Minimal Forbidden Words and Digital Lines

Gabriele Fici

16th Meeting on Tomography and Applications
Discrete Tomography, Neuroscience and Image Reconstruction

Milan, Italy, May 2–4, 2022

Let $\Sigma_d = \{0, 1, \cdots, d-1\}$ be an alphabet of cardinality d. A word over Σ_d is a sequence of letters from Σ_d. We will assume in this talk $d = 2$.

A word $w = w_1 \cdots w_{|w|}$ has period $p > 0$ if $w_i = w_j$ whenever $i = j \mod p$.

For example, the periods of the word $w = 0010010$ are 3, 6, $7 = |w|$ and every $p > |w|$.

Proposition

A word has period $p \leq |w|$ if and only if its prefix of length $|w| - p$ equals its suffix of length $|w| - p$, i.e., it has a border of length $|w| - p$.

Definition

A word is unbordered (aka bifix-free) if it has no border (nonempty factor appearing as a prefix and as a suffix). Equivalently, if its smallest period equals its length.

Definition

A word is primitive if it is not a power of another word.
That is, w is primitive if $w = v^n$ implies $n = 1$.

unbordered $\implies$ primitive.

primitive $\not\Longrightarrow$ unbordered (e.g. 010)

Definition

A central word is a word having two coprime periods p and q and length equal to $p+q-2$.

Central words are binary words. The first few central words are: ε, 0, 1, 00, 11, 000, 010, 101, 111, 0000, 1111, 00000, 00100, 01010, 10101, 11011, 11111, etc.

Every central word w is a palindrome, i.e., it coincides with its reversal $\widetilde{w}$.

Basics in Combinatorics on Words

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

For example, $010010 = (010)01(0) = (0)10(010)$.

Basics in Combinatorics on Words

Definition

Two words w and w' are conjugates if they are rotations of one another, i.e., there exist words u, v such that $w = uv$ and $w' = vu$.

The conjugacy class of a word w (aka necklace) has $|w|$ distinct elements if and only if w is primitive. In this case, the lexicographically smallest (for the order induced by $0 < 1$) word in the class is called a Lyndon word (and it is always unbordered).

Example

Let's write all the conjugates of 01001 in lexicographic order:

$$00101$$
$$01001$$
$$01010$$
$$10010$$
$$10100$$

Proposition

A word is a conjugate of its reversal if and only if it is a concatenation of two palindromes.

Proof.

If $w = uv$ and $\widetilde{w} = vu$ then $\widetilde{w} = \widetilde{uv} = \tilde{v}\,\tilde{u} = vu$, hence $v = \tilde{v}$ and $u = \tilde{u}$. Conversely, if $w = uv$, $u = \tilde{u}$, $v = \tilde{v}$, then $\widetilde{w} = \widetilde{uv} = \tilde{v}\,\tilde{u} = vu$. $\square$

Christoffel Words

Proposition

Let C be a central word. Then the words $0C1$ and $1C0$ are conjugates, since they can be written as concatenations of two palindromes.

Proof.

If C is a power of a letter, the claim is immediate. Otherwise, $C = P01Q = Q10P$, with P, Q palindromes. Hence, $0C1 = 0P0 \cdot 1Q1$ and $1C0 = 1Q1 \cdot 0P0$. □

The words $0C1$ and $1C0$ are called, respectively, primitive (lower and upper) Christoffel words.

For example, the central word $C = 010$ yields the primitive Christoffel words 00101 and 10100.

We also consider words of length 1 to be primitive Christoffel words.

Christoffel Words

Definition

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

We have:

$$\text{primitive Christoffel} = \text{balanced} + \text{unbordered}$$

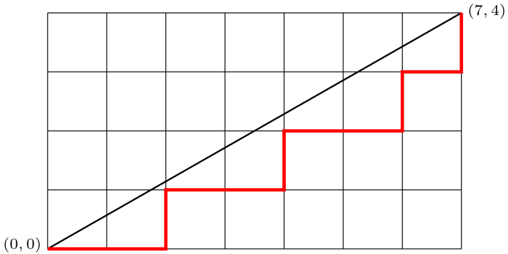
In fact, the set of primitive lower Christoffel words is precisely the set of balanced Lyndon words.

