# The Standard Factorization of Lyndon Words: an Average Point of View

Frédérique Bassino [a] Julien Clément [a] Cyril Nicaud [a]

[a] *Institut Gaspard Monge*
*Université de Marne-la-Vallée*
*77454 Marne-la-Vallée Cedex 2 - France*

**Abstract**

A non-empty word $w$ of $\{a, b\}^*$ is a Lyndon word if and only if it is strictly smaller for the lexicographical order than any of its proper suffix. Such a word $w$ is either a letter or admits a standard factorization $uv$ where $v$ is its smallest proper suffix. For any Lyndon word $v$, we show that the set of Lyndon words having $v$ as right factor of the standard factorization is rational and compute explicitly the associated generating function. Next we establish that, for the uniform distribution over the Lyndon words of length $n$, the average length of the right factor $v$ of the standard factorization is asymptotically $3n/4$.

*Key words:* Lyndon word, standard factorization, average-case analysis, analytic combinatorics, success run

## 1 Introduction

Given a totally ordered alphabet $A$, a *Lyndon word* is a word that is strictly smaller, for the lexicographical order, than any of its conjugates (*i.e.*, all words obtained by a circular permutation on the letters). Lyndon words were introduced by Lyndon [19] under the name of "standard lexicographic sequences" in order to give a base for the free Lie algebra over $A$; the standard factorization plays a central role in this framework (see [17], [22], [23]). More precisely to a Lyndon word $w$ is associated a binary tree $T(w)$ recursively built in the following way: if $w$ is a letter, then $T(w)$ is a leaf labeled by $w$, otherwise $T(w)$

*Email addresses:* bassino@univ-mlv.fr (Frédérique Bassino), clementj@univ-mlv.fr (Julien Clément), nicaud@univ-mlv.fr (Cyril Nicaud).

is an internal node having $T(u)$ and $T(v)$ as children where $u \cdot v$ is the standard factorization of $w$. This structure encodes a nonassociative operation, either a commutator in the free group [6], or a Lie bracketing [17]; both constructions leads to bases of the free Lie algebra. The average complexity of the algorithms computing these bases is basically determined by the average height of these trees.

One of the basic properties of the set of Lyndon words is that every word is uniquely factorizable as a non increasing product of Lyndon words. As there exists a bijection between Lyndon words over an alphabet of cardinality $k$ and irreducible polynomials over $\mathbb{F}_k$ [14], lots of results are known about this factorization: the average number of factors, the average length of the longest factor [10] and of the shortest [21].

Several algorithms deal with Lyndon words. Duval gives in [8] an algorithm that computes, in linear time, the factorization of a word into Lyndon words. There exists [13] an algorithm generating all Lyndon word up to a given length in lexicographical order. This algorithm runs in a constant average time (see [3]).

In Section 2, we define more formally Lyndon words and give some enumerative properties of these sets of words. Then we introduce the standard factorization of a Lyndon word $w$ which is the unique couple of Lyndon words $u$, $v$ such that $w = uv$ and $v$ is of maximal length.

In Section 3, we study the set of Lyndon words of $\{a, b\}^*$ having a given right factor in their standard factorization and prove that it is a rational language. We also compute its associated generating function. But as the set of Lyndon words is not context-free [1], we are not able to directly derive asymptotic properties from these generating functions. Consequently in Section 4 we use probabilistic techniques and results from analytic combinatorics (see [11]) in order to compute the average length of the factors of the standard factorization of Lyndon words.

Section 5 is devoted to algorithms and experimental results. We give an algorithm to generate randomly for uniform distribution a Lyndon word of a given length and another one related to the standard factorization of a Lyndon word. Finally experiments are given which confirm our results and give hints of further studies.

## 2 Preliminary

We denote $A^*$ the free monoid over the alphabet $A = \{a, b\}$ obtained by all finite concatenations of elements of $A$. The length $|w|$ of a word $w$ is the number of the letters $w$ is product of, $|w|_a$ is the number of occurrences of the letter $a$ in $w$. We consider the lexicographical order $<$ over all non-empty words of $A^*$ defined by the extension of the order $a < b$ over $A$.

We record two properties of this order

(i) For any word $w$ of $A^*$, $u < v$ if and only if $wu < wv$.
(ii) Let $x, y \in A^*$ be two words such that $x < y$. If $x$ is not a prefix of $y$ then for every $x', y' \in A^*$ we have $xx' < yy'$.

By definition, a *Lyndon word* is a primitive word (*i.e*, it is not a power of another word) that is minimal, for the lexicographical order, in its conjugate class (*i.e*, the set of all words obtained by a circular permutation). The set of Lyndon words of length $n$ is denoted by $\mathcal{L}_n$ and $\mathcal{L} = \cup_n \mathcal{L}_n$.

$$\mathcal{L} = \{a, b, ab, aab, abb, aaab, aabb, abbb, \\ aaaab, aaabb, aabab, aabbb, ababb, abbbb, \dots\}$$

Equivalently, $w \in \mathcal{L}$ if and only if

$$\forall u, v \in A^+, \quad w = uv \Rightarrow w < vu.$$

A non-empty word is a Lyndon word if and only if it is strictly smaller than any of its proper suffixes.

**Proposition 1** *A word $w \in A^+$ is a Lyndon word if and only if either $w \in A$ or $w = uv$ with $u, v \in \mathcal{L}$, $u < v$.*

**Theorem 2 (Lyndon)** *Any word $w \in A^+$ can be written uniquely as a non-increasing product of Lyndon words:*

$$w = \ell_1 \ell_2 \dots \ell_n, \quad \ell_i \in \mathcal{L}, \quad \ell_1 \geq \ell_2 \geq \cdots \geq \ell_n.$$

*Moreover, $\ell_n$ is the smallest suffix of $w$.*

The number $\mathrm{Card}(\mathcal{L}_n)$ of Lyndon words of length $n$ over $A$ (see [17]) is

$$\mathrm{Card}(\mathcal{L}_n) = \frac{1}{n} \sum_{d|n} \mu(d)\, \mathrm{Card}(A)^{n/d},$$

where $\mu$ is the Moebius function defined on $\mathbb{N} \setminus \{0\}$ by $\mu(1) = 1$, $\mu(n) = (-1)^i$ if $n$ is the product of $i$ distinct primes and $\mu(n) = 0$ otherwise.

When $\mathrm{Card}(A) = 2$, we obtain the following estimate

$$\mathrm{Card}(\mathcal{L}_n) = \frac{2^n}{n}\left(1 + O\left(2^{-n/2}\right)\right).$$

For $w \in \mathcal{L} \setminus A$ a Lyndon word not reduced to a letter, the pair $(u, v)$, $u, v \in \mathcal{L}$ such that $w = uv$ and $v$ of maximal length is called the *standard factorization*. The words $u$ and $v$ are called the *left factor* and *right factor* of the *standard factorization*.

Equivalently, the right factor $v$ of the standard factorization of a Lyndon word $w$ which is not reduced to a letter can be defined as the smallest proper suffix of $w$.

**Example 3** $aaabaab = aaab \cdot aab, aaababb = a \cdot aababb, aabaabb = aab \cdot aabb$.

## 3 Counting Lyndon words with a given right factor

In this section, we prove that the set of Lyndon words with a given right factor in their standard factorization is a rational language and compute its generating function. The techniques used in the following basically come from combinatorics on words.

Let $w = vab^i$ be a word containing one $a$ and ending with a sequence of $b$. The word $R(w) = vb$ is the *reduced word* of $w$.

For any Lyndon word $v$, we define the set

$$\mathcal{X}_v = \{v_0 = v, v_1 = R(v), v_2 = R^2(v), \ldots, v_k = R^k(v)\}.$$

where $k = |v|_a$ is the number of occurrences of $a$ in $v$. Note that $\mathrm{Card}\,(\mathcal{X}_v) = |v|_a + 1$ and $v_k = b$.

**Example 4** *(1)* $v = aabab$: $\mathcal{X}_{aabab} = \{aabab, aabb, ab, b\}$.
*(2)* $v = a$: $\mathcal{X}_a = \{a, b\}$.
*(3)* $v = b$: $\mathcal{X}_b = \{b\}$.

By construction, $v$ is the smallest element of $\mathcal{X}_v^+$ for the lexicographical order.

**Lemma 5** *Every word $x \in \mathcal{X}_v$ is a Lyndon word.*

**PROOF.** If $v = a$, then $\mathcal{X}_v = \{a, b\}$, else any element of $\mathcal{X}_v$ ends by a $b$. In this case, if $x \notin \mathcal{L}$, there exists a decomposition $x = x_1x_2b$ such that $x_2bx_1 \leq x_1x_2b$

and $x_1 \neq \varepsilon$. Thus $x_2 a$ is not a left factor of $x_1 x_2 b$ and $x_2 a < x_1 x_2 a$. By construction of $\mathcal{X}_v$, as $x \neq v$, there exists a word $w$ such that $v = x_1 x_2 a w$. We get that $x_2 a w x_1 < x_1 x_2 a w$. This is impossible since $v \in \mathcal{L}$. $\square$

A *code* $C$ over $A^*$ is a set of non-empty words such any word $w$ of $A^*$ can be written in at most one way as a product of elements of $C$. A set of words is *prefix* if none of its elements is the prefix of another one. Such a set is a code, called a *prefix code*. A code $C$ is said to be *circular* if any word of $A^*$ written along a circle admits at most one decomposition as product of words of $C$. These codes can be characterized as the bases of very pure monoids, *i.e.*, if $w^n \in C^*$ then $w \in C^*$. For a general reference about codes, see [2].

