



Unveiling the Connection Between the Lyndon Factorization and the Canonical Inverse Lyndon Factorization via a Border Property

Paola Bonizzoni¹ ✉ 


Dip. di Informatica, Sistemistica e Comunicazione, University of Milano-Bicocca, Milan, Italy

Clelia De Felice² ✉ 

Dip. di Informatica, University of Salerno, Fisciano, Italy

Brian Riccardi ✉ 

Dip. di Informatica, Sistemistica e Comunicazione, University of Milano-Bicocca, Milan, Italy

Rocco Zaccagnino ✉ 

Dip. di Informatica, University of Salerno, Fisciano, Italy

Rosalba Zizza ✉ 

Dip. di Informatica, University of Salerno, Fisciano, Italy

Abstract

The notion of Lyndon word and Lyndon factorization has shown to have unexpected applications in theory as well as in developing novel algorithms on words. A counterpart to these notions are those of inverse Lyndon word and inverse Lyndon factorization. Differently from the Lyndon words, the inverse Lyndon words may be bordered. The relationship between the two factorizations is related to the inverse lexicographic ordering, and has only been recently explored. More precisely, a main open question is how to get an inverse Lyndon factorization from a classical Lyndon factorization under the inverse lexicographic ordering, named CFL_{in} . In this paper we reveal a strong connection between these two factorizations where the border plays a relevant role. More precisely, we show two main results. We say that a factorization has the border property if a nonempty border of a factor cannot be a prefix of the next factor. First we show that there exists a unique inverse Lyndon factorization having the border property. Then we show that this unique factorization with the border property is the so-called canonical inverse Lyndon factorization, named ICFL. By showing that ICFL is obtained by compacting factors of the Lyndon factorization over the inverse lexicographic ordering, we provide a linear time algorithm for computing ICFL from CFL_{in} .

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words; Mathematics of computing → Combinatorial algorithms

Keywords and phrases Lyndon words, Lyndon factorization, Combinatorial algorithms on words

Digital Object Identifier 10.4230/LIPIcs.MFCS.2024.31

Funding *Paola Bonizzoni*: from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement PANGAIA No. 872539, and from MUR 2022YRB97K, PINC, Pangenome INformatiCs: from Theory to Applications.

Clelia De Felice: from MUR 2022YRB97K, PINC, Pangenome INformatiCs: from Theory to Applications.

Brian Riccardi: from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement PANGAIA No. 872539, and from MUR 2022YRB97K, PINC, Pangenome INformatiCs: from Theory to Applications.

Rocco Zaccagnino: from INdAM-GNCS Project 2023.

Rosalba Zizza: from INdAM-GNCS Project 2023.

¹ Corresponding author

² Corresponding author



1 Introduction

The theoretical investigation of combinatorial properties of well-known word factorizations is a research topic that recently have witnessed special interest especially for improving the efficiency of algorithms [5]. Among these, the *Lyndon Factorization* introduced by Chen, Fox, Lyndon in [12], named CFL, undoubtedly stands. Any word w admits a unique factorization $\text{CFL}(w)$, that is a lexicographically non-increasing sequence of factors which are *Lyndon words*. A Lyndon word w is strictly lexicographically smaller than each of its proper cyclic shifts, or, equivalently, than each of its nonempty proper suffixes [24]. Interesting applications of the use of the Lyndon factorization and Lyndon words are the development of the bijective Burrows-Wheeler Transforms [2, 6, 21] and a novel algorithm for sorting suffixes [5]. In particular, the notion of a Lyndon word has been re-discovered various times as a theoretical tool to locate short motifs [15] and relevant k -mers in bioinformatics applications [26]. In this line of research, Lyndon-based word factorizations have been explored to define a novel feature representation for biological sequences based on theoretical combinatorial properties proved to capture sequence similarities [7].

The notion of a *Lyndon word* has a counterpart that is the notion of an *inverse Lyndon word*, i.e., a word lexicographically greater than its suffixes. Inverting the relation between a word and its suffixes, as between Lyndon words and inverse Lyndon words, leads to different properties. Indeed, although a word could admit more than one *inverse Lyndon factorization*, that is a factorization into a nonincreasing product of *inverse Lyndon words*, in [8] the *Canonical Inverse Lyndon Factorization*, named ICFL, was introduced. ICFL maintains the main properties of CFL: it is unique and can be computed in linear time. In addition, it maintains a similar *Compatibility Property*, used for obtaining the sorting of the suffixes of w (“global suffixes”) by using the sorting of the suffixes of each factor of $\text{CFL}(w)$ (“local suffixes”) [25]. Most notably, $\text{ICFL}(w)$ has another interesting property [8, 9, 10]: we can provide an upper bound on the length of the longest common prefix of two substrings of a word w starting from different positions.

A relationship between $\text{ICFL}(w)$ and $\text{CFL}(w)$ has been proved by using the notion of *grouping* [8]. First, let $\text{CFL}_{in}(w)$ be the Lyndon factorization of w with respect to the inverse lexicographic order, it is proved that $\text{ICFL}(w)$ is obtained by concatenating the factors of a non-increasing maximal chain with respect to the prefix order, denoted by \mathcal{PMC}_w , in $\text{CFL}_{in}(w)$ (see Section 6). Despite this result, the connection between $\text{CFL}_{in}(w)$ and the inverse Lyndon factorization still remained obscure, mainly by the fact that a word may have multiple inverse Lyndon factorizations.

In this paper, we explore this connection between CFL_{in} and the inverse Lyndon factorizations. Our first main contribution consists in showing that there is a unique inverse Lyndon factorization of a word that has *border property*. The border property states that any nonempty border of a factor cannot be a prefix of the next factor. We further highlight the aforementioned connection by proving that the inverse Lyndon factorization with the border property is a *compact* factorization (Definition 6.7), i.e., each inverse Lyndon factor is the concatenation of *compact factors*. In turn, a compact factor is the concatenation of the longest sequence of identical words in a \mathcal{PMC} . We then show the second contribution of this paper: this unique factorization is ICFL itself and then provide a simpler linear time algorithm for computing ICFL. Our algorithm is based on a new property that characterizes $\text{ICFL}(w)$: the last factor in $\text{ICFL}(w)$ is the longest suffix of w that is an inverse Lyndon word. Recall that the Lyndon factorization of w has a similar property: the last factor is the longest suffix of w that is a Lyndon word.

2 Words

Throughout this paper we follow [4, 13, 22, 23, 27] for the notations. We denote by Σ^* the *free monoid* generated by a finite alphabet Σ and we set $\Sigma^+ = \Sigma^* \setminus \{1\}$, where 1 is the empty word. For a word $w \in \Sigma^*$, we denote by $|w|$ its *length*. A word $x \in \Sigma^*$ is a *factor* of $w \in \Sigma^*$ if there are $u_1, u_2 \in \Sigma^*$ such that $w = u_1xu_2$. If $u_1 = 1$ (resp. $u_2 = 1$), then x is a *prefix* (resp. *suffix*) of w . A factor (resp. prefix, suffix) x of w is *proper* if $x \neq w$. Two words x, y are *incomparable* for the prefix order, denoted as $x \not\asymp y$, if neither x is a prefix of y nor y is a prefix of x . Otherwise, x, y are *comparable* for the prefix order. We write $x \leq_p y$ if x is a prefix of y and $x \geq_p y$ if y is a prefix of x . The notion of a pair of words comparable (or incomparable) for the suffix order is defined symmetrically.