Christoffel Words

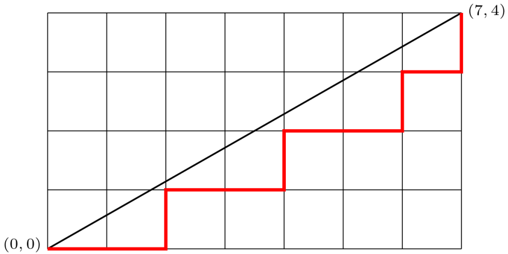
Definition

Given a pair of natural numbers (a, b), the lower (resp. upper) (a, b)-Christoffel word is the digital approximation from below (resp. from above) of the Euclidean segment joining $(0, 0)$ to (a, b). It has slope b/a.

For example, the lower $(7, 4)$-Christoffel word is 00100100101.



For example, the lower $(7, 4)$-Christoffel word is 00100100101.



Another way to obtain the primitive lower Christoffel word $0C1$ of slope b/a is by constructing C in the following way: take the positive multiples of a and b smaller than ab, sort them and write 1 or 0 accordingly:

	4	**7**	8	12	**14**	16	20	**21**	24	
0	0	1	0	0	1	0	0	1	0	1

Christoffel Words

Let $w_{a,b}$ (resp. $W_{a,b}$) be the lower (resp. upper) (a,b)-Christoffel word.

Some remarks:

1. $w_{a,b} = 0C1$ for a palindrome C; C is central iff $w_{a,b}$ is primitive iff a,b are coprime;
2. If $a' = na$, $b' = nb$, then $w_{a',b'} = (w_{a,b})^n$ and $W_{a',b'} = (W_{a,b})^n$;
3. $W_{a,b} = \widetilde{w_{a,b}} = 1C0$ is the reversal of $w_{a,b}$ and is conjugated to it;
4. The length of $w_{a,b}$ (resp. $W_{a,b}$) is $a+b$ (there are a 1s and b 0s);

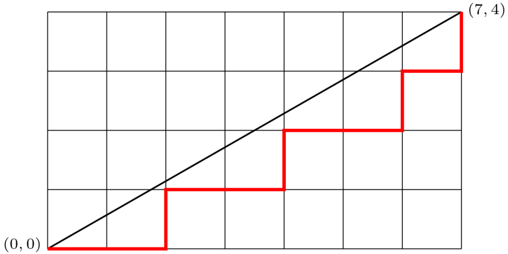
Christoffel Words

Let $a, b > 0$. The lower Christoffel word $w_{a,b} = w_1 w_2 \cdots w_{a+b}$ can be defined by

$$w_i = \begin{cases} 0 & \text{if } ib > (i-1)b, \text{ mod}(a+b) \\ 1 & \text{if } ib < (i-1)b, \text{ mod}(a+b) \end{cases}$$

Example

Let $a = 7$ and $b = 4$.
We have $\{i4 \text{ mod}(11) \mid i = 1, 2, \ldots, 11\} = \{4, 8, 1, 5, 9, 2, 6, 10, 3, 7, 0\}$.
Hence, $w_{7,4} = 00100100101$.



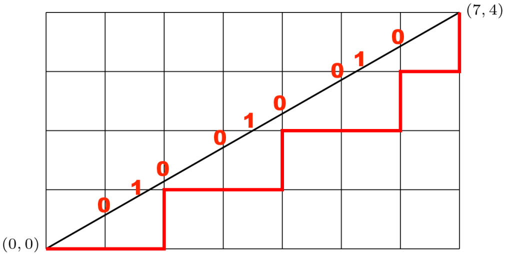
Christoffel Words

If $w_{a,b}$ is a primitive lower Christoffel word (i.e., a and b are coprime) of length > 1, $w_{a,b} = 0C1$ for a central word C.

The central word C encodes the intersections of the Euclidean segment from $(0,0)$ to (a,b) with the grid (0=vertical, 1=horizontal).

Example

Let $a = 7$ and $b = 4$.
$w_{7,4} = 00100100101 = 0 \cdot 010010010 \cdot 1$.