**Proposition 6** *The set $\mathcal{X}_v$ is a prefix circular code.*

**PROOF.** If $x, y \in \mathcal{X}_v$ with $|x| < |y|$, then, by construction of $\mathcal{X}_v$, $x > y$. So $x$ is not a left factor of $y$ and $\mathcal{X}_v$ is a prefix code.

Moreover, for every $n \geq 1$, if $w$ is a word such that $w^n \in \mathcal{X}_v^*$ then $w \in \mathcal{X}_v^*$. Indeed if $w \notin \mathcal{X}_v^*$, then either $w$ is a proper prefix of a word of $\mathcal{X}_v$ or $w$ has a prefix in $\mathcal{X}_v^*$. If $w$ is a proper prefix of a word of $\mathcal{X}_v$, it is a prefix of $v$ and it is strictly smaller than any word of $\mathcal{X}_v$. As $w^n \in \mathcal{X}_v^*$, $w$ or one of its prefix is a suffix of a word of $\mathcal{X}_v$. But all elements of $\mathcal{X}_v$ are Lyndon words greater than $v$, so their suffixes are strictly greater than $v$ and $w$ can not be a prefix of a word of $\mathcal{X}_v$.

Now if $w = w_1 w_2$ where $w_1$ is the longest prefix of $w$ in $\mathcal{X}_v^+$, then $w_2$ is a non-empty prefix of a word $\mathcal{X}_v$, so $w_2$ is strictly smaller than any word of $\mathcal{X}_v$. As $w^n \in \mathcal{X}_v^*$, $w_2$ or one of its prefix is a suffix of a word of $\mathcal{X}_v$, but all elements of $\mathcal{X}_v$ are Lyndon words greater than $v$, so their suffixes are strictly greater than $v$ and $w$ can not have a prefix in $\mathcal{X}_v^+$.

As a conclusion, since $\mathcal{X}_v$ is a code and for every $n \geq 1$, if $w^n \in \mathcal{X}_v^*$ then $w^n \in \mathcal{X}_v^*$, $\mathcal{X}_v$ is circular code. $\square$

**Proposition 7** *Let $\ell \in \mathcal{L}$ be a Lyndon word, $\ell \geq v$ if and only if $\ell \in \mathcal{X}_v^+$.*

**PROOF.** If $\ell \geq v$, let $\ell_1$ be the longest prefix of $\ell$ which belongs to $\mathcal{X}_v^*$, and $\ell_2$ such that $\ell = \ell_1 \ell_2$. If $\ell_2 \neq \varepsilon$, we have the inequality $\ell_2 \ell_1 > \ell \geq v$, thus $\ell_2 \ell_1 > v$. The word $v$ is not a prefix of $\ell_2$ since $\ell_2$ has no prefix in $\mathcal{X}_v$, hence we have $\ell_2 = \ell_2' b \ell_2''$ and $v = \ell_2' a v''$. Then, by construction of $\mathcal{X}_v$, $\ell_2' b \in \mathcal{X}_v$ which is impossible. Thus $\ell_2 = \varepsilon$ and $\ell \in \mathcal{X}_v^+$.

Conversely, if $\ell \in \mathcal{X}_v^+$, as a product of words greater than $v$, $\ell \geq v$. $\square$

Denote for a language $L \subset A^*$

$$a^{-1}L = \{w \in A^* \,|\, aw \in L\}.$$

**Theorem 8** *Let $v \in \mathcal{L}$ and $w \in A^*$. Then $awv$ is a Lyndon word with $aw \cdot v$ as standard factorization if and only if $w \in \mathcal{X}_v^* \setminus (a^{-1}\mathcal{X}_v)\mathcal{X}_v^*$. Hence the set $\mathcal{F}_v$ of Lyndon words having $v$ as right standard factor is a rational language.*

**PROOF.** Assume that $awv$ is a Lyndon word and its standard factorization is $aw \cdot v$. By Theorem 2, $wv$ can be written uniquely as

$$wv = \ell_1\ell_2 \dots \ell_n, \quad \ell_i \in \mathcal{L}, \quad \ell_1 \geq \ell_2 \geq \cdots \geq \ell_n.$$

As $v$ is the smallest (for the lexicographical order) suffix of $awv$, and consequently of $wv$, we get $\ell_n = v$; if $w = \varepsilon$, then $n = 1$, else $n \geq 2$ and for $1 \leq i \leq n-1$, $\ell_i \geq v$. Thus, $w \in \mathcal{X}_v^*$.

Moreover if $w \in (a^{-1}\mathcal{X}_v)\mathcal{X}_v^*$, then $aw \in \mathcal{X}_v^+ \cap \mathcal{L}$. Hence $aw \geq v$ which is contradictory with the definition of the standard factorization. So $w \in \mathcal{X}_v^* \setminus (a^{-1}\mathcal{X}_v)\mathcal{X}_v^*$.

Conversely, if $w \in \mathcal{X}_v^* \setminus (a^{-1}\mathcal{X}_v)\mathcal{X}_v^*$, then

$$w = x_1x_2 \dots x_n, \quad x_i \in \mathcal{X}_v \quad \text{and} \quad aw \notin \mathcal{X}_v^+.$$

From Proposition 1, the product $\ell\ell'$ of two Lyndon words such that $\ell < \ell'$ is a Lyndon word. Replacing as much as possible $x_ix_{i+1}$ by their product when $x_i < x_{i+1}$, $w$ can be rewritten as

$$w = y_1y_2 \dots y_m, \quad y_i \in \mathcal{X}_v^+ \cap \mathcal{L}, \quad y_1 \geq y_2 \geq \cdots \geq y_m.$$

As $aw \notin \mathcal{X}_v^+$, for any integer $1 \leq i \leq m$, $ay_1 \dots y_i \notin \mathcal{X}_v^+$.

Now we prove by induction that $aw \in \mathcal{L}$. As $y_1 \in \mathcal{L}$ and $a < y_1$, $ay_1 \in \mathcal{L}$. Suppose that $ay_1 \dots y_i \in \mathcal{L}$. Then, as $y_{i+1} \in \mathcal{L} \cap \mathcal{X}_v^+$, and $ay_1 \dots y_i \in \mathcal{L} \setminus \mathcal{X}_v^+$, from Proposition 7, we get $ay_1 \dots y_i < v \leq y_{i+1}$. Hence $ay_1 \dots y_{i+1} \in \mathcal{L}$. So, $aw \in \mathcal{L}$.

As $aw \in \mathcal{L} \setminus \mathcal{X}_v^+$, $aw < v$ and $awv \in \mathcal{L}$. Setting $v = y_{m+1}$, we have

$$wv = y_1y_2 \dots y_my_{m+1}, \quad y_i \in \mathcal{X}_v^* \cap \mathcal{L}, \quad y_1 \geq y_2 \geq \dots \geq y_{m+1}.$$

Moreover any proper suffix $s$ of $awv$ is a suffix of $wv$ and can be written as $s = y_i'y_{i+1} \dots y_{m+1}$ where $y_i'$ is a suffix of $y_i$. As $y_i \in \mathcal{L}$, $y_i' \geq y_i$. As $y_i \in \mathcal{X}_v^+$, $y_i \geq v$ and thus $s \geq v$. Thus, $v$ is the smallest suffix of $awv$ and $aw \cdot v$ is the standard factorization of the Lyndon word $awv$.

Finally as the set of rational languages is closed by complementation, concatenation, Kleene star operation and left quotient, for any Lyndon word $v$, the set $\mathcal{F}_v$ of Lyndon words having $v$ right standard factor is a rational language. $\square$

We define the generating functions $X_v(z)$ of $\mathcal{X}_v$ and $X_v^*(z)$ of $\mathcal{X}_v^*$:

$$X_v(z) = \sum_{w \in \mathcal{X}_v} z^{|w|} \quad \text{and} \quad X_v^*(z) = \sum_{w \in \mathcal{X}_v^*} z^{|w|}.$$

As the set $\mathcal{X}_v$ is a code, the elements of $\mathcal{X}_v^*$ are sequences of elements of $\mathcal{X}_v$ (see [11]):

$$X_v^*(z) = \frac{1}{1 - X_v(z)}.$$

Denote by $F_v(z) = \sum_{x \in \mathcal{F}_v} z^{|x|}$ the generating function of the set

$$\mathcal{F}_v = \{awv \in \mathcal{L} \,|\, aw \cdot v \text{ is the standard factorization}\}.$$

**Theorem 9** *Let $v$ be a Lyndon word. The generating function of the set $\mathcal{F}_v$ of Lyndon words having a right standard factor $v$ can be written*

$$F_v(z) = z^{|v|} \left(1 + \frac{2z - 1}{1 - X_v(z)}\right).$$

**PROOF.** First of all, note that any Lyndon word of $\{a, b\}^*$ which is not a letter ends with the letter $b$, so $F_a(z) = 0$. And as $\mathcal{X}_a = \{a, b\}$, the formula given for $F_v(z)$ holds for $v = a$.