We recall that, given a nonempty word w , a *border* of w is a word which is both a proper prefix and a suffix of w [14]. The longest proper prefix of w which is a suffix of w is also called *the border* of w [14, 23]. A word $w \in \Sigma^+$ is *bordered* if it has a nonempty border. Otherwise, w is *unbordered*. A nonempty word w is *primitive* if $w = x^k$ implies $k = 1$. An unbordered word is primitive. A *sesquipower* of a word x is a word $w = x^n p$ where p is a proper prefix of x and $n \geq 1$. Two words x, y are called *conjugate* if there exist words u, v such that $x = uv, y = vu$. The conjugacy relation is an equivalence relation. A conjugacy class is a class of this equivalence relation.

► **Definition 2.1.** Let $(\Sigma, <)$ be a totally ordered alphabet. The *lexicographic* (or *alphabetic order*) \prec on $(\Sigma^*, <)$ is defined by setting $x \prec y$ if

- x is a proper prefix of y , or
- $x = ras, y = rbt, a < b$, for $a, b \in \Sigma$ and $r, s, t \in \Sigma^*$.

In the next part of the paper we will implicitly refer to totally ordered alphabets. For two nonempty words x, y , we write $x \ll y$ if $x \prec y$ and x is not a proper prefix of y [3]. We also write $y \succ x$ if $x \prec y$. Basic properties of the lexicographic order are recalled below.

► **Lemma 2.2.** For $x, y \in \Sigma^+$, the following properties hold.

- (1) $x \prec y$ if and only if $zx \prec zy$, for every word z .
- (2) If $x \ll y$, then $xu \ll yv$ for all words u, v .
- (3) If $x \prec y \prec xz$ for a word z , then $y = xy'$ for some word y' such that $y' \prec z$.
- (4) If $x \ll y$ and $y \ll z$, then $x \ll z$.

Let t, j, r_j be positive integers, with $1 \leq j \leq t$. Let $\mathcal{S}_1, \dots, \mathcal{S}_t$ be sequences, with $\mathcal{S}_j = (s_{j,1}, \dots, s_{j,r_j})$. We let $(\mathcal{S}_1, \dots, \mathcal{S}_t)$ stand for the sequence $(s_{1,1}, \dots, s_{1,r_1}, \dots, s_{t,1}, \dots, s_{t,r_t})$.

3 Lyndon words

► **Definition 3.1.** A *Lyndon word* $w \in \Sigma^+$ is a word which is primitive and the smallest one in its conjugacy class for the lexicographic order.

► **Example 3.2.** Let $\Sigma = \{a, b\}$ with $a < b$. The words $a, b, aaab, abbb, aabab$ and $aababaabb$ are Lyndon words. On the contrary, $abab, aba$ and $abaab$ are not Lyndon words.

► **Proposition 3.3.** Each Lyndon word w is unbordered.

A class of conjugacy is also called a *necklace* and often identified with the minimal word for the lexicographic order in it. We will adopt this terminology. Then a word is a necklace if and only if it is a power of a Lyndon word. A *prenecklace* is a prefix of a necklace. Then

clearly any nonempty prenecklace w has the form $w = (uv)^k u$, where uv is a Lyndon word, $u \in \Sigma^*$, $v \in \Sigma^+$, $k \geq 1$, that is, w is a sesquipower of a Lyndon word uv . The following result has been proved in [16]. It shows that the nonempty prefixes of Lyndon words are exactly the nonempty prefixes of the powers of Lyndon words with the exclusion of c^k , where c is the maximal letter and $k \geq 2$.

► **Proposition 3.4.** *A word is a nonempty prefix of a Lyndon word if and only if it is a sesquipower of a Lyndon word distinct of c^k , where c is the maximal letter and $k \geq 2$.*

In the following $L = L_{(\Sigma^*, <)}$ will be the set of Lyndon words, totally ordered by the relation $<$ on $(\Sigma^*, <)$.

► **Theorem 3.5.** *Any word $w \in \Sigma^+$ can be written in a unique way as a nonincreasing product $w = \ell_1 \ell_2 \cdots \ell_h$ of Lyndon words, i.e., in the form*

$$w = \ell_1 \ell_2 \cdots \ell_h, \text{ with } \ell_j \in L \text{ and } \ell_1 \succeq \ell_2 \succeq \cdots \succeq \ell_h \quad (3.1)$$

The sequence $\text{CFL}(w) = (\ell_1, \dots, \ell_h)$ in Eq. (3.1) is called the *Lyndon decomposition* (or *Lyndon factorization*) of w . It is denoted by $\text{CFL}(w)$ because Theorem 3.5 is usually credited to Chen, Fox and Lyndon [12]. The following result, proved in [16], is necessary for our aims.

► **Corollary 3.6.** *Let $w \in \Sigma^+$, let ℓ_1 be its longest prefix which is a Lyndon word and let w' be such that $w = \ell_1 w'$. If $w' \neq 1$, then $\text{CFL}(w) = (\ell_1, \text{CFL}(w'))$.*

Sometimes we need to emphasize consecutive equal factors in CFL. We write $\text{CFL}(w) = (\ell_1^{n_1}, \dots, \ell_r^{n_r})$ to denote a tuple of $n_1 + \dots + n_r$ Lyndon words, where $r > 0$, $n_1, \dots, n_r \geq 1$. Precisely $\ell_1 \succ \dots \succ \ell_r$ are Lyndon words, also named *Lyndon factors* of w . There is a linear time algorithm to compute the pair (ℓ_1, n_1) and thus, by iteration, the Lyndon factorization of w [17, 23]. Linear time algorithms may also be found in [16] and in the more recent paper [19].

4 Inverse Lyndon words

For the material in this section see [8, 9, 10]. Inverse Lyndon words are related to the inverse alphabetic order. Their definition is recalled below.

► **Definition 4.1.** *Let $(\Sigma, <)$ be a totally ordered alphabet. The inverse $<_{in}$ of $<$ is defined by*

$$\forall a, b \in \Sigma \quad b <_{in} a \Leftrightarrow a < b$$

The inverse lexicographic or inverse alphabetic order on $(\Sigma^*, <)$, denoted $<_{in}$, is the lexicographic order on $(\Sigma^*, <_{in})$.

► **Example 4.2.** Let $\Sigma = \{a, b, c, d\}$ with $a < b < c < d$. Then $dab \prec dabd$ and $dabda \prec dac$. We have $d <_{in} c <_{in} b <_{in} a$. Therefore $dab \prec_{in} dabd$ and $dac \prec_{in} dabda$.

Of course for all $x, y \in \Sigma^*$ such that $x \bowtie y$,

$$y \prec_{in} x \Leftrightarrow x \prec y.$$

Moreover, in this case $x \ll y$. This justifies the adopted terminology.

From now on, $L_{in} = L_{(\Sigma^*, <_{in})}$ denotes the set of the Lyndon words on Σ^* with respect to the inverse lexicographic order. Following [18], a word $w \in L_{in}$ will be named an *anti-Lyndon word*. Correspondingly, an *anti-prenecklace* will be a prefix of an *anti-necklace*, which in turn will be a necklace with respect to the inverse lexicographic order.

In the following, we denote by $\text{CFL}_{in}(w)$ the Lyndon factorization of w with respect to the inverse order $<_{in}$.

► **Definition 4.3.** A word $w \in \Sigma^+$ is an inverse Lyndon word if $s \prec w$, for each nonempty proper suffix s of w .

► **Example 4.4.** The words $a, b, aaaaa, bbba, baaab, bbaba$ and $bbababbaa$ are inverse Lyndon words on $\{a, b\}$, with $a < b$. On the contrary, $aaba$ is not an inverse Lyndon word since $aaba \prec ba$. Analogously, $aabba \prec ba$ and thus $aabba$ is not an inverse Lyndon word.

