon Stevin 10 (suppl.) (2003) 787–805, ISSN 1370-1444.

J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1-2) (1997) 171–203, ISSN 0304-3975.

A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, Theoret. Comput. Sci. 136 (2) (1994) 361–385, ISSN 0304-3975.

C. Reutenauer, Private communication, 2008.

E. M. Coven, G. A. Hedlund, Sequences with minimal block growth, Math. Systems Theory 7 (1973) 138–153, ISSN 0025-5661.

H. Tietze, Bemerkungenuber konvexe und nicht-konvexe Figuren, J. Reine Angew. Math 160 (222) (1929) 67–69.