Assume that $v \neq a$. From Theorem 8, $F_v(z)$ can be written as

$$F_v(z) = z^{|av|} \sum_{w \in \mathcal{X}_v^* \setminus a^{-1}\mathcal{X}_v^+} z^{|w|}.$$

In order to transform this combinatorial description involving $\mathcal{X}_v^* \setminus a^{-1}\mathcal{X}_v^+$ into an enumerative formula of the generating function $F_v(z)$, we prove first that $a^{-1}\mathcal{X}_v^+ \subset \mathcal{X}_v^*$ and, next that the set $a^{-1}\mathcal{X}_v^+$ can be described as a disjoint union of rational sets.

If $x \in \mathcal{X}_v \setminus \{b\}$, then $x$ is greater than $v$ and as $x$ is a Lyndon word, its proper suffixes are strictly greater than $v$; consequently, writing $a^{-1}x$ as a non-increasing sequence of Lyndon word $\ell_1, \ldots, \ell_m$, we get, since $\ell_m \geq v$, that for all $i$, $\ell_i$ is greater than $v$. Consequently from Proposition 7, for all $i$, $\ell_i \in \mathcal{X}_v^*$ and as a product of elements of $\mathcal{X}_v^+$, $a^{-1}x \in \mathcal{X}_v^+$. Therefore $a^{-1}\left(\mathcal{X}_v \setminus \{b\}\right)\mathcal{X}_v^* \subset \mathcal{X}_v^*$.

Moreover if $x_1, x_2 \in \mathcal{X}_v$ and $x_1 \neq x_2$, as $\mathcal{X}_v$ is a prefix code,

$$a^{-1}x_1\mathcal{X}_v^* \cap a^{-1}x_2\mathcal{X}_v^* = \emptyset.$$

Thus $a^{-1}\left(\mathcal{X}_v \setminus \{b\}\right)\mathcal{X}_v^*$ is the disjoint union of the sets $\left(a^{-1}x_i\right)\mathcal{X}_v^*$ when $x_i$ ranges over $\mathcal{X}_v \setminus \{b\}$. Consequently the generating function of the set $\mathcal{F}_v$ of Lyndon words having $v$ as right factor satisfies

$$F_v(z) = z^{|v|+1}\frac{1 - \frac{\mathcal{X}_v(z)-z}{z}}{1-\mathcal{X}_v(z)}$$

and finally the announced equality. □

Note that the function $F_v(z)$ is rational for any Lyndon word $v$. But the right standard factor runs over the set of Lyndon words which is not context-free [1]. Therefore in order to study the average length of the factors in the standard factorization of Lyndon words, we adopt another point of view.

## 4 Main result

Making use of probabilistic techniques and results from analytic combinatorics (see [11]), we establish the following result.

**Theorem 10** *Under the uniform distribution over the Lyndon words of length $n$, the average length of the right factor of the standard factorization is*

$$\frac{3n}{4}\left(1 + O\left(\frac{\log^3 n}{n}\right)\right).$$

**Remark 11** *The error term comes from successive approximations at different steps of the proof and, for this reason, it is probably overestimated (see experimental results in Section 5).*

First we partition the set $\mathcal{L}_n$ in the two following subsets: $a\mathcal{L}_{n-1}$ and $\mathcal{L}_n \backslash a\mathcal{L}_{n-1}$.

Note that $a\mathcal{L}_{n-1} \subset \mathcal{L}_n$, that is, if $w$ is a Lyndon word then $aw$ is also a Lyndon word. Moreover if $w \in a\mathcal{L}_{n-1}$, the standard factorization is $w = a \cdot v$ with $v \in \mathcal{L}_{n-1}$. As

$$\mathrm{Card}\left(\mathcal{L}_{n-1}\right) = \frac{2^{n-1}}{n-1}\left(1 + O\left(2^{-n/2}\right)\right),$$

the contribution of the set $a\mathcal{L}_{n-1}$ to the mean value of the length of the right factor is

$$(n-1) \times \frac{\mathrm{Card}\,(a\mathcal{L}_{n-1})}{\mathrm{Card}\,(\mathcal{L}_n)} = \frac{n}{2}\left(1 + O\left(2^{-n/2}\right)\right).$$

The remaining part of this paper is devoted to the standard factorization of the words of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ which requires a careful analysis.

**Proposition 12** *The contribution of the set $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ to the mean value of the length of right factor is*

$$\frac{n}{4}\left(1 + O\left(\frac{\log^3 n}{n}\right)\right).$$

This proposition basically asserts that in average for the uniform distribution over $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$, the length of the right factor is asymptotically $n/2$.

The idea is to build a transformation $\varphi$, which is an involution on almost all the set $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$, such that the sum of the lengths of standard right factors of $w$ and $\varphi(w)$ is about the length $|w|$ of $w$. The word $\varphi(w)$ is obtained from $w$ by exchanging particular suffixes of the factors of the standard factorization of $w$ so that standard factors of $w$ and $\varphi(w)$ have the same prefixes.

### *4.1 Max-run decomposition of words of $\mathcal{L} \setminus a\mathcal{L}$*

For any Lyndon word $w$ which is not reduced to a letter, there exists a positive integer $k = k(w)$ such that $a^k b$ is a prefix of $w$. It is also the length of the longest runs of $a$'s in $w$. Let $\mathfrak{L}_k$ be the set of Lyndon words with a first run of length $k$.

Denote by $\mathcal{X}_k$ the set $\mathcal{X}_{a^{k-1}b}$ namely $\mathcal{X}_k = \{a^i b \mid 0 \le i \le k-1\}$. We partition each set $\mathfrak{L}_k \setminus a\mathfrak{L}_{k-1}$ in two sets $\mathfrak{L}_k^1$ and $\mathfrak{L}_k^2$

$$\mathfrak{L}_k^1 = a^k b\mathcal{X}_{k-1}^*(a^{k-1}b\mathcal{X}_{k-1}^*)^+ \cap (\mathcal{L} \setminus a\mathcal{L}), \qquad \mathfrak{L}_k^2 = a^k b\mathcal{X}_k^*(a^k b\mathcal{X}_k^*)^+ \cap (\mathcal{L} \setminus a\mathcal{L})$$

depending on the standard factorization. Indeed the standard factorization of a word $w$ of $\mathfrak{L}_k \setminus a\mathfrak{L}_{k-1}$ can only be one of the following

$$\begin{aligned} w &= a^k bu \cdot a^{k-1}bv \\ w &= a^k bu \cdot a^k bv. \end{aligned}$$

This means that the right factor of a Lyndon word $w$ can only begin with $a^{k-1}b$ or $a^k b$. Denote by $K$ the length of the first run of $a$'s of the right factor of $w$. Remark that $K = k-1$ if $w \in \mathfrak{L}_k^1$ and $K = k$ if $w \in \mathfrak{L}_k^2$.

With these notations, we introduce a decomposition of words of $\mathcal{L} \setminus a\mathcal{L}$ called *max-run decomposition* throughout this paper. Any word $w$ of $\mathcal{L} \setminus a\mathcal{L}$ can be written

$$w = f_1 \dots f_m \quad \text{with } k = k(w),\ f_1 \in a^k b \mathcal{X}_K^* \text{ and } f_i \in a^K b \mathcal{X}_K^* \text{ for } 2 \le i \le m.$$

The standard factorization always occurs at a point of the max-run decomposition: there exists $j \in \{2, \dots, m\}$ such that the standard factorization of $w$ is

$$\prod_{i=1}^{j-1} f_i \ \cdot \ \prod_{i=j}^{m} f_i.$$

We will study this decomposition by means of analytical tools and present now definitions and results which play a central role hereafter. Let $X_k(z)$ and $X_k^*(z)$ be the generating functions respectively associated to $\mathcal{X}_k$ and $\mathcal{X}_k^*(z)$ namely

$$X_k(z) = \sum_{i=1}^{k} z^i \qquad \text{and} \qquad X_k^*(z) = \frac{1}{1 - X_k(z)}.$$

The smallest pole of $X_k^*(z)$ that is, from the Rouché theorem (see [5]), the only one in the unit disc is

$$\rho_k = \frac{1}{2} + \epsilon_k, \quad \text{with} \quad \epsilon_k = \frac{1}{2^{k+2}} + \frac{k+1}{2^{2k+3}} + O\left(\frac{k^2}{2^{3k}}\right).$$

The value of $\epsilon_k$ is obtained by the bootstrapping method as in [16] using the fact that $\rho_k$ is a root of $1 - 2z + z^{k+1}$.

Denoting by $[z^n]F(z)$ the coefficient of $z^n$ in $F(z)$ and using the *standard extraction formula* for rational series with a simple pole (see [11]), we can write

$$[z^n]\frac{P(z)}{1 - X_k(z)} = \frac{P(\rho_k)}{X_k'(\rho_k)}\,\rho_k^{-(n+1)} + O(1) \tag{1}$$

provided that $\rho_k$ is not a root of the polynomial $P(z)$. Therefore we also need the following estimate of the derivative $X_k'$ of $X_k$ at $z = \rho_k$

$$(X_k'(\rho_k))^{-1} = \frac{1}{4}\left(\frac{1 - 4\epsilon_k^2}{1 - 2k\epsilon_k}\right) = \frac{1}{4}\left(1 + \frac{k}{2^{k+1}}\right) + O\left(\frac{k^2}{2^{2k}}\right). \tag{2}$$

In the following subsets of Lyndon words will be enumerated by means of the elegant construction of primitive cycles [12].