The following result, proved in [8, 10], and also in [11], summarizes some properties of the inverse Lyndon words.

► **Proposition 4.5.** Let $w \in \Sigma^+$. Then we have

1. The word w is an anti-Lyndon word if and only if it is an unbordered inverse Lyndon word.
2. The word w is an inverse Lyndon word if and only if w is a nonempty anti-prenecklace.
3. If w is an inverse Lyndon word, then any nonempty prefix of w is an inverse Lyndon word.

► **Definition 4.6.** An inverse Lyndon factorization of a word $w \in \Sigma^+$ is a sequence (m_1, \dots, m_k) of inverse Lyndon words such that $m_1 \cdots m_k = w$ and $m_i \ll m_{i+1}$, $1 \leq i \leq k-1$.

As the following example in [8] shows, a word may have different inverse Lyndon factorizations.

► **Example 4.7.** Let $\Sigma = \{a, b, c, d\}$ with $a < b < c < d$, $z = dabdadacddbdc$. It is easy to see that $(dab, dadacd, db, dc)$, $(dabda, dac, ddbdc)$, $(dab, dadac, ddbdc)$ are all inverse Lyndon factorizations of z .

5 The border property

In this section we prove the main result of this paper, namely, for any nonempty word w , there exists a unique inverse Lyndon factorization of w which has a special property, named the border property.

► **Definition 5.1** (Border property). Let $w \in \Sigma^+$. A factorization (m_1, \dots, m_k) of w has the border property if each nonempty border z of m_i is not a prefix of m_{i+1} , $1 \leq i \leq k-1$.

We first prove a fundamental property of the inverse Lyndon factorizations of w which have the border property.

► **Lemma 5.2.** Let $w \in \Sigma^+$, let (m_1, \dots, m_k) be an inverse Lyndon factorization of w having the border property. If α is a nonempty border of m_j , $1 \leq j \leq k-1$, then there exists a nonempty prefix β of m_{j+1} such that $|\beta| \leq |\alpha|$ and $\alpha \ll \beta$.

Proof. Let $w \in \Sigma^+$, let (m_1, \dots, m_k) be an inverse Lyndon factorization of w having the border property, let α be a nonempty border of m_j , $1 \leq j \leq k-1$. We distinguish two cases: either $|m_{j+1}| < |\alpha|$ or $|m_{j+1}| \geq |\alpha|$.

Assume $|m_{j+1}| < |\alpha|$. By hypothesis (m_1, \dots, m_k) is an inverse Lyndon factorization, hence $m_j \ll m_{j+1}$, that is, there are $r, s, t \in \Sigma^*$, $a, b \in \Sigma$, such that $a < b$ and $m_j = ras$, $m_{j+1} = rbt$. Obviously $|ra| \leq |m_{j+1}| < |\alpha|$, thus there is $s' \in \Sigma^*$ such that $\alpha = ras'$. Consequently, $\alpha = ras' \ll rbt = m_{j+1}$ and our claim holds with $\beta = m_{j+1}$.

Assume $|m_{j+1}| \geq |\alpha|$. Let β be the nonempty prefix of m_{j+1} such that $|\beta| = |\alpha|$. Clearly $\beta \neq \alpha$ because (m_1, \dots, m_k) has the border property. Since α and β are two different nonempty words of the same length, either $\beta \ll \alpha$ or $\alpha \ll \beta$. The first case leads to a

contradiction because if $\beta \ll \alpha$, then $m_{j+1} \ll m_j$ by Lemma 2.2 and this contradicts the fact that (m_1, \dots, m_k) is an inverse Lyndon factorization. Thus, $\alpha \ll \beta$ and the proof is complete. \blacktriangleleft

► **Proposition 5.3.** *For each $w \in \Sigma^+$, there exists a unique inverse Lyndon factorization of w having the border property.*

Proof. The proof is by induction on $|w|$. If $|w| = 1$, then the statement clearly holds. Thus assume $|w| > 1$. Let $F_1(w) = (f_1, \dots, f_k)$ and $F_2(w) = (f'_1, \dots, f'_v)$ be two inverse Lyndon factorizations of w having the border property. Thus

$$f_1 \cdots f_k = f'_1 \cdots f'_v = w \quad (5.1)$$

If $|f_k| = |f'_v|$ and $v = 1$ or $k = 1$, clearly $f_k = f'_v$ and $F_1(w) = F_2(w)$. Analogously, if $|f_k| = |f'_v|$, $v > 1$ and $k > 1$, then $f_k = f'_v$ and $F_1'(w') = (f_1, \dots, f_{k-1})$, $F_2'(w') = (f'_1, \dots, f'_{v-1})$ would be two inverse Lyndon factorizations of w' having the border property, where w' is such that $w = w'f_k$. Of course, $|w'| < |w|$. By induction hypothesis, $F_1'(w') = F_2'(w')$, hence $F_1(w) = F_2(w)$.

By contradiction, let $|f_k| \neq |f'_v|$. Assume $|f_k| < |f'_v|$ (similar arguments apply if $|f_k| > |f'_v|$). The word f_k is a proper suffix of f'_v . Clearly $k > 1$. Let g be the smallest integer such that $f_{g+1} \cdots f_k$ is a proper suffix of f'_v , $1 \leq g \leq k - 1$, that is,

$$f'_v = \alpha f_{g+1} \cdots f_k \quad (5.2)$$

where $\alpha \in \Sigma^+$ is a suffix of f_g .

Notice that

$$\alpha \not\ll f_{g+1}. \quad (5.3)$$

Indeed, if $\alpha \ll f_{g+1}$, then, by Eq. (5.2), we would have $f'_v = \alpha f_{g+1} \cdots f_k \ll f_{g+1} \cdots f_k$, which is impossible because f'_v is an inverse Lyndon word.

The word α is a nonempty proper suffix of f_g since otherwise we would have $\alpha = f_g \ll f_{g+1}$, contrary to Eq. (5.3). Since f_g is an inverse Lyndon word and α is a nonempty proper suffix of f_g , either $\alpha \leq_p f_g$ or $\alpha \ll f_g$.

If $\alpha \leq_p f_g$, then α is a nonempty border of f_g , then, by Lemma 5.2, there exists a nonempty prefix β of f_{g+1} such that $|\beta| \leq |\alpha|$ and $\alpha \ll \beta$. Thus, $\alpha \ll f_{g+1}$ which contradicts Eq. (5.3). Assume $\alpha \ll f_g$. Since $f_g \ll f_{g+1}$, by Lemma 2.2 we have $\alpha \ll f_{g+1}$ which contradicts once again Eq. (5.3). This finishes the proof. \blacktriangleleft

► **Example 5.4.** Let $\Sigma = \{a, b, c, d\}$ with $a < b < c < d$, let $z = dabdadacddbdc$. Notice that only the last one of the inverse Lyndon factorizations of z from Example 4.7 fulfils the border property, and the others do not.

6 Groupings and compact factorizations

In this section we prove a structural property of an inverse Lyndon factorization having the border property, namely it is a *compact factorization*. This result is crucial to characterize the relationship between $\text{CFL}_{in}(w)$ and the factorization into inverse Lyndon words of w . First we report the notion of *grouping* given in [8]. We refer to [8, 10] for a detailed and complete discussion on this topic.