Christoffel Words

Let $a, b > 0$. Consider the $(a+b) \times (a+b)$ matrix $\mathcal{A}_{a,b}$ 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 $a+b$.

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

Christoffel Words

Let $a, b > 0$. Consider the $(a+b) \times (a+b)$ matrix $\mathcal{A}_{a,b}$ 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 $a+b$.

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

Christoffel Words

Let $a, b > 0$. Consider the $(a+b) \times (a+b)$ matrix $\mathcal{A}_{a,b}$ 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 $a+b$.

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

Christoffel Words

Let $a, b > 0$. Consider the $(a+b) \times (a+b)$ matrix $\mathcal{A}_{a,b}$ 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 $a+b$.

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

The first row is the lower Christoffel word $w_{a,b}$. Every row is obtained from the previous one by swapping a 01 factor with 10. The last row is the upper Christoffel word $W_{a,b}$.

Christoffel Words

Let $a, b > 0$. Consider the $(a+b) \times (a+b)$ matrix $\mathcal{A}_{a,b}$ 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 $a+b$.

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

The first row is the lower Christoffel word $w_{a,b}$. Every 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 lexicographically ordered.

Christoffel Words

Definition

A word is circularly balanced if all its conjugates are balanced.

We have:

Proposition

A word is a conjugate of a Christoffel word (not necessarily primitive) if and only if it is circularly balanced.

Decompositions of Christoffel Words

Let $0C1$ be a primitive lower Christoffel word, hence a balanced Lyndon word.

If the central word C is not a power of a single letter, we can write $C = P01Q = Q10P$, with P and Q central words.

Hence, we have the following factorizations:

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

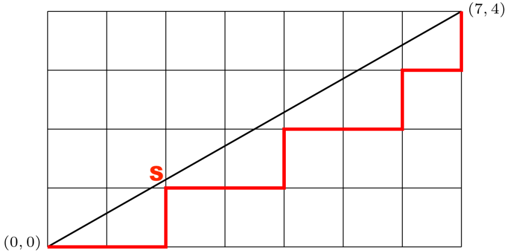
If instead $C = 0^n$ (the case $C = 1^n$ is analogous) we have:

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

Decompositions of Christoffel Words

The standard factorization divides a primitive lower Christoffel word in two shorter primitive lower Christoffel words.

It determines the point S closest to the Euclidean segment.

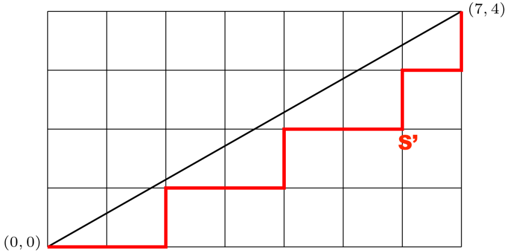


$0Q1 \cdot 0P1 = 001 \cdot 00100101$

Decompositions of Christoffel Words

The palindromic factorization, instead, divides a primitive lower Christoffel word in two palindromes.

It determines the point S' farthest from the Euclidean segment.



$0P0 \cdot 1Q1 = 00100100 \cdot 101$

Special Balanced Words

Actually, *any* balanced word with a 0s and b 1s is a digital approximation of the Euclidean segment from $(0, 0)$ to (a, b). Indeed, all these words are "between" the lower and the upper (a, b)-Christoffel Word.

For example, if $a = 5$ and $b = 3$ we have the 8 words in the conjugacy class of the primitive lower Christoffel word 00100101 and 4 other non circularly balanced words: 00101010, 01010100, 10001001, 10010001.

As a non-coprime example, if $a = 4$ and $b = 2$ we have the 3 words in the conjugacy class of the lower Christoffel word 001001 and 5 other non circularly balanced words: 001010, 010001, 010100, 100001, 100010.

Problem

Given a and b, how many balanced word are there with a 0s and b 1s?

Special Balanced Words

Definition

A balanced word v is right special (resp. left special) if both $v0$ and $v1$ are balanced (resp. if both $0v$ and $1v$ are balanced).

A balanced word is bispecial if it is both left and right special.