**Proposition 13 (Primitive cycles)** *Let $\mathcal{C}$ be a code, with generating function $C(z) = \sum_{w \in \mathcal{C}} z^{|w|}$. Then the generating function of the primitive cycles of elements of $\mathcal{C}$ is*

$$\sum_{m \ge 1} \frac{\mu(m)}{m} \log\left(\frac{1}{1 - C(z^m)}\right).$$

This equation can be used directly to obtain several interesting generating functions of sets of words

(i) the set of Lyndon words taking $\mathcal{C} = \{a, b\}$, $C(z) = 2z$.
(ii) the set of Lyndon words beginning with strictly less than $k$ $a$'s taking $\mathcal{C} = \mathcal{X}_k$, $C(z) = X_k(z)$.
(iii) the set of Lyndon words beginning with exactly $k$ $a$'s taking $\mathcal{C} = a^k b(\mathcal{X}_k)^*$, $C(z) = \frac{z^{k+1}}{1-X_k(z)}$.

*Length $k$ of longest runs.*

First we study the precise distribution of the length of the longest runs of $a$'s in a Lyndon word $w$. This question is strongly related to the notion of success run in probability theory [9]

**Proposition 14** *The probability $p_{n,k}$ that $a^i b$, with $1 \leq i < k$, is a prefix of a Lyndon word of length $n$ is*

$$p_{n,k} = (1 + 2\epsilon_k)^{-n} + O\left(2^{-n/2}\right) \text{ with } \epsilon_k = \frac{1}{2^{k+2}} + \frac{k+1}{2^{2k+3}} + O\left(\frac{k^2}{2^{3k}}\right).$$

**PROOF.** Denote $\mathfrak{L}_{<k}$ the set of Lyndon words beginning with strictly less than $k$ $a$'s

$$\mathfrak{L}_{<k} = \{w \in \mathcal{L} \mid w \geq a^{k-1}b\}.$$

The number of words of length $n$ in $\mathfrak{L}_{<k}$ is the number of primitive cycles of elements in $\mathcal{X}_k$ of total length $n$. From Proposition 13, we get

$$L_{<k}(z) = \sum_{m \geq 1} \frac{\mu(m)}{m} \log\left(\frac{1}{1 - X_k(z^m)}\right),$$

where $\mu$ is the Moebius function. We set $L_{<k}(z) = \sum_{n \geq 1} \ell_{n,k} z^n$. Then, differentiating with respect to $z$, we obtain

$$\sum_{n \geq 1} n\ell_{n,k} z^{n-1} = \sum_{m \geq 1} \mu(m) \frac{X_k'(z^m)}{1 - X_k(z^m)} z^{m-1}.$$

Hence we have

$$n\ell_{n,k} = \sum_{m|n} \mu\left(\frac{n}{m}\right) [z^m] \frac{X_k'(z)}{1 - X_k(z)} z.$$

Introducing $\rho_k$ and using Equation (1), we get

$$\ell_{n,k} = \frac{1}{n} \sum_{m|n} \mu\left(\frac{n}{m}\right) \left(\rho_k^{-m} + O(1)\right).$$

Moreover as the number of divisors (see [15]) of $n$ is $O(n^\delta)$ for any positive $\delta$, we can write for any positive $\delta < 1$

$$\ell_{n,k} = \frac{1}{n} \sum_{m|n} \mu\left(\frac{n}{m}\right) \rho_k^{-m} + O\left(n^{\delta-1}\right).$$

Finally replacing $\rho_k$ by $1/2 + \epsilon_k$, we obtain

$$\ell_{n,k} = \frac{2^n}{n}(1+2\epsilon_k)^{-n} + O\left(\frac{2^{n/2}}{n}\right).$$

Making use of the following equalities

$$p_{n,k} = \frac{\ell_{n,k}}{\mathrm{Card}(\mathcal{L}_n)} \qquad \text{and} \qquad \mathrm{Card}(\mathcal{L}_n) = \frac{2^n}{n}\left(1 + O\left(2^{-n/2}\right)\right),$$

we get the announced result. □

Next result gives an interval to which belongs almost surely the length of the longest runs of $a$'s in a Lyndon word. In this way we restrict our combinatorial model over Lyndon words, leaving apart only a negligible portion of them.

**Lemma 15** *The length $k$ of the longest runs of $a$'s in a word $w \in \mathcal{L}_n$ satisfies*

$$\Pr\{k(w) \in [\log_2 n - \log_2\log_2 n - 1, 2\log_2 n[\} = 1 - O\left(\frac{1}{n}\right).$$

**PROOF.** From Proposition 14, one has for the length $k(w)$ of the longest run of $a$'s in a word $w$ of $\mathcal{L}_n$

$$\Pr\{k(w) < k\} = (1+2\epsilon_k)^{-n} + O\left(2^{-n/2}\right). \tag{3}$$

The inequality $\log(1+x) > x\log 2$ is true for $0 < x < 1$ gives after simple algebra the result that is the value of $k$ for which $\Pr\{k(w) < k\} \leq \frac{1}{n}$, namely $k = \log_2 n - \log_2\log_2 n - 1$.

Again, in Equation (3), the inequality $\log(1+x) < x$ (true for all $x$) and the estimation $2\epsilon_k = 2^{-(k+1)}\left(1 + O\left(k2^{-k}\right)\right)$ give the values of $k$ for which $\Pr\{k(w) < k\} \leq 1 - \frac{1}{n}$, namely $k = 2\log_2 n$. □

**Remark 16** *As* $\mathrm{Card}(\mathcal{L}_n \setminus a\mathcal{L}_{n-1}) = \Omega(\mathrm{Card}(\mathcal{L}_n))$, *Lemma 15 can be stated for* $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ *instead of* $\mathcal{L}_n$.

In what follows $\mathcal{I}_n$ *denotes the interval* $[\log_2 n - \log_2\log_2 n - 1, 2\log_2 n[$.

*Number of factors of the max-run decomposition.*

Now we establish a bound on the number of factors in the max-run decomposition.

**Lemma 17** *Let $w$ be a Lyndon word of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ with $k(w) \in \mathcal{I}_n$. The number $m$ of factors in the max-run decomposition satisfies*

$$\Pr\{m \geq 2\log_2 n\} = O\left(\frac{\log n}{n}\right).$$

**PROOF.** Denote $\mathfrak{L}^1_{k,\geq m}$ the set of words of $\mathfrak{L}^1_k$ with more than $m$ runs of $a$'s of length $k-1$ and $\mathfrak{L}^2_{k,\geq m}$ the set of words of $\mathfrak{L}^2_k$ with more than $m$ runs of $a$'s of length $k$. We want to estimate the ratio

$$\frac{\sum_{k\in\mathcal{I}_n} \mathrm{Card}((\mathfrak{L}^1_{k,\geq m_0} \cup \mathfrak{L}^2_{k,\geq m_0}) \cap A^n)}{\sum_{k\in\mathcal{I}_n} \mathrm{Card}((\mathfrak{L}^1_k \cup \mathfrak{L}^2_k) \cap A^n)}$$

for $m_0 = 2\log_2 n$.

First of all $(\mathfrak{L}^1_k \cup \mathfrak{L}^2_k) \cap A^n$ is the set of Lyndon words of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ beginning with a longest run of $a$'s of length $k$. From $\mathrm{Card}(\mathcal{L}_n \setminus a\mathcal{L}_{n-1}) = \frac{2^{n-1}}{n}(1+O(\frac{1}{n}))$ and Lemma 15, we get

$$\sum_{k\in\mathcal{I}_n} \mathrm{Card}((\mathfrak{L}^1_k \cup \mathfrak{L}^2_k) \cap A^n) = \frac{2^{n-1}}{n}\left(1 + O\left(\frac{1}{n}\right)\right). \tag{4}$$

In order to estimate the remaining part of the ratio, we introduce the set $\mathcal{W}_{k,m}$ of words beginning by a longest run of $a$'s of length $k$ and containing at least $m$ longest runs of $a$'s

$$\mathcal{W}_{k,m} = a^k b(\mathcal{X}^*_k a^k b)^{m-1}\mathcal{X}^*_{k+1}.$$

Then, denoting $W_{k,m}(z)$ the generating function of $\mathcal{W}_{k,m}$,

$$\mathrm{Card}((\mathfrak{L}^1_{k+1,\geq m} \cup \mathfrak{L}^2_{k,\geq m}) \cap A^n) \leq [z^n]W_{k,m}(z).$$

Indeed $(\mathfrak{L}^2_{k,\geq m} \cap A^n) \subset (\mathcal{W}_{k,m} \cap \mathcal{L}_n)$ and $a^{-1}(\mathfrak{L}^1_{k+1,\geq m} \cap A^n) \subset (\mathcal{W}_{k,m} \cap A^{n-1}) \setminus \mathcal{L}_{n-1}$. Moreover

$$\mathrm{Card}((\mathcal{W}_{k,m} \cap A^{n-1}) \setminus \mathcal{L}_{n-1}) \leq \mathrm{Card}((\mathcal{W}_{k,m} \cap A^n) \setminus \mathcal{L}_n),$$

since by adding a $b$ just after the first occurrence of $a^k b$ we define an injection from the first set on the second one. Thus, setting $\mathcal{I}_n = [k_1, k_2[$, we obtain

$$\sum_{k\in\mathcal{I}_n} \mathrm{Card}((\mathfrak{L}^1_{k,\geq m} \cup \mathfrak{L}^2_{k,\geq m}) \cap A^n) \leq \sum_{k=k_1-1}^{k_2} [z^n]W_{k,m}(z).$$

Moreover considering the ambiguous language $(a^k b \mathcal{X}_{k+1}^*)^m$, we get the following bound

$$[z^n]W_{k,m}(z) \leq [z^n]\left(\frac{z^{k+1}}{1-X_{k+1}(z)}\right)^m. \tag{5}$$

Since we shall consider $m = 2\log_2 n$, here we can not use directly a formula like in (1) to extract coefficients for this rational function. So using the saddle point method, we establish a bound for its coefficients.