Let $\text{CFL}_{in}(w) = (\ell_1, \dots, \ell_h)$, where $\ell_1 \succeq_{in} \ell_2 \succeq_{in} \dots \succeq_{in} \ell_h$. Consider the partial order \geq_p , where $x \geq_p y$ if y is a prefix of x . Recall that a *chain* is a set of a pairwise comparable elements. We say that a chain is maximal if it is not strictly contained in any other chain. A

non-increasing (*maximal*) chain in $\text{CFL}_{in}(w)$ is the sequence corresponding to a (maximal) chain in the multiset $\{\ell_1, \dots, \ell_h\}$ with respect to \geq_p . We denote by \mathcal{PMC}_w , or simply \mathcal{PMC} when it is understood, a non-increasing maximal chain in $\text{CFL}_{in}(w)$. Looking at the definition of the (inverse) lexicographic order, it is easy to see that a \mathcal{PMC} is a sequence of consecutive factors in $\text{CFL}_{in}(w)$. Moreover $\text{CFL}_{in}(w)$ is the concatenation of its \mathcal{PMC} . The formal definitions are given below.

► **Definition 6.1.** Let $w \in \Sigma^+$, let $\text{CFL}_{in}(w) = (\ell_1, \dots, \ell_h)$ and let $1 \leq r < s \leq h$. We say that $\ell_r, \ell_{r+1}, \dots, \ell_s$ is a non-increasing maximal chain for the prefix order in $\text{CFL}_{in}(w)$, abbreviated \mathcal{PMC} , if $\ell_r \geq_p \ell_{r+1} \geq_p \dots \geq_p \ell_s$. Moreover, if $r > 1$, then $\ell_{r-1} \not\geq_p \ell_r$, if $s < h$, then $\ell_s \not\geq_p \ell_{s+1}$. Two \mathcal{PMC} $\mathcal{C}_1 = \ell_r, \ell_{r+1}, \dots, \ell_s$, $\mathcal{C}_2 = \ell_{r'}, \ell_{r'+1}, \dots, \ell_{s'}$ are consecutive if $r' = s + 1$ (or $r = s' + 1$).

► **Definition 6.2.** Let $w \in \Sigma^+$, let $\text{CFL}_{in}(w) = (\ell_1, \dots, \ell_h)$. We say that $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ is the decomposition of $\text{CFL}_{in}(w)$ into its non-increasing maximal chains for the prefix order if the following holds

- (1) Each \mathcal{C}_j is a non-increasing maximal chain in $\text{CFL}_{in}(w)$.
- (2) \mathcal{C}_j and \mathcal{C}_{j+1} are consecutive, $1 \leq j \leq s - 1$.
- (3) $\text{CFL}_{in}(w)$ is the concatenation of the sequences $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s$.

► **Definition 6.3.** Let $w \in \Sigma^+$. We say that (m_1, \dots, m_k) is a grouping of $\text{CFL}_{in}(w)$ if the following holds

- (1) (m_1, \dots, m_k) is an inverse Lyndon factorization of w
- (2) Each factor m_j , is the product of consecutive factors in a \mathcal{PMC} in $\text{CFL}_{in}(w)$.

► **Example 6.4.** Let $\Sigma = \{a, b, c, d\}$, $a < b < c < d$, and $w = dabadabdabdadac$. We have $\text{CFL}_{in}(w) = (daba, dab, dab, dadac)$. The decomposition of $\text{CFL}_{in}(w)$ into its \mathcal{PMC} is $((daba, dab, dab), (dadac))$. Moreover, $(daba, dab, dadac)$ is a grouping of $\text{CFL}_{in}(w)$ but for the inverse Lyndon factorization $(dabadab, dab, dadac)$ this is no longer true.

Next, let $y = dabadabdabdabdadac$. We have $\text{CFL}_{in}(y) = (daba, dab, dab, dab, dadac)$. The decomposition of $\text{CFL}_{in}(y)$ into its \mathcal{PMC} is $((daba, dab, dab, dab), (dadac))$. Moreover, $(daba, (dab)^3, dadac)$ and $(dabadab, (dab)^2, dadac)$ are two groupings of $\text{CFL}_{in}(y)$.

For our aims, we need to consider the words that are concatenations of equal factors in CFL_{in} . This approach leads to a refinement of the partition of CFL_{in} into non-increasing maximal chains for the prefix order, as defined below.

► **Definition 6.5 (Compact sequences).** Let $\mathcal{C} = (\ell_1, \dots, \ell_h)$ be a non-increasing maximal chain for the prefix order in $\text{CFL}_{in}(w)$. The decomposition of \mathcal{C} into maximal compact sequences is the sequence $(\mathcal{G}_1, \dots, \mathcal{G}_n)$ such that

- (1) $\mathcal{C} = (\mathcal{G}_1, \dots, \mathcal{G}_n)$
- (2) For every i , $1 \leq i \leq n$, \mathcal{G}_i consists of the longest sequence of consecutive identical elements in \mathcal{C}

Let $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ be the decomposition of $\text{CFL}_{in}(w)$ into its non-increasing maximal chains for the prefix order. The decomposition of $\text{CFL}_{in}(w)$ into its maximal compact sequences is obtained by replacing each \mathcal{C}_j in $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ with its decomposition into maximal compact sequences.

► **Definition 6.6 (Compact factor).** Let $(\mathcal{G}_1, \dots, \mathcal{G}_n)$ be the decomposition of $\text{CFL}_{in}(w)$ into its maximal compact sequences. For every i , $1 \leq i \leq n$, the concatenation g_i of the elements in \mathcal{G}_i is a compact factor in $\text{CFL}_{in}(w)$.

► **Definition 6.7** (Compact factorization). *Let $w \in \Sigma^+$. We say that (m_1, \dots, m_k) is a compact factorization of w if (m_1, \dots, m_k) is an inverse Lyndon factorization of w and each m_j , $1 \leq j \leq k$, is a concatenation of compact factors in $\text{CFL}_{in}(w)$.*

► **Example 6.8.** Consider again $y = \text{dabadabdabdadbac}$ over $\Sigma = \{a, b, c, d\}$, $a < b < c < d$, as in Example 6.4. The decomposition of $\text{CFL}_{in}(y) = (\text{daba}, \text{dab}, \text{dab}, \text{dab}, \text{dadac})$ into its maximal compact sequences is $((\text{daba}), (\text{dab}, \text{dab}, \text{dab}), (\text{dadac}))$. The compact factors in $\text{CFL}_{in}(w)$ are $\text{daba}, (\text{dab})^3, \text{dadac}$. Moreover, $(\text{daba}, (\text{dab})^3, \text{dadac})$ is a compact factorization whereas $(\text{dabadab}, (\text{dab})^2, \text{dadac})$ is a grouping of $\text{CFL}_{in}(y)$ which is not a compact factorization.

► **Proposition 6.9.** *Let $w \in \Sigma^+$. If (m_1, \dots, m_k) is an inverse Lyndon factorization of w having the border property, then (m_1, \dots, m_k) is a compact factorization of w .*

Proof. Let $w \in \Sigma^+$, let (m_1, \dots, m_k) be an inverse Lyndon factorization of w having the border property. Let $\text{CFL}_{in}(w) = (\ell_1, \dots, \ell_h)$, where $\ell_1 \succeq_{in} \ell_2 \succeq_{in} \dots \succeq_{in} \ell_h$ and ℓ_1, \dots, ℓ_h are anti-Lyndon words. First we prove that (m_1, \dots, m_k) is a grouping of $\text{CFL}_{in}(w)$ by induction on $|w|$. If $|w| = 1$ the statement clearly holds, thus assume $|w| > 1$.