Theorem (F., 2014)

A balanced word v is bispecial if and only if $0v1$ is a lower Christoffel word.

Actually, 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 strictly bispecial, that is, all of $0v1$, $1v0$, $0v0$, $1v1$ are balanced (Berstel, de Luca, 1997).

Example

Let $0v1$ be the Christoffel word $0 \cdot 0100 \cdot 1$. The word 0100 is bispecial but not strictly biscpecial. 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.

Minimal Forbidden Words

Definition

Let L be a language (finite or infinite set of words) closed under taking factors (factorial). 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 $\mathsf{MF}(L)$ denote the set of minimal forbidden words of L. A word $w = avb \in \mathsf{MF}(L)$ if and only if

1. $avb \notin L$;
2. $av, vb \in L$.

Minimal Forbidden Words

A special case is when L is the set of factors of a single word w. In this case we talk of minimal forbidden words of the word w.

Example

Let $w = 01001$. The minimal forbidden words (MFWs) of w are:

$$\mathsf{MF}(w) = \{000, 0010, 101, 11\}.$$

Theorem

There is a bijection between factorial languages and their sets of minimal forbidden words.

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

Let now Bal be the set of balanced words.

Theorem (F., 2014)

$MF(Bal) = \{bwa \mid \{a, b\} = \{0, 1\}$, awb is a non-primitive Chr. word$\}$.

Example

000101 is not balanced, but all its proper factors are. Indeed, 100100 is the square of the primitive upper Christoffel word 100.

Example

000100101 is not balanced, but all its proper factors are. Indeed, 100100100 is the cube of the primitive upper Christoffel word 100.

Corollary (F., 2014)

For every $n > 0$, there are exactly $n - \phi(n) - 1$ words in $MF(Bal)$ that start with 0, and they are all Lyndon words.

Minimal Forbidden Words

In 2011, Provençal studied the language of minimal almost balanced words, MABs, i.e., minimal words with the property that there exists a unique pair of unbalanced factors.

Example

000101 is almost balanced with unique unbalanced pair $000, 101$, but all its proper factors are balanced. Hence it is MAB.

Example

000100101 is not MAB, since $000, 101$ and $000100, 100101$ are distinct pairs of unbalanced factors.

Theorem (Provençal, 2011)

*$MAB = \{u^2v^2, \widetilde{u^2v^2} \mid$
uv is the standard decomposition of a primitive lower Christoffel word$\}$.*

Minimal Forbidden Words

Theorem

MAB $= \{bwa \mid \{a,b\} = \{0,1\}$, awb *is the square of a primitive Chr. word*$\} \subseteq$ *MF*(Bal).

Proof.

Let $uv = x = 0C1$ be a standard decomposition. Let $u = 0Q1$, $v = 0P1$ for P, Q central words, so that $x = 0Q10P1 = 0P01Q1$. Then $u^2v^2 = 0Q10Q10P10P1 = 0Q10P01Q10P1$, hence $1C0 = 1Q10P01Q10P0 = \widetilde{x}^2$. The other cases are similar. □

In other words, if aCb is a primitive Christoffel word, and $C = PabQ = QbaP$, then we have:

1. $aPabQb$ is a Christoffel word;
2. $aQabPb$ is a MAB word.

From straight lines to convex lines

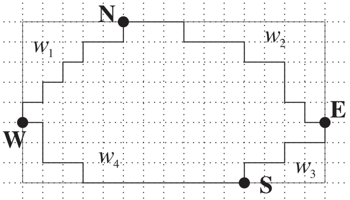
We saw that balanced words are good approximations of segments in the plane.

We now discuss which words are good approximations of *convex lines* in the plane.

Digitally Convex Words

Given a convex figure in the plane, we can digitalize it by considering its intersections with the grid $\mathbb{Z} \times \mathbb{Z}$.

We then separate the binary sequence coding this intersections in 4 parts: WN, NE, ES, SW.



W is the lowest on the Left side;
N is the leftmost on the Top side;
E is the highest on the Right side;
S is the rightmost on the Bottom side;
So that $w \equiv w_1 w_2 w_3 w_4$.

Definition

A word is WN digitally convex (or, simply, digitally convex) if it is a factor of a word that encodes a WN word.