**Lemma 18** *Let $F(z)$ be analytic function such that $F(1) = 1$ and $F'(1) \neq 0$, and $G(z) = (1-F(z))^{-m}$. When $m = O(\log n)$ there exists $c < 1$ such that for $n$ large enough*

$$[z^n]\frac{1}{(1-F(z))^m} \leq (1+c)\left(\frac{en}{mF'(1)}\right)^m.$$

**PROOF.** Using saddle point bound [7,20] on function $\frac{1}{(1-F(z))^m}$ yields that

$$[z^n]\frac{1}{(1-F(z))^m} \leq \frac{1}{(1-F(\xi(n)))^m}(\xi(n))^{-n} \tag{6}$$

where $\xi$ is the unique positive solution in $]0,1[$ of the equation

$$\xi\frac{G'(\xi)}{G(\xi)} = n.$$

The last equation is equivalent to

$$\xi\frac{F'(\xi)}{1-F(\xi)} = \frac{n}{m}.$$

Thus replacing in (6) gives

$$[z^n]\frac{1}{(1-F(z))^m} \leq \left(\frac{\xi n}{mF'(\xi)}\right)^m \frac{1}{\xi^n}.$$

Setting $\xi = 1-x$ and studying Taylor coefficients of $F(1-x)$, we obtain

$$x = \frac{m}{n}(1+o(1)).$$

Using the standard estimate $(1-x)^n \sim e^{-nx}$, one can write for all $c > 0$ and $n$ large enough

$$[z^n]\frac{1}{(1-F(z))^m} \leq (1+c)\left(\frac{ne}{mF'(1)}\right)^m,$$

concluding the proof of the lemma. □

Since $\rho_{k+1}$ is the smallest root of $X_{k+1}(z) - 1$ and

$$[z^n]\left(\frac{z^{k+1}}{1 - X_{k+1}(z)}\right)^m = \frac{1}{\rho_{k+1}^{n-m(k+1)}}[z^{n-m(k+1)}]\frac{1}{(1 - X_{k+1}(\rho_{k+1}z))^m}$$

applying Lemma 18 and using inequality (5) one has for $c > 0$ and $m = O(\log n)$

$$[z^n]W_{k,m}(z) \leq (1+c)\frac{1}{\rho_{k+1}^n}\left(\frac{ne\rho_{k+1}^{k+1}}{mX'_{k+1}(\rho_{k+1})}\right)^m.$$

Denoting by $b_k$ the last quantity, we get $\sum_{k=k_1-1}^{k_2}[z^n]W_{k,m}(z) \leq \sum_{k=k_1-1}^{k_2} b_k$. Since $\rho_{k+1}^{k+1} = 2^{-(k+1)}(1 + O(k2^{-k}))$ and by Equation 2, we have

$$\left(\frac{ne\rho_{k+1}^{k+1}}{mX'_{k+1}(\rho_{k+1})}\right)^m = \left(\frac{ne}{2^{k+3}m}\right)^m\left(1 + O\left(\frac{mk}{2^k}\right)\right).$$

Moreover for $k \in \mathcal{I}_n$,

$$\rho_k^{-n} = \frac{2^n}{(1 + 2^{-(k+1)} + O(k2^{-2k}))^n} = 2^n e^{-n/2^{k+1}}\left(1 + O\left(\frac{nk}{2^{2k}}\right)\right). \qquad (7)$$

This entails for $k = O(\log n)$ and $m = O(\log n)$,

$$b_k = (1+c)\,2^n e^{-n/2^{k+2}}\left(\frac{ne}{2^{k+3}m}\right)^m\left(1 + O\left(\frac{\log^3 n}{n}\right)\right).$$

When $n$ and $m$ are fixed, $b_k$ is maximal for $k = \log_2(n/m) - 2$ and is equal to $O(2^{n-m})$. So for $m_0 = 2\log_2 n$,

$$\sum_{k\in\mathcal{I}_n}\mathrm{Card}((\mathfrak{L}^1_{k,\geq m_0} \cup \mathfrak{L}^2_{k,\geq m_0}) \cap A^n) \leq \sum_{k=k_1-1}^{k_2} b_k = O\left(\frac{2^n}{n^2}\log n\right).$$

Finally using (4), we obtain

$$\frac{\sum_{k\in\mathcal{I}_n}\mathrm{Card}\left((\mathfrak{L}^1_{k,\geq m_0} \cup \mathfrak{L}^2_{k,\geq m_0}) \cap A^n\right)}{\sum_{k\in\mathcal{I}_n}\mathrm{Card}((\mathfrak{L}^1_k \cup \mathfrak{L}^2_k) \cap A^n)} = O\left(\frac{\log n}{n}\right),$$

concluding the proof. □

*Nature of the factors of the max-run decomposition.*

Our goal in the following is to distinguish for the lexicographical order the factors of the max-run decomposition. Recall that any word of $w \in \mathcal{L} \setminus a\mathcal{L}$ can be written $w = f_1 \dots f_m$ where $f_1 = a^{k(w)}bw_1$, $f_i = a^K bw_i$ for $i > 1$, $w_i \in \mathcal{X}_K^*$

for all $i$ and $K = k(w)$ or $k(w) - 1$. The $w_i$ are called the *interleaving words*. We first prove that all interleaving words are of length at least $K$.

We introduce the set $\mathcal{P}_K$ of words $w \in \mathcal{X}_K^*$ such that denoting by $w[i]$ the $i$-th letter of $w$

$$K \leq |w| \leq 2K - 1 \quad \text{and} \quad \forall i \in \{K, \ldots, |w| - 1\}, w[i] = a.$$

For example for $K = 3$, we have $\mathcal{X}_3 = \{b, ab, aab\}$ and the set $\mathcal{P}_3$ is

$$\mathcal{P}_3 = \{baab, abaab, bbaab, bab, abab, bbab, bbb, abb, aab\}.$$

The following formula stresses the role of the last word of $\mathcal{X}_K$ in the factorization of words of $\mathcal{P}_K$

$$\mathcal{P}_K = \left(\cup_{j=0}^{K-2} A^j b\, a^{K-1} b\right) \cup \left(\cup_{j=1}^{K-2} A^j b\, a^{K-2} b\right) \cup \cdots \cup \left(\cup_{j=K-2}^{K-2} A^j b\, ab\right) \cup A^{K-1} b.$$

Usual translation to generating functions entails

$$\begin{aligned} P_K(z) &= \sum_{j=2}^{K} z^{j+1} \left( \sum_{i=K-j}^{K-2} (2z)^i \right) + z(2z)^{K-1} \\ &= z(2z)^{K-1} \left( \frac{z^{K+1} + 4\, z^2\, (\frac{1}{2})^{K+1} - 4\, z\, (\frac{1}{2})^{K+1}}{(2\, z - 1)\, (z - 1)} - \frac{z}{z-1} + 1 \right). \end{aligned}$$

The closed form of this formula is not as important as the fact that

$$\text{for } x = O\left(\frac{1}{2^K}\right), \quad P_K\left(\frac{1}{2} + x\right) = 1 + O(Kx) \quad \text{and} \quad P_k'\left(\frac{1}{2} + x\right) = O(K). \tag{8}$$

**Lemma 19** *Let $w$ be a Lyndon word of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ with $k(w) \in \mathcal{I}_n$. In its max-run decomposition, all interleaving words are of length at least $K$ with probability*

$$1 - O\left(\frac{\log^2 n}{n}\right).$$

**PROOF.** We distinguish two cases depending on the values of $K$, namely $k$ and $k - 1$. More precisely we prove that all longest runs $a^k b$ in Lyndon words of length $n$ are followed by words of $\mathcal{P}_k$ with high probability. If the longest run $a^k b$ is unique then all runs of $a^{k-1} b$ are also followed by words of $\mathcal{P}_{k-1}$ with high probability.