The words m_1 and ℓ_1 are comparable for the prefix order, hence either m_1 is a proper prefix of ℓ_1 or ℓ_1 is a prefix of m_1 . Suppose that m_1 is a proper prefix of ℓ_1 . Thus, there are j , $1 < j \leq k$, and $x, y \in \Sigma^*$, $x \neq 1$, such that $m_j = xy$ and $\ell_1 = m_1 \dots m_{j-1}x$. Necessarily it turns out $j = 2$ because otherwise $m_1 \ll m_{j-1}$, hence, by Lemma 2.2, $\ell_1 \ll m_{j-1}x$ and this contradicts the fact that ℓ_1 is an anti-Lyndon word. In conclusion $\ell_1 = m_1x$ and $m_2 = xy$. We know that $m_1 \ll m_2$, that is, there are $r, s, t \in \Sigma^*$, $a, b \in \Sigma$, such that $a < b$ and $m_1 = ras$, $m_2 = rbt = xy$. If $|x| \leq |r|$, then x is a nonempty border of ℓ_1 and if $|x| > |r|$, then there is a word t' such that $x = rbt'$ which implies $\ell_1 \ll x$. Both cases again contradict the fact that ℓ_1 is an anti-Lyndon word.

Therefore, ℓ_1 is a prefix of m_1 . If $m_1 = \ell_1 \dots \ell_h = w$, then $k = 1$ and the statement is proved. Otherwise, let i be the largest integer such that $m_1 = \ell_1 \dots \ell_{i-1}x$, $x, y \in \Sigma^*$, $\ell_i = xy$, $1 < i \leq h$, $y \neq 1$. Let $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ be the decomposition of $\text{CFL}_{in}(w)$ into its non-increasing maximal chains for the prefix order. We claim that $\ell_1 \dots \ell_{i-1}$ is a prefix of the concatenation of the elements of \mathcal{C}_1 , thus $(\ell_1, \dots, \ell_{i-1})$ is a chain for the prefix order. If $i = 1$ we are done. Let $i > 1$. By contradiction, assume that there is j , $1 < j < i$, such that $\ell_j \notin \mathcal{C}_1$. Therefore, $\ell_1 \ll \ell_j$ which implies $m_1 \ll \ell_j \dots \ell_{i-1}x$ and this contradicts the fact that m_1 is an inverse Lyndon word.

We now prove that $x = 1$. Assume $x \neq 1$. As a preliminary step, we prove that there is no nonempty prefix β of m_2 such that $|\beta| \leq |x|$ and $x \ll \beta$. In fact, if such a prefix existed, there would be $r, s, t \in \Sigma^*$, $a, b \in \Sigma$, such that $a < b$ and $x = ras$, $\beta = rbt$. Notice that y is a nonempty prefix of $m_2 \dots m_k$, thus y and $\beta = rbt$ are comparable for the prefix order. If $|\ell_i| = |xy| \leq |xr|$, then $0 < |y| \leq |r|$ and y would be a nonempty prefix of r . Thus y would be a nonempty border of ℓ_i . If $|\ell_i| = |xy| > |xr|$, then there would be a word t' such that $\ell_i = rasrbt'$ which would imply $\ell_i \ll rbt'$. Both cases contradict the fact that ℓ_i is an anti-Lyndon word.

Now either ℓ_i is a prefix of ℓ_1 or $\ell_1 \ll \ell_i$. If ℓ_i were a prefix of ℓ_1 , then x would be a nonempty border of m_1 . By Lemma 5.2 there would exist a nonempty prefix β of m_2 such that $|\beta| \leq |x|$ and $x \ll \beta$ which contradicts our preliminary step.

If it were true that $\ell_1 \ll \ell_i$ then there would be $r, s, t \in \Sigma^*$, $a, b \in \Sigma$, such that $a < b$ and $\ell_1 = ras$, $\ell_i = rbt = xy$. If $|x| > |r|$, then there would be a word t' such that $x = rbt'$ which would imply $m_1 \ll x$ and this contradicts the fact that m_1 is an inverse Lyndon word. If $|x| \leq |r|$, then x would be a prefix of r and x would be a nonempty border of m_1 . By Lemma 5.2 again, there would exist a nonempty prefix β of m_2 such that $|\beta| \leq |x|$ and $x \ll \beta$ which contradicts again our preliminary step.

Let $w' \in \Sigma^*$ be such that $w = m_1 w'$. We know that $w' \neq 1$ and clearly $|w'| < |w|$. Of course (m_2, \dots, m_k) is an inverse Lyndon factorization of w having the border property. Moreover, by Corollary 3.6, $\text{CFL}_{in}(w') = (\ell_i, \dots, \ell_h)$ and $(\mathcal{C}'_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ is the decomposition of $\text{CFL}_{in}(w')$ into its non-increasing maximal chains for the prefix order, where \mathcal{C}'_1 is defined by $\mathcal{C}_1 = (\ell_1, \dots, \ell_{i-1}, \mathcal{C}'_1)$. By induction hypothesis, (m_2, \dots, m_k) is a grouping of $\text{CFL}_{in}(w')$ and consequently (m_1, \dots, m_k) is a grouping of $\text{CFL}_{in}(w)$.

Finally, to obtain a contradiction, suppose that (m_1, \dots, m_k) is a grouping of $\text{CFL}_{in}(w)$ having the border property such that (m_1, \dots, m_k) is not a compact factorization of w . To adapt the notation to the proof, set $\text{CFL}_{in}(w) = (\ell_1^{n_1}, \dots, \ell_r^{n_r})$, where $r > 0$, $n_1, \dots, n_r \geq 1$ and ℓ_1, \dots, ℓ_r are anti-Lyndon words. By Definitions 6.3 and 6.7, there exist integers j, h, p_h, q_h , $1 \leq j \leq k-1$, $1 \leq h \leq r$, $p_h \geq 1$, $q_h \geq 1$, $p_h + q_h \leq n_h$, such that m_j ends with $\ell_h^{p_h}$ and m_{j+1} starts with $\ell_h^{q_h}$. Thus, by Definition 6.3, ℓ_h is a prefix of m_j . Moreover, ℓ_h is a proper prefix of m_j . Indeed otherwise $\ell_h = m_j \leq_p m_{j+1}$ which is impossible because $m_j \ll m_{j+1}$ ((m_1, \dots, m_k) is an inverse Lyndon factorization). Thus ℓ_h is a nonempty border of m_j . The word ℓ_h is also a prefix of m_{j+1} and this contradicts the fact that (m_1, \dots, m_k) has the border property. \blacktriangleleft

7 The canonical inverse Lyndon factorization: The algorithm

In this section we state another relevant result of the paper related to the main one stated in Section 5. We have shown that a nonempty word w can have more than one inverse Lyndon factorization but w has a unique inverse Lyndon factorization with the border property (Example 4.7, Proposition 5.3). Below we highlight that this unique factorization is the canonical one defined in [8, 10].

This special inverse Lyndon factorization is denoted by ICFL because it is the counterpart of the Lyndon factorization CFL of w , when we use (I)inverse words as factors. Indeed, in [8] it has been proved that $\text{ICFL}(w)$ can be computed in linear time and it is uniquely determined for a word w .

In Proposition 7.7 we show another interesting property of ICFL: the last factor of the factorization is the longest suffix that is an inverse Lyndon word. Based on this result we provide a new simpler linear algorithm for computing ICFL.

We begin by recalling previously proved results on ICFL, namely Proposition 7.7 in [8] and Proposition 9.5 in [10]. They are merged into Proposition 7.1.

► **Proposition 7.1.** *For any $w \in \Sigma^+$, $\text{ICFL}(w)$ is a grouping of $\text{CFL}_{in}(w)$. Moreover, $\text{ICFL}(w)$ has the border property.*

Corollary 7.2 is a direct consequence of Propositions 5.3, 6.9 and 7.1.