In the figure, the WN word is $w_1 = 1010101001$.

Digitally Convex Words

Theorem (Chen, Fox, Lyndon, 1958)

Any word factorizes uniquely in non-increasing Lyndon words. This factorization is called the Lyndon factorization of w.

Example

Let $w = 01000010000010000100000001$. The Lyndon factorization of w is

$$01 \cdot 00001 \cdot 00000100001 \cdot 0000001.$$

Example

Let $w = 1100$. The Lyndon factorization of w is

$$1 \cdot 1 \cdot 0 \cdot 0.$$

Digitally Convex Words

Theorem (Brlek, Lachaud, Provençal, Reutenauer, 2009)

w is digitally convex if and only if all the Lyndon words in the Lyndon factorization of w are balanced (hence primitive lower Christoffel).

Example

Let $w = 0101001001$. The Lyndon factorization of w is

$$01 \cdot 01 \cdot 001 \cdot 001.$$

Therefore w is digitally convex.

Notice that a digitally convex word is not necessarily balanced, e.g., $1100 = 1 \cdot 1 \cdot 0 \cdot 0$.

Digitally Convex Words

Let DC be the set of digitally convex words. Any word in DC starts with 0 or it is a power of 1 concatenated with a word in DC starting with 0.

Theorem

The number of digitally convex words starting with 0 is given by the Euler transform of the Euler totient function ϕ (sequence A061255 in OEIS)

$$|DC(n)| = \frac{1}{n}\sum_{k=1}^{n}|DC(n-k)|\sum_{d|k} d\phi(d)$$

n	1	2	3	4	5	6	7	8	9	10	11	12
$\lvert$DC$(n)\rvert$	1	1	2	4	7	13	21	37	60	98	157	251

Minimal Forbidden Words of Digitally Convex Words

The set of digitally convex words is a factorial language.

In 2011, Provençal studied the minimal forbidden words of the set DC of digitally convex word.

Theorem (Provençal, 2011)

MF(*DC*) $= \{u(uv)^k v \mid$
$k \geq 1$, uv *is the standard factorization of a primitive lower Chris. word.*$\}$

Theorem

MF(*DC*) *is the set of minimal forbidden words starting with* 0 *of balanced words. Hence,*
MF(*DC*) $= \{0w1 \mid 1w0$ *is a non-primitive Christoffel word*$\} \subseteq$ *Lyn*

Minimal Forbidden Words of Digitally Convex Words

Example

1. Start with two coprime numbers, e.g., $4, 7$;
2. Build an upper $(k4, k7)$-Christoffel word, $k > 1$, e.g., $k = 2$: $10100100100 \cdot 10100100100$;
3. Swap the first and the last letter: 0010010010010100100101;
4. The word so obtained is a non-balanced Lyndon word; all its factors are digitally convex.

Sesquipowers of Christoffel Words

An interesting class of digitally convex words is given by the words of the form $w^n w'$, where w is a primitive lower Christoffel word and w' is a prefix of w. These words are called sesquipowers (or fractional powers) of primitive lower Christoffel words.

Let $\mathsf{SC}(n)$ be the set of sesquipowers of primitive lower Christoffel words in lexicographic order, e.g. $\mathsf{SC}(5) =$

$\{00000, 00001, 00010, 00100, 00101, 01010, 01011, 01101, 01110, 01111, 11111\}$

Let $\mathcal{F}(n)$ be the nth Farey sequence (pairs of coprime integers a, b between 0 and n) in increasing order, e.g.

$$\mathcal{F}(5) = \left\{\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}\right\}$$

There is a bijection between $\mathsf{SC}(n)$ and $\mathcal{F}(n)$, preserving the orders, given by

$$w^n w' \mapsto \frac{|w|_1}{|w|}$$

Sesquipowers of Christoffel Words

Definition (F., Lipták, 2011)

A word is prefix normal if no factor has more 0s than the prefix of the same length.

Example

001100 is prefix normal, while 00101001001 is not (00100 has more 0s than 00101).

Theorem

A word is a sesquipower of a primitive lower Christoffel word if and only if it is balanced and prefix normal.

THANK YOU