We consider the code $\mathcal{C} = a^k b \mathcal{P}_k\ \mathcal{X}_k^*$ and the set $\mathcal{C}_k^{\leftrightarrow}$ of primitive cycles over the code $\mathcal{C}$, *i.e.* the set of Lyndon words beginning with $k$ $a$'s and such that all occurences of $a^k b$ are followed by words of $\mathcal{P}_k$. Applying Proposition 13 with

$C(z) = \frac{z^{k+1}P_k(z)}{1-X_k(z)}$ yields the generating function of $\mathcal{C}_k^{\leftrightarrow}$

$$C_k^{\leftrightarrow}(z) = \sum_{m\geq 1} \frac{\mu(m)}{m} \log\left(\frac{1-X_k(z^m)}{1-X_k(z^m)-z^{m(k+1)}P_k(z^m)}\right).$$

Moreover for words with a unique longest run of length $k$, all possible occurences of $a^{k-1}b$ in $a^{-1}w$ must be separated by words of $\mathcal{P}_{k-1}$. So instead of $\mathcal{C}_k^{\leftrightarrow}$ we are bound to study the set

$$\mathcal{D}_k^{\leftrightarrow} = \left(\mathcal{C}_k^{\leftrightarrow} \setminus a^k b\mathcal{P}_k\mathcal{X}_k^*\right) \cup a^k b\mathcal{P}_{k-1}\mathcal{X}_{k-1}^* \left(a^{k-1}b\mathcal{P}_{k-1}\mathcal{X}_{k-1}\right)^*.$$

Its generating function can be written

$$D_k^{\leftrightarrow}(z) = C_k^{\leftrightarrow}(z) - \Delta_k(z) \text{ with } \Delta_k(z) = \frac{z^{k+1}P_k(z)}{1-X_k(z)} - \frac{z^{k+1}P_{k-1}(z)}{1-X_{k-1}(z)-z^kP_{k-1}(z)}.$$

We shall compare the cardinality of the set $\mathcal{L}_n^{\leftrightarrow} = \cup_{k\in\mathcal{I}_n}(\mathcal{D}_k^{\leftrightarrow} \cap A^n)$ namely

$$\mathrm{Card}(\mathcal{L}_n^{\leftrightarrow}) = \sum_{k\in\mathcal{I}_n} [z^n] D_k^{\leftrightarrow}(z)$$

with the number of Lyndon words of length $n$. Let $\varrho_k$ be the smallest root of $1-X_{k-1}(z)-z^kP_{k-1}(z)$. We can prove as in Section 4.1 that $\varrho_k$ is simple and belongs to $[1/2, 1[$. Using the bootstrapping method and the estimates (8) of $P_k$ and $P_k'$, we obtain

$$\varrho_k = \frac{1}{2} + \frac{1}{2^{k+2}} + O\left(\frac{k}{2^{2k}}\right) = \rho_k + O\left(\frac{k}{2^{2k}}\right). \tag{9}$$

Let $c_{n,k}^{\leftrightarrow} = [z^n]C_k^{\leftrightarrow}(z)$. By usual coefficient extraction we have

$$c_{n,k}^{\leftrightarrow} = \frac{1}{n}\left(\frac{1}{\varrho_{k+1}^n} - \frac{1}{\rho_k^n}\right) + O\left(\frac{2^{n/2}}{n}\right).$$

From Equations (7) and (9) we get

$$\sum_{k\in\mathcal{I}_n} c_{n,k}^{\leftrightarrow} = \frac{2^n}{n}\left(e^{-n/2^{k_2+2}} - e^{-n/2^{k_1+1}}\right) + \frac{2^n}{n}\sum_{k\in\mathcal{I}_n} e^{-n/2^{k+2}} O\left(\frac{nk}{2^{2k}}\right).$$

By definition of $\mathcal{I}_n$, $e^{-n/2^{k_2+2}} - e^{-n/2^{k_1+1}} = 1+O(1/n)$. Moreover as $\left(\frac{n}{2^k}\right)^2 \exp(-\frac{n}{2^{k+2}})$ is uniformly bounded for $k > 0$, we obtain

$$\sum_{k\in\mathcal{I}_n} c_{n,k}^{\leftrightarrow} = \frac{2^n}{n}\left(1 + O\left(\frac{\log^2 n}{n}\right)\right). \tag{10}$$

On the other hand, using again coefficient extraction of rational functions and using Equations (2), (7), (9) and we have

$$[z^n]\frac{z^{k+1}P_{k-1}(z)}{1-X_{k-1}(z)-z^kP_{k-1}(z)} = \frac{\varrho_k^{k+1}P_{k-1}(\varrho_k)\ \varrho_k^{-(n+1)}}{X'_{k-1}(\varrho_k)+(k+1)\varrho_k^kP_k(\varrho_k)+\varrho_k^{k+1}P'_{k-1}(\varrho_k)}+O(1)$$

$$[z^n]\frac{z^{k+1}P_k(z)}{1-X_k(z)} = \frac{\rho_k^{k+1}P_k(\rho_k)}{X'_k(\rho_k)}\frac{1}{\rho_k^{n+1}}+O(1).$$

As $X'_k(z)=X'_{k-1}(z)+(k+1)z^k$, we get by Equations (7) and (9)

$$[z^n]\Delta_k(z) = 2^n\frac{1}{2^{k+1}}e^{-n/2^{k+2}}O\left(\frac{nk}{2^{2k}}\right).$$

Again as $\left(\frac{n}{2^k}\right)^3\exp(-\frac{n}{2^{k+2}})$ is uniformly bounded for $k>0$, we obtain

$$\sum_{k\in\mathcal{I}_n}[z^n]\Delta_k(z) = O\left(\frac{2^n\log^2 n}{n^2}\right).$$

Consequently using Equation (10), we get $\mathrm{Card}(\mathcal{L}_n^{\leftrightarrow})=\frac{2^n}{n}\left(1+O\left(\frac{\log^2 n}{n}\right)\right)$ and

$$\frac{\mathrm{Card}(\mathcal{L}_n)}{\mathrm{Card}(\mathcal{L}_n^{\leftrightarrow})} = 1+O\left(\frac{\log^2 n}{n}\right).$$

Thus almost all Lyndon words of length $n$ belong to $\mathcal{L}_n^{\leftrightarrow}$. Finally since $\mathrm{Card}(\mathcal{L}_n\setminus a\mathcal{L}_{n-1})=\Omega(\mathrm{Card}(\mathcal{L}_n))$ the property on the length of the interleaving words also holds on $\mathcal{L}_n\setminus a\mathcal{L}_{n-1}$ with an error term of same order. □

To compare for the lexicographical order two factors beginning with a longest run of $a$'s of length $K$, it remains to distinguish at most $m=2\log_2 n$ interleaving words of $\mathcal{P}_K\mathcal{X}_K^*$.

**Lemma 20** *Let $w$ be a Lyndon word of $\mathcal{L}_n\setminus a\mathcal{L}_{n-1}$ with $k(w)\in\mathcal{I}_n$ having a max-run decomposition into $m=O(\log n)$ factors. The $m$ interleaving words have pairwise distinct prefixes in $\mathcal{P}_K$ with probability greater than*

$$1-O\left(\frac{\log^3 n}{n}\right).$$

**PROOF.** From previous lemma, interleaving words are longer than $K$ with probability $1-O(\log^2 n/n)$. Thus we focus on the subsets $\mathcal{Q}_{n,K}^{\leftrightarrow}$ of $\mathcal{L}_n\setminus a\mathcal{L}_{n-1}$ of Lyndon words with a max-run decomposition where all interleaving words are in $\mathcal{P}_K\mathcal{X}_K^*$. We shall prove that all the prefixes in $\mathcal{P}_K$ of these words are

pairwise distinct with high probability and that the restriction on the length of the interleaving words does not affect the order of the error term.

Given a sequence of positive integers $\boldsymbol{m} = (m_1, \ldots, m_\ell)$ and an increasing sequence of positive integers $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\ell)$, define the set $\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ as the set of Lyndon words $w \in \mathcal{Q}^{\leftrightarrow}_{n,K}$ such that

(i) $w$ admits a decomposition into $m = \sum_{i=1}^{\ell} m_i$ factors;
(ii) for $i \in \{1, \ldots, \ell\}$, $w$ has exactly $m_i$ interleaving words with prefixes of length $\omega_i$ in $\mathcal{P}_K$.

This defines a partition of $\mathcal{Q}^{\leftrightarrow}_{n,K}$ according to $\boldsymbol{m}$ and $\boldsymbol{\omega}$. Denote by $\mathcal{Q}^{\neq}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ the subset of words of $\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ with interleaving words having pairwise distinct prefixes in $\mathcal{P}_K$.

Let $\mathcal{S}$ be a set of $m$ distinct words of $\mathcal{P}_K$ of total length $N = \sum \omega_i m_i$. There are $(m-1)(m-1)!$ possible way of ordering $\mathcal{S}$ so that the first word is not the smallest, yielding a word of $\mathfrak{L}^1_k$, and $(m-1)!$ possible ways of ordering $\mathcal{S}$ so that the first word is the smallest, yielding a word of $\mathfrak{L}^2_k$. So completing the words up to length $n$ with $a^K b$ (possibly $a^{K+1}b$ at the beginning) before each word of $\mathcal{P}_K$ and words of $\mathcal{X}^*_K$ after each word of $\mathcal{P}_K$, we obtain

$$\left[z^{n-m(K+1)-N}\right] (m-1)!\, \frac{1+(m-1)z}{(1-X_K(z))^m}$$

Lyndon words for a given set $\mathcal{S}$. If the words of $\mathcal{S}$ are not distinct, then the last quantity is just an upper bound for the number of Lyndon words one can obtain.