► **Corollary 7.2.** *For each $w \in \Sigma^+$, $\text{ICFL}(w)$ is a compact factorization and it is the unique inverse Lyndon factorization of w having the border property.*

Since $\text{ICFL}(w)$ is the unique inverse Lyndon factorization with the border property, from now on these two notions will be synonymous. Proposition 7.3 has been proved in [11].

► **Proposition 7.3.** *Let $w \in \Sigma^+$, let $\text{CFL}_{in}(w) = (\ell_1, \dots, \ell_h)$ and let $(\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s)$ be the decomposition of $\text{CFL}_{in}(w)$ into its non-increasing maximal chains for the prefix order. Let w_1, \dots, w_s be words such that $\text{CFL}_{in}(w_j) = \mathcal{C}_j$, $1 \leq j \leq s$. Then $\text{ICFL}(w)$ is the concatenation of the sequences $\text{ICFL}(w_1), \dots, \text{ICFL}(w_s)$, that is,*

$$\text{ICFL}(w) = (\text{ICFL}(w_1), \dots, \text{ICFL}(w_s)) \quad (7.1)$$

31:10 Unveiling the Connection Between CFL and ICFL via a Border Property

We can now state some results useful to prove the correctness of our algorithm. First we observe that, thanks to Corollary 7.2 and Proposition 7.3, to compute ICFL we can limit ourselves to the case in which CFL_{in} is a chain with respect to the prefix order.

► **Lemma 7.4.** *Let ℓ_1, \dots, ℓ_h be anti-Lyndon words over Σ that form a non-increasing chain for the prefix order, that is, $\ell_1 \geq_p \ell_2 \geq_p \dots \geq_p \ell_h$. If $\ell_1 \neq \ell_2$, then $\ell_1 \not\prec_p \ell_2 \cdots \ell_h$.*

Proof. By contradiction, assume that ℓ_1 is a prefix of $\ell_2 \cdots \ell_h$. Then, $\ell_1 = \ell_2 \cdots \ell_t z$ where either $z = 1$ and $2 < t \leq h$ or z is a nonempty prefix of ℓ_{t+1} , $2 \leq t < h$. Thus either ℓ_t or z is a nonempty border of ℓ_1 , a contradiction in both cases. ◀

► **Remark 7.5.** [10] Let x, y be two different borders of $w \in \Sigma^+$. If x is shorter than y , then x is a border of y .

► **Proposition 7.6.** *Let $w \in \Sigma^+$ and assume that $\text{CFL}_{in}(w)$ form a non-increasing chain for the prefix order. If (m_1, \dots, m_k) is a factorization of w such that each m_j , $1 \leq j \leq k$, is a concatenation of compact factors in $\text{CFL}_{in}(w)$, then (m_1, \dots, m_k) has the border property.*

Proof. Let $w \in \Sigma^+$ and assume that $\text{CFL}_{in}(w)$ form a non-increasing chain for the prefix order. Let (m_1, \dots, m_k) be a factorization of w such that each m_j , $1 \leq j \leq k$, is a concatenation of compact factors in $\text{CFL}_{in}(w)$. The proof is by induction on k . If $k = 1$, then the conclusion follows immediately. Assume $k > 1$.

Let $w' \in \Sigma^+$ be such that $w = m_1 w'$. It is clear that (m_2, \dots, m_k) is a factorization of w' such that each m_j , $2 \leq j \leq k$, is a concatenation of compact factors in $\text{CFL}_{in}(w')$. Thus, by the induction hypothesis, (m_2, \dots, m_k) has the border property. It remains to prove that each nonempty border of m_1 is not a prefix of m_2 . The proof is straightforward if m_1 is unbordered, thus assume that m_1 is bordered.

Let $\text{CFL}_{in}(w) = (\ell_1^{n_1}, \dots, \ell_r^{n_r})$, where $\ell_1^{n_1}, \dots, \ell_r^{n_r}$ are the compact factors in $\text{CFL}_{in}(w)$, that is, ℓ_1, \dots, ℓ_r are anti-Lyndon words such that $\ell_1 \geq_p \dots \geq_p \ell_r$. Since m_1 is a concatenation of compact factors in $\text{CFL}_{in}(w)$, there is h , $1 \leq h < r$ such that

$$m_1 = \ell_1^{n_1} \cdots \ell_h^{n_h}.$$

Notice that ℓ_h is a nonempty border of m_1 . Furthermore, since ℓ_h is unbordered, ℓ_h is the shortest nonempty border of m_1 .

If there were a word z which is a nonempty border of m_1 and also a prefix of m_2 , by Remark 7.5, ℓ_h would be a prefix of m_2 . Therefore, ℓ_h would be a prefix of the word $\ell_{h+1}^{n_{h+1}} \cdots \ell_r^{n_r}$ which contradicts Lemma 7.4. ◀

► **Proposition 7.7.** *Let $w \in \Sigma^+$ and let $\text{ICFL}(w) = (m_1, \dots, m_k)$ be the unique inverse Lyndon factorization of w having the border property. Then m_k is the longest suffix of w which is an inverse Lyndon word.*

Proof. Let $w \in \Sigma^+$ and let (m_1, \dots, m_k) be the unique inverse Lyndon factorization of w having the border property. If $k = 1$ we are done. Thus suppose $k > 1$. By contradiction, suppose that m_k is not the longest suffix of w that is an inverse Lyndon word. Let s be such longest suffix. Thus, there exist a nonempty suffix x of m_j , $1 \leq j < k$ such that $s = x m_{j+1} \cdots m_k$. Furthermore x must be a proper suffix of m_j or we would have $s = m_j \cdots m_k \ll m_{j+1} \cdots m_k$ contradicting the hypothesis that s is inverse Lyndon.

We claim that $x \ll m_{j+1}$. Indeed, since m_j is an inverse Lyndon word, it holds $x \preceq m_j$. Thus, if $x \ll m_j$ or $x = m_j$, it immediately follows that $x \ll m_{j+1}$. Otherwise, $x \leq_p m_j$ and x is a nonempty border of m_j . By Lemma 5.2 applied to (m_1, \dots, m_k) , with $x = \alpha$, there must exist a prefix β of m_{j+1} such that $x \ll \beta$, hence $x \ll m_{j+1}$.

Since $x \ll m_{j+1}$, we have $s = xm_{j+1} \cdots m_k \ll m_{j+1} \cdots m_k$, contradicting the hypothesis that s is an inverse Lyndon word. \blacktriangleleft

► **Remark 7.8.** Let $w \in \Sigma^+$ and let $\text{ICFL}(w) = (m_1, \dots, m_k)$ with $k > 1$. Let w' be such that $w = w'm_k$. Then, by Propositions 5.3 and 7.7, we obtain $\text{ICFL}(w) = (\text{ICFL}(w'), m_k)$.

Proposition 7.9 allows us to determine the longest suffix m' of a word w such that m' is an inverse Lyndon word.

► **Proposition 7.9.** *Let $w \in \Sigma^+$ be an inverse Lyndon word, and let $\ell \in \Sigma^+$ be an anti-Lyndon word. Then:*

1. *If $\ell \ll w$, then for every $k \geq 1$, $\ell^k w$ is not an inverse Lyndon word.*
2. *If ℓw is not an inverse Lyndon word, then $\ell \ll w$. Furthermore, for every $k \geq 1$, w is the longest suffix of $\ell^k w$ that is an inverse Lyndon word.*

Proof. By Lemma 2.2, the proof of 1. is immediate. Suppose ℓw is not inverse Lyndon. Then, there exists a proper suffix s of ℓw such that $\ell w \preceq s$, hence $\ell w \ll s$. Since ℓ is anti-Lyndon, for every proper suffix x of ℓ it follows $x \ll \ell$ and consequently $xw \ll \ell w$. Thus, s must be a suffix of w . Since w is an inverse Lyndon word, one of the following three cases holds: (1) $w = s$; (2) $s <_p w$; (3) $s \ll w$. By $\ell w \ll s$, in each of the three cases it is evident that $\ell w \ll w$. Thus there are $r, t, t' \in \Sigma^*$ and $a, b \in \Sigma$ with $a < b$ such that $\ell w = rat$, $w = rbt'$. If $|\ell| \geq |ra|$, then clearly $\ell \ll w$. Otherwise, $|\ell| \leq |r|$ and there is $r' \in \Sigma^*$ such that $r = \ell r'$. Consequently, by $\ell w = rat = \ell r'at$, we obtain $w = r'at$. Hence $w = rbt' = \ell r'bt' = r'at$ which contradicts the fact that w is an inverse Lyndon word.