Let us fix a sequence of positive integers $\boldsymbol{m} = (m_1, \ldots, m_\ell)$ and an increasing sequence of positive integers $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\ell)$ and denote by $P_{K,n} = \mathrm{Card}(\mathcal{P}_K \cap A^n)$. As

$$\forall (a_1, a_2, \ldots, a_p) \in [0,1]^p, \quad \prod_{i=1}^{p} (1-a_i) \geq 1 - \sum_{i=1}^{p} a_i,$$

we have the following chain of inequalities provided the sets $\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ and $\mathcal{Q}^{\neq}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ are not empty

$$\frac{\mathrm{Card}(\mathcal{Q}^{\neq}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})}{\mathrm{Card}(\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})} \geq \frac{\prod_{i=1}^{\ell} \binom{P_{K,\omega_i}}{m_i}}{\prod_{i=1}^{\ell} P_{K,\omega_i}^{m_i}} \geq \prod_{i=1}^{\ell} \left(1 - \frac{m_i^2}{P_{K,\omega_i}}\right).$$

Finally since $\sum m_i^2 \leq (\sum m_i)^2$ and $P_{K,\omega_i} \geq 2^{K-1}$ for all $i$, we have

$$\frac{\mathrm{Card}(\mathcal{Q}^{\neq}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})}{\mathrm{Card}(\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})} \geq 1 - \frac{m^2}{2^{K-1}}.$$

So for $m = O(\log n)$ and $K > \log_2 n - \log_2 \log_2 n - 2$ the ratio becomes

$$\frac{\mathrm{Card}(\mathcal{Q}^{\neq}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})}{\mathrm{Card}(\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}})} = 1 - O\left(\frac{\log^3 n}{n}\right).$$

Since $\mathcal{Q}^{\leftrightarrow}_{n,K}$ is the disjoint union of $\mathcal{Q}^{\leftrightarrow}_{n,K,\boldsymbol{m},\boldsymbol{\omega}}$ for all $(\boldsymbol{m},\boldsymbol{\omega})$, the result is also true for $\mathcal{Q}^{\leftrightarrow}_{n,K}$. Finally as the error term $O(\log^3 n/n)$ is uniform for all subsets $\mathcal{Q}^{\leftrightarrow}_{n,K}$ and the error term $O(\log^2 n/n)$ coming from the hypothesis on the length of the interleaving words is of smaller order, the property holds for words $w$ of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ with $k(w) \in \mathcal{I}_n$ and $m = O(\log n)$. □

### *4.2 An involution over $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$*

We now introduce an involution on almost all the set $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ such that the sum of the lengths of the right factors of $w$ and its image is approximatively $|w|$.

To achieve this goal we partition the set $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ in two subsets,

$$\mathcal{L}_n \setminus a\mathcal{L}_{n-1} = \mathcal{G}_n \cup \mathcal{B}_n.$$

The set $\mathcal{G}_n$ is the set of words in $\mathcal{L}_n \setminus \mathcal{L}_{n-1}$ whose max-run decomposition $a^k b w_1 \dots a^K b w_m$ verifies

(i) $k \in \mathcal{I}_n$;
(ii) $m < 2\log_2 n$;
(iii) the interleaving words $w_i$ have pairwise distinct prefixes in $\mathcal{P}_K$.

For any word $w = a^k bu \cdot a^K bv \in \mathcal{G}_n$ we define $\varphi(w)$ as

$$\varphi(w) = a^k bu'v''a^K bv'u'',$$

with $u = u'u''$, $v = v'v''$ and $u'$ and $v'$ in $\mathcal{P}_K$.

The key fact is that, globally, $\varphi$ preserves the runs of $a$'s and the prefixes in $\mathcal{P}_K$ of the interleaving words of the max-run decomposition.

If $\ell$ is a Lyndon word, we denote by $\mathtt{right}(\ell)$ the right factor of $\ell$.

**Lemma 21** *Under the uniform distribution over $\mathcal{G}_n$ the average length of the right factor of the standard factorization is*

$$\frac{n}{2}\left(1 + O\left(\frac{\log n}{n}\right)\right).$$

**PROOF.** We prove that $\varphi$ is an involution on $\mathcal{G}_n$ and the sum of the lengths of the right factors of a word $w \in \mathcal{G}_n$ and $\varphi(w)$ is about $|w|$.

Let $w \in \mathcal{G}_n$ with standard factorization

$$w = a^k b w_1 \dots a^K b w_{d-1} \ \cdot \ a^K b w_d \dots a^K b w_m$$

with $w_i \in \mathcal{P}_K \mathcal{X}_K^*$ for $1 \leq i \leq m$, then

$$\varphi(w) = a^k b w_1' w_d'' a^K b w_{d+1} \dots a^K b w_m a^K w_d' w_1'' a^K b w_2 \dots a^K b w_{d-1}$$

with $w_1 = w_1' w_1''$, $w_d = w_d' w_d''$ and $w_1', w_d'$ in $\mathcal{P}_K$.

By definition of $\varphi$, $\varphi(w) \in a^k b \mathcal{P}_K \mathcal{X}_K^* (a^K b \mathcal{P}_K \mathcal{X}_K^*)^+$. Moreover for a word $w$ of $\mathcal{G}_n$, the position of the smallest proper suffix of $\varphi(w)$ can be easily determined. Indeed $\varphi$ preserves the relative order between $a^k b w_1' < a^K b w_d' < a^K b w_i$ for $i \neq 1, d$. Thus $\varphi(w)$ is a Lyndon word and the standard factorization of $\varphi(w)$ is

$$\varphi(w) = a^k b w_1' w_d'' a^K b w_{d+1} \dots a^K b w_m \ \cdot \ a^K w_d' w_1'' a^K b w_2 \dots a^K b w_{d-1}.$$

So $\varphi(w) \in \mathcal{G}_n$ and $\varphi$ is an involution on $\mathcal{G}_n$: $\varphi(\varphi(w)) = w$ for $w \in \mathcal{G}_n$.

Moreover for any word $w$ of $\mathcal{G}_n$

$$|\mathtt{right}(w)| + |\mathtt{right}(\varphi(w))| = |w| - (k - K) + |w_d'| - |w_1'|,$$

where $k - K \in \{0, 1\}$.

By definition, the lengths of prefixes $w_d'$ and $w_1'$ are in $[K, 2K-1]$, so $||w_d'| - |w_1'|| < K$. As $k \in \mathcal{I}_n$ and $k - K \in \{0, 1\}$ we get that $||w_d'| - |w_1'|| = O(\log n)$. Finally as $\varphi$ is an involution on $\mathcal{G}_n$ we obtain

$$\begin{aligned} 2 \sum_{w \in \mathcal{G}_n} |\mathtt{right}(w)| &= \sum_{w \in \mathcal{G}_n} (|\mathtt{right}(w)| + |\mathtt{right}(\varphi(w))|) \\ &= (n + O(\log n)) \ \mathrm{Card}(\mathcal{G}_n), \end{aligned}$$

concluding the proof. □

Now we compute the total contribution of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ to the mean value of the standard right factor

$$\sum_{w \in \mathcal{L}_n \setminus a\mathcal{L}_{n-1}} |\mathtt{right}(w)| = \sum_{w \in \mathcal{G}_n} |\mathtt{right}(w)| + \sum_{w \in \mathcal{B}_n} |\mathtt{right}(w)|.$$

Using Lemma 21 and the fact that $|\mathtt{right}(w)| \leq |w|$ for any Lyndon word $w$,

we get

$$\sum_{w\in\mathcal{L}_n\setminus a\mathcal{L}_{n-1}} |\mathtt{right}(w)| = \frac{n}{2}\left(1+O\left(\frac{\log n}{n}\right)\right)\mathrm{Card}(\mathcal{G}_n)+O(n)\times\mathrm{Card}(\mathcal{B}_n).$$

Moreover Lemmas 15, 17 and 20 match exactly the conditions (i), (ii) and (iii) which characterize the set $\mathcal{G}_n$. It leads to the estimate

$$\mathrm{Card}(\mathcal{G}_n) = \mathrm{Card}(\mathcal{L}_n\setminus a\mathcal{L}_{n-1})\left(1-O\left(\frac{\log^3 n}{n}\right)\right).$$

Consequently we get

$$\sum_{w\in\mathcal{L}_n\setminus a\mathcal{L}_{n-1}} |\mathtt{right}(w)| = \frac{n}{2}\ \mathrm{Card}(\mathcal{L}_n\setminus a\mathcal{L}_{n-1})\left(1+O\left(\frac{\log^3 n}{n}\right)\right).$$

Finally as

$$\mathrm{Card}(\mathcal{L}_n\setminus a\mathcal{L}_{n-1}) = \mathrm{Card}(\mathcal{L}_n)\left(\frac{1}{2}+O\left(\frac{1}{n}\right)\right),$$

the total contribution of of $\mathcal{L}_n\setminus a\mathcal{L}_{n-1}$ to the mean value of the standard right factor is

$$\frac{n}{4}\left(1+O\left(\frac{\log^3 n}{n}\right)\right),$$

concluding the proof of Proposition 12 and Theorem 10.

## 5 Algorithms and experimental results

In this section we give two linear algorithms to generate random Lyndon words of a given length $n$ and to compute the standard factorization of a Lyndon word.

There exists an algorithm SmallestConjugate($u$), proposed by Booth [18,4], that gives the smallest conjugate a random Lyndon word of length $n$ in linear time. We use it to make a reject algorithm which is efficient to generate randomly a Lyndon word of length $n$:

```
RandomLyndonWord(n)        // return a random Lyndon word
  string u, v;
  do
    u = RandomWord(n);        // u is a random word of A^n
    v = SmallestConjugate(u);        // v is the smallest conjugate of u
  until (length(v) == n);        // v is primitive
  return v;
```

The algorithm RandomLyndonWord computes uniformly a Lyndon word.

**Lemma 22** *The average complexity of RandomLyndonWord(n) is linear.*

**PROOF.** Each execution of the **do** ...**until** loop is done in linear time. The condition is not satisfied when $u$ is a conjugate of a periodic word $v^p$ with $p > 1$. This happens with probability $O(\frac{n}{2^{n/2}})$. Thus the loop is executed a bounded number of times in the average. □

**Lemma 23** *Let $\ell$ be a Lyndon word which is not a letter. Let $\ell_1 \dots \ell_k$ be the factorization of $a^{-1}\ell$ into a nonincreasing sequence of Lyndon words. The right factor of $\ell$ in its standard factorization is $\ell_k$.*

**PROOF.** By Theorem 2, $\ell_k$ is the smallest suffix of $a^{-1}\ell$, thus it is the smallest proper suffix of $\ell$. □

Let the Duval(string $u$, int $k$, array $pos$) be the function which computes the factorization of $u$ into a nonincreasing sequence of Lyndon words by Duval's algorithm [8]. It stores in an array $pos$ of size $k$ the positions of the factors. The following algorithm computes the right factor of a Lyndon word $\ell$ which is not a letter

```
RightFactor(string l[1..n])
  array pos;
  int k;
  Duval(l[2..n], k, pos);    // omit the first letter and apply Duval's algorithm
  return l[pos[k]..n];       // return the last factor
```

This algorithm is linear in time since Duval's algorithm is linear.

Figures 1, 3 and 2 present some experimental results obtained with our algorithms.

### Open problems

The results obtained in this paper are only a first step toward the average case-analysis of the tree obtained from a Lyndon word by successive standard factorizations. In order to study the height of these trees, a better insight of the distribution of the right factors of words of $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ is needed.

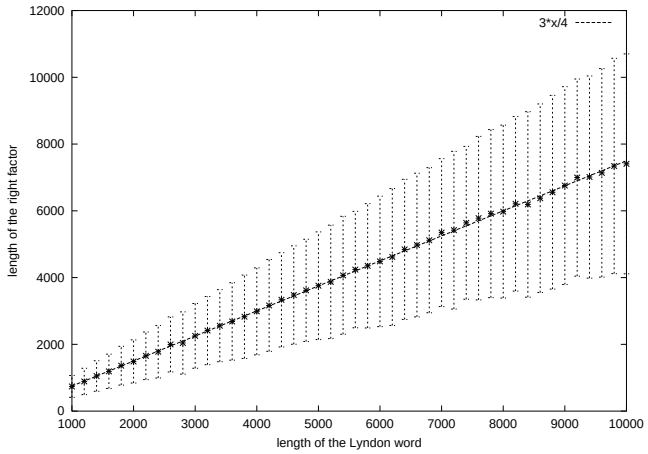


Fig. 1. Average length of the right factor of random Lyndon words of length from $1,000$ to $10,000$. Each plot is computed with $1,000$ words. The error bars represent the standard deviation.

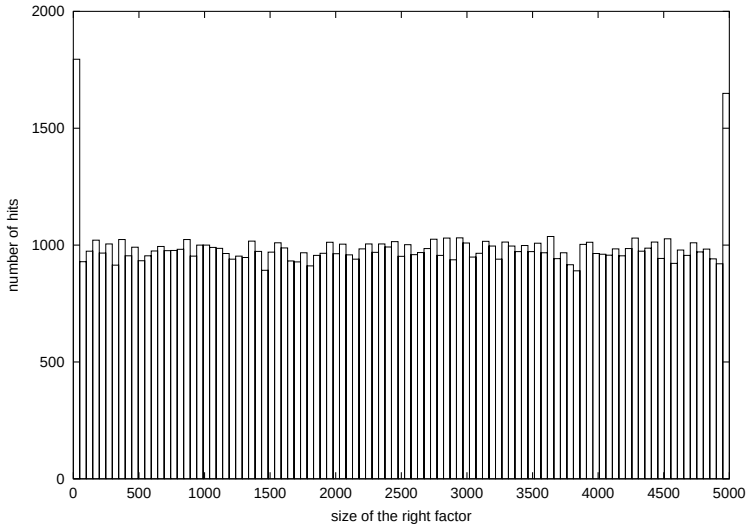


Fig. 2. Distribution of the length of the right factor over $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$. We generated $100,000$ random Lyndon words of length $5,000$.

Figure 3 hints a very strong equi-repartition property of the length of the right factor over this set. This suggests a particular subdivision process at each node of the factorization tree which needs further investigations.

## References

[1] J. Berstel, L. Boasson, The set of lyndon words is not context-free, Bull. Eur. Assoc. Theor. Comput. Sci. EATCS 63 (1997) 139–140.

[2] J. Berstel, D. Perrin, Theory of codes, Academic Press, 1985.

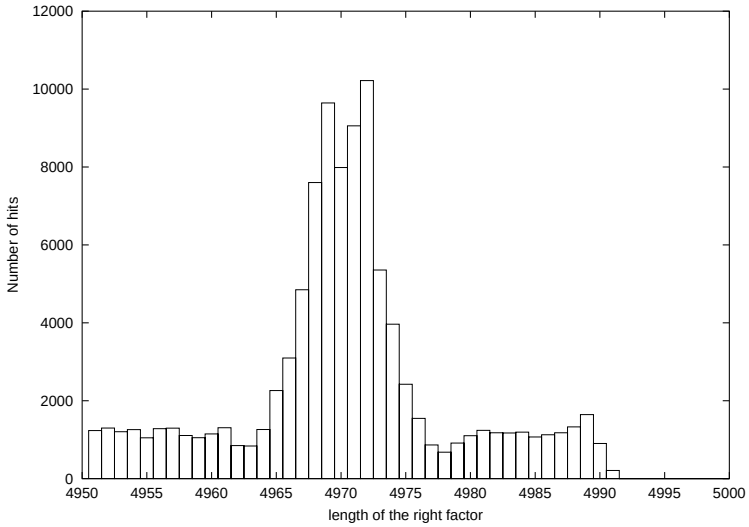


Fig. 3. Zoom on the distribution of the length of the right factor over $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$.

[3] J. Berstel, M. Pocchiola, Average cost of Duval's algorithm for generating Lyndon words, Theoret. Comput. Sci. 132 (1-2) (1994) 415–425.

[4] K. S. Booth, Lexicographically least circular substrings, Inform. Process. Lett. 10(4-5) (1980) 240–242.

[5] H. Cartan, Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes, Hermann, 1985.

[6] K. Chen, R. Fox, R. Lyndon, Free differential calculus IV: The quotient groups of the lower central series, Ann. Math. 58 (1958) 81–95.

[7] N. G. de Bruijn, Asymptotic Method in Analysis, North Holland, 1961.

[8] J.-P. Duval, Factorizing words over an ordered alphabet, Journal of Algorithms 4 (1983) 363–381.

[9] W. Feller, An introduction to Probability Theory and Its Applications, 3rd Edition, Vol. 1, Wiley, 1968.

[10] P. Flajolet, X. Gourdon, D. Panario, The complete analysis of a polynomial factorization algorithm over finite fields, Journal of Algorithms 40 (2001) 37–81.

[11] P. Flajolet, R. Sedgewick, Analytic combinatorics–symbolic combinatorics, Book in preparation, (Individual chapters are available as INRIA Research reports at http://www.algo.inria.fr/flajolet/publist.html) (2002).

[12] P. Flajolet, M. Soria, The cycle construction, SIAM J. Disc. Math. 4 (1991) 58–60.

[13] H. Fredricksen, J. Maiorana, Necklaces of beads in $k$ colors and $k$-ary de Bruijn sequences, Discrete Math. 23 (3) (1978) 207–210.

[14] S. Golomb, Irreducible polynomials, synchronizing codes, primitive necklaces and cyclotomic algebra, in: Proc. Conf Combinatorial Math. and Its Appl., Univ. of North Carolina Press, Chapel Hill, 1969, pp. 358–370.

[15] G. Hardy, E. Wright, An Introduction to the Number Theory, Oxford University Press, 1938.

[16] D. Knuth, The average time for carry propagation, Indagationes Mathematicae 40 (1978) 238–242.

[17] M. Lothaire, Combinatorics on Words, Vol. 17 of Encyclopedia of mathematics and its applications, Addison-Wesley, 1983.

[18] M. Lothaire, Applied Combinatorics on Words, (in preparation), available at http://www-igm.univ-mlv.fr/~berstel/Lothaire.

[19] R. Lyndon, On Burnside problem I, Trans. American Math. Soc. 77 (1954) 202–215.

[20] A. M. Odlyzko, Handbook of Combinatorics, Elsevier, 1995, Ch. Asymptotic enumeration methods.

[21] D. Panario, B. Richmond, Smallest components in decomposable structures: exp-log class, Algorithmica 29 (2001) 205–226.

[22] C. Reutenauer, Free Lie algebras, Oxford University Press, 1993.

[23] F. Ruskey, J. Sawada, Generating Lyndon brackets: a basis for the $n$-th homogeneous component of the free Lie algebra, Journal of Algorithms 46 (2003) 21–26.