For every $k \geq 1$, w is a suffix of $\ell^k w$ that is an inverse Lyndon word. Let x be a proper nonempty suffix of ℓ . Of course $x \ll \ell$. The word xw is not an inverse Lyndon word, otherwise we would have $\ell \ll w \preceq xw \ll \ell w$, a contradiction. Moreover, by Lemma 2.2, for any j , $1 \leq j < k$, we have $x\ell^j w \ll \ell^j w$ and $x\ell^j w$ is not an inverse Lyndon word. Finally, by 1., $\ell^k w$ is not an inverse Lyndon word. \blacktriangleleft

■ **Algorithm 1** Compute $\text{ICFL}(w)$, the unique compact factorization of w having the border property.

```

1: function FACTORIZE( $w$ )
2:    $(\ell_1^{e_1}, \dots, \ell_n^{e_n}) \leftarrow \text{COMPACTFACTORS}(w)$            ▷ Compute compact factors of  $w$ 
3:    $\mathcal{F} \leftarrow \emptyset$ 
4:    $m' \leftarrow \ell_n^{e_n}$ 
5:   for  $t = n - 1$  downto 1 do                               ▷ Work one compact factor at a time
6:     if  $\ell_t \ll m'$  then                                       ▷ Proposition 7.9
7:        $\mathcal{F} \leftarrow (m', \mathcal{F})$ 
8:        $m' \leftarrow \ell_t^{e_t}$ 
9:     else
10:       $m' \leftarrow \ell_t^{e_t} \cdot m'$ 
11:    $\mathcal{F} \leftarrow (m', \mathcal{F})$ 
12:   return  $\mathcal{F}$ 

```

We now describe Algorithm 1. Function $\text{FACTORIZE}(w)$ will compute the unique compact factorization of w having the border property. First, at line 2, the decomposition of w into its compact factors is computed. Then, the factorization of w is carried out from right to left. Specifically, in accordance with Proposition 7.7, the for-loop at lines 5–10 will search for the longest suffix m' of w that is an inverse Lyndon word. The update of m' is managed

31:12 Unveiling the Connection Between CFL and ICFL via a Border Property

by iteratively applying Proposition 7.9 at line 6. Once such longest suffix is found (that is, when the condition at line 6 is true) it is added to the growing factorization \mathcal{F} and a new search for the longest suffix for the remaining portion of the string is initiated. Otherwise, line 10, the suffix is extended. In the end, the complete factorization is returned.

► **Example 7.10.** Let $\Sigma = \{a, b, c, d\}$, $a < b < c < d$, and let us run $\text{FACTORIZE}(w)$ on $w = \text{dabadabdabdadac}$. First, at line 2, we get the sequence $(\ell_1, \ell_2^2, \ell_3) = (\text{daba}, (\text{dab})^2, \text{dadac})$. Then, at lines 3–4 we set $\mathcal{F} = \emptyset$ and $m' = \ell_3 = \text{dadac}$. We begin the for-loop at lines 5–10 in which i is set to 2 and 1, in turn. With $i = 2$ the test of line 6 succeeds, since $\ell_2 = \text{dab} \ll \text{dadac} = m'$, and so we set $\mathcal{F} = (\text{dadac})$ and $m' = \ell_2^2 = (\text{dab})^2$. At the second iteration, with $i = 1$, the test of line 6 again succeeds, since $\ell_1 = \text{daba} \ll (\text{dab})^2 = m'$, thus we set $\mathcal{F} = ((\text{dab})^2, \text{dadac})$ and $m' = \ell_1 = \text{daba}$. We now fall out of the loop to line 11 where we set $\mathcal{F} = (\text{daba}, (\text{dab})^2, \text{dadac}) = \text{ICFL}(w)$.

7.1 Correctness and complexity

We now prove that Algorithm 1 is correct, that is that it will compute the unique inverse Lyndon factorization of w having the border property, namely $\text{ICFL}(w)$. Formally:

► **Lemma 7.11.** *Let $w \in \Sigma^+$, and let \mathcal{F} be the result of $\text{FACTORIZE}(w)$. Then, $\mathcal{F} = \text{ICFL}(w)$.*

Proof. Let $(\ell_1^{e_1}, \dots, \ell_n^{e_n})$ be the decomposition of w into its compact factors, and let $L_t = \ell_t^{e_t} \dots \ell_n^{e_n}$. We will denote by m'_t (resp. \mathcal{F}_t) the value of m' (resp. \mathcal{F}) at the end of iteration t . We will prove the following loop invariant: at the end of iteration t , sequence (m'_t, \mathcal{F}_t) is a compact factorization of L_t having the border property. The claimed result will follow by Corollary 7.2.

Initialization. Prior to entering the loop, $(m'_n, \mathcal{F}_n) = (\ell_n^{e_n})$, where the last equality follows from Proposition 7.7.

Maintenance. Let $t \leq n - 1$. By the induction hypothesis, $\text{ICFL}(L_{t+1}) = (m'_{t+1}, \mathcal{F}_{t+1})$.

Suppose $\ell_t \ll m'_{t+1}$. Then, by 1. of Proposition 7.9, $\ell_t \cdot m'_{t+1}$ is not inverse Lyndon and m'_{t+1} is the longest suffix of $\ell_t^{e_t} \cdot m'_{t+1}$ that is an inverse Lyndon word. Thus, by Proposition 7.7 m'_{t+1} is the last factor of any compact factorization of $\ell_t^{e_t} \cdot m'_{t+1}$. Hence, $(m'_t, \mathcal{F}_t) = (\ell_t^{e_t}, m'_{t+1}, \mathcal{F}_{t+1})$ is a compact factorization of L_t having the border property.

Now, consider the case where $\ell_t \not\ll m'_{t+1}$. Then, by the contrapositive of 2. of Proposition 7.9, $\ell_t \cdot m'_{t+1}$ is inverse Lyndon and thus, again by 2. of Proposition 7.9, $\ell_t^{e_t} \cdot m'_{t+1}$ is inverse Lyndon. Therefore, $(m'_t, \mathcal{F}_t) = (\ell_t^{e_t} \cdot m'_{t+1}, \mathcal{F}_{t+1})$ is a compact factorization having the border property.

Termination. After iteration $t = 1$, sequence $(m'_1, \mathcal{F}_1) = \text{ICFL}(L_1) = \text{ICFL}(w)$.

Finally, line 11 sets $\mathcal{F} = (m'_1, \mathcal{F}_1) = \text{ICFL}(w)$. ◀

Function $\text{FACTORIZE}(w)$ has time complexity that is linear in the length of w . Indeed, the sequence of compact factors obtained at line 2 can be computed in linear time in the length of w by a simple modification of Duval's algorithm (see [23]). After that, each iteration t of loop 5–10 can be implemented to run in time $\mathcal{O}(|\ell_t|)$. Indeed, condition $\ell_t \ll m'$ can be checked by naively comparing ℓ_t against m' . Furthermore, the update of m' and \mathcal{F} can be done in constant time: in fact, ℓ_t , $\ell_t^{e_t}$, m' and \mathcal{F} can all be implemented as pairs of indexes (in case of the former three) or as a list of indexes (in case of the latter) of w .

8 Conclusions

We discover the special connection between the Lyndon factorization under the inverse lexicographic ordering, named CFL_{in} and the canonical inverse Lyndon factorization, named ICFL: there exists a unique inverse Lyndon factorization having the border property and this unique factorization is ICFL. Moreover each inverse factor of ICFL is obtained by concatenating compact factors of CFL_{in} . These properties give a constrained structure to ICFL that deserve to be further explored to characterize properties of words. In particular, we believe the characterization of ICFL as a compact factorization, proved in the paper, could highlight novel properties related the compression of a word, as investigated in [20]. In particular, the number of compact factors seems to be a measure of repetitiveness of the word to be also used in speeding up suffix sorting of a word.

Finally, we believe that the characterization of ICFL in terms of CFL_{in} may be used to extend to ICFL the *conservation property* proved in [10] for CFL. This property shows that the Lyndon factorization of a word w preserves common factors with the factorization of a superstring of w . This extends the conservation of Lyndon factors explored for the product $u \cdot v$ of two words u and v [1, 20].

References

- 1 Alberto Apostolico and Maxime Crochemore. Fast parallel Lyndon factorization with applications. *Mathematical systems theory*, 28(2):89–108, 1995.
- 2 Hideo Bannai, Juha Kärkkäinen, Dominik Köppl, and Marcin Piatkowski. Constructing and indexing the bijective and extended Burrows-Wheeler transform. *Information and Computation*, 297:105153, 2024. doi:10.1016/j.ic.2024.105153.
- 3 Hideo Bannai, I Tomohiro, Shunsuke Inenaga, Yuto Nakashima, Masayuki Takeda, and Kazuya Tsuruta. A new characterization of maximal repetitions by Lyndon trees. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 562–571, 2015.
- 4 Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Encyclopedia of Mathematics and its Applications 129, Cambridge University Press, 2009.
- 5 Nico Bertram, Jonas Ellert, and Johannes Fischer. Lyndon words accelerate suffix sorting. In Petra Mutzel, Rasmus Pagh, and Grzegorz Herman, editors, *29th Annual European Symposium on Algorithms, ESA 2021, September 6-8, 2021, Lisbon, Portugal (Virtual Conference)*, volume 204 of *LIPICs*, pages 15:1–15:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
- 6 Elena Biagi, Davide Cenzato, Zsuzsanna Lipták, and Giuseppe Romana. On the number of equal-letter runs of the Bijective Burrows-Wheeler Transform. In *CEUR Workshop Proceedings*, volume 3587, pages 129–142. R. Piskac c/o Redaktion Sun SITE, Informatik V, RWTH Aachen, 2023.
- 7 Paola Bonizzoni, Matteo Costantini, Clelia De Felice, Alessia Petescia, Yuri Pirola, Marco Previtali, Raffaella Rizzi, Jens Stoye, Rocco Zaccagnino, and Rosalba Zizza. Numeric Lyndon-based feature embedding of sequencing reads for machine learning approaches. *Inf. Sci.*, 607:458–476, 2022.
- 8 Paola Bonizzoni, Clelia De Felice, Rocco Zaccagnino, and Rosalba Zizza. Inverse Lyndon words and inverse Lyndon factorizations of words. *Adv. Appl. Math.*, 101:281–319, 2018.
- 9 Paola Bonizzoni, Clelia De Felice, Rocco Zaccagnino, and Rosalba Zizza. Lyndon words versus inverse Lyndon words: Queries on suffixes and bordered words. In Alberto Leporati, Carlos Martín-Vide, Dana Shapira, and Claudio Zandron, editors, *Language and Automata Theory and Applications - 14th International Conference, LATA 2020, Milan, Italy, March 4-6, 2020, Proceedings*, volume 12038 of *Lecture Notes in Computer Science*, pages 385–396. Springer, 2020. doi:10.1007/978-3-030-40608-0_27.

- 10 Paola Bonizzoni, Clelia De Felice, Rocco Zaccagnino, and Rosalba Zizza. On the longest common prefix of suffixes in an inverse Lyndon factorization and other properties. *Theor. Comput. Sci.*, 862:24–41, 2021.
- 11 Paola Bonizzoni, Clelia De Felice, Rocco Zaccagnino, and Rosalba Zizza. From the Lyndon factorization to the Canonical Inverse Lyndon factorization: back and forth. under submission, ArXiv, 2024.
- 12 Kuo-Tsai Chen, Ralph H. Fox, and Roger C. Lyndon. Free Differential calculus, IV. The Quotient Groups of the Lower Central Series. *Ann. Math.*, 68:81–95, 1958.
- 13 Christian Choffrut and Juhani Karhumäki. Combinatorics of Words. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages, Vol. 1*, pages 329–438. Springer-Verlag, Berlin, Heidelberg, 1997.
- 14 Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on strings*. Cambridge University Press, 2007.
- 15 Olivier Delgrange and Eric Rivals. Star: an algorithm to search for tandem approximate repeats. *Bioinformatics*, 20(16):2812–2820, 2004.
- 16 Jean-Pierre Duval. Factorizing Words over an Ordered Alphabet. *J. Algorithms*, 4(4):363–381, 1983.
- 17 Harold Fredricksen and James Maiorana. Necklaces of beads in k colors and k -ary de Bruijn sequences. *Discrete Math.*, 23(3):207–210, 1978.
- 18 Daniele A. Gewurz and Francesca Merola. Numeration and enumeration. *Eur. J. Comb.*, 33(7):1547–1556, 2012.
- 19 Sukhpal Singh Ghuman, Emanuele Giaquinta, and Jorma Tarhio. Alternative Algorithms for Lyndon Factorization. In *Proceedings of the Prague Stringology Conference 2014, Prague, Czech Republic, September 1-3, 2014*, pages 169–178, 2014.
- 20 Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Faster Lyndon factorization algorithms for SLP and LZ78 compressed text. *Theoretical Computer Science*, 656:215–224, 2016.
- 21 Dominik Köppl, Daiki Hashimoto, Diptarama Hendrian, and Ayumi Shinohara. In-place bijective Burrows-Wheeler Transforms. In Inge Li Gørtz and Oren Weimann, editors, *31st Annual Symposium on Combinatorial Pattern Matching, CPM 2020, June 17-19, 2020, Copenhagen, Denmark*, volume 161 of *LIPICs*, pages 21:1–21:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICs.CPM.2020.21.
- 22 M. Lothaire. *Algebraic Combinatorics on Words, Encyclopedia Math. Appl*, volume 90. Cambridge University Press, 1997.
- 23 M. Lothaire. *Applied Combinatorics on Words*. Cambridge University Press, 2005.
- 24 Roger Lyndon. On Burnside’s problem. *Trans. Amer. Math. Soc.*, 77:202–215, 1954.
- 25 Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, and Marinella Sciortino. Suffix array and Lyndon factorization of a text. *J. Discrete Algorithms*, 28:2–8, 2014.
- 26 Igor Martayan, Bastien Cazaux, Antoine Limasset, and Camille Marchet. Conway-Bromage-Lyndon (CBL): an exact, dynamic representation of k -mer sets. *bioRxiv*, 2024. doi:10.1101/2024.01.29.577700.
- 27 Christophe Reutenauer. Free Lie algebras. In *Handbook of Algebra, London Mathematical Society Monographs*. Oxford Science Publications, 1993.