The average lengths of the factors of the standard factorization of Lyndon words

Frédérique Bassino, Julien Clément and Cyril Nicaud,
Institut Gaspard Monge
Université de Marne-la-Vallée
77454 Marne-la-Vallée Cedex 2 - France
email: {bassino, clementj, nicaud}@univ-mlv.fr

No Institute Given

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 suffixes. 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 $3n/4$. Finally we present algorithms on Lyndon words derived from our work together with experimental results.

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 [Lyn54] 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 [Lot83], [Reu93], [RSar]).

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 $\mathbb{F}_k$ [Gol69], lot of results are known about this factorization: the average number of factors, the average length of the longest factor [FGP01] and of the shortest [PR01].

Several algorithms deal with Lyndon words. Duval gives in [Duv83] an algorithm that computes, in linear time, the factorization of a word into Lyndon words; he also presents in [Duv88] an algorithm for generating all Lyndon word up to a given length in lexicographical order. This algorithm runs in a constant average time (see [BP94]).

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 [BB97], 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 [FS02]) 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 which is based on the proof of Theorem 2 of Section 3. To the best of our knowledge these algorithms, although simple and not necessarily new, are not found elsewhere. Finally experiments are given which confirm our results and give hints of further studies.

The results contained in this paper constitute a first step in the study of the average behavior of the binary Lyndon trees obtained from Lyndon words by a recursive application of the standard factorization.

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 $\mathcal{L} = \cup_n \mathcal{L}_n$.

$$\begin{aligned}\mathcal{L} = \{&a, b, ab, aab, abb, aaab, aabb, abbb,\\ &aaaab, aaabb, aabab, aabbb, ababb, abbbb, \ldots\}\end{aligned}$$

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

$$w = l_1 l_2 \dots l_n, \quad l_i \in \mathcal{L}, \quad l_1 \geq l_2 \geq \cdots \geq l_n.$$

Moreover, l_n is the smallest suffix of w.

The number $\mathrm{Card}(\mathcal{L}_n)$ of Lyndon words of length n over A (see [Lot83]) is

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

where μ 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

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

Definition 1 (Standard factorization). *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.
Examples.

$$aaabaab = aaab \cdot aab, \qquad aaababb = a \cdot aababb, \qquad aabaabb = aab \cdot aabb.$$

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

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

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

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 1 *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_2a is not a left factor of x_1x_2b and $x_2a < x_1x_2a$. By construction of $\mathcal{X}_v$, as $x \neq v$, there exists a word w such that $v = x_1x_2aw$. We get that $x_2awx_1 < x_1x_2aw$. This is impossible since $v \in \mathcal{L}$.

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 [BP85].

Proposition 2 *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 $w = w_1w_2$ 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.

Proposition 3 *Let $l \in \mathcal{L}$ be a Lyndon word, $l \geq v$ if and only if $l \in \mathcal{X}_v^+$.*

Proof. If $l \geq v$, let l_1 be the longest prefix of l which belongs to $\mathcal{X}_v^*$, and l_2 such that $l = l_1l_2$. If $l_2 \neq \varepsilon$, we have the inequality $l_2l_1 > l \geq v$, thus $l_2l_1 > v$. The word v is not a prefix of l_2 since l_2 has no prefix in $\mathcal{X}_v$, hence we have $l_2 = l_2'bl_2''$ and $v = l_2'av''$. Then, by construction of $\mathcal{X}_v$, $l_2'b \in \mathcal{X}_v$ which is impossible. Thus $l_2 = \varepsilon$ and $l \in \mathcal{X}_v^+$.

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

Theorem 2 *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 1, wv can be written uniquely as

$$wv = l_1 l_2 \dots l_n, \quad l_i \in \mathcal{L}, \quad l_1 \geq l_2 \geq \cdots \geq l_n.$$

As v is the smallest (for the lexicographical order) suffix of awv, and consequently of wv, we get $l_n = v$; if $w = \varepsilon$, then $n = 1$, else $n \geq 2$ and for $1 \leq i \leq n-1$, $l_i \geq v$. 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

$$w = x_1 x_2 \dots x_n, \quad x_i \in \mathcal{X}_v \quad \text{and} \quad aw \notin \mathcal{X}_v^+.$$

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

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

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 3, 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

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

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.

Remark. The proof of Theorem 2 leads to a linear algorithm that computes the right factor of a Lyndon word using the fact that the factorization of Theorem 1 can be achieved in linear time and space (by an algorithm of Duval [Duv83], see Section 5).

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 [FS02]):

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

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

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

Theorem 3 *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 2, $F_v(z)$ 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 $l_1, \ldots, l_m$, we get, since $l_m \geq v$, that for all i, l_i is greater than v. Consequently from Proposition 3, for all i, $l_i \in \mathcal{X}_v$ 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 [BB97]. 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 of results from analytic combinatorics (see [FS02]), we establish the following result.

Theorem 4 *The average length for the uniform distribution over the Lyndon words of length n of the right factor of the standard factorization is asymptotically*

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

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

First we partition the set $\mathcal{L}_n$ of Lyndon words of length n in the two following subsets: $a\mathcal{L}_{n-1}$ and $\mathcal{L}'_n = \mathcal{L}_n \setminus a\mathcal{L}_{n-1}$.

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

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

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

Proposition 4 *The contribution of the set $\mathcal{L}'_n$ 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 $\mathcal{L}'_n$, the length of the right factor is asymptotically $n/2$.

The idea is to build a transformation φ, which is a bijection on a set $\mathcal{D}_n \subset \mathcal{L}'_n$, such that the sum of the lengths of standard right factors of w and $\varphi(w)$ is about $|w|$ the length of w. Indeed with such a relation we can compute the contribution of $\mathcal{D}_n$ to the expectation of parameter `right`. Then if the contribution of $\mathcal{L}'_n \setminus \mathcal{D}_n$ to the parameter `right` is negligible we are able to conclude for the expectation of parameter `right`.

It remains to exhibit/construct such a bijection φ and determine a "good" set $\mathcal{D}_n$. This is done in the following way: assume that w is a Lyndon word in $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$. *Let us denote by k the length of the first run of a's of the standard right* *factor.* We partition the set $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$ in two depending on the factorization. Indeed the standard factorization of w can only be one of the following

$$w = a^{k+1}b\,u \;\cdot\; a^k b\,v \text{ (first kind)}$$
$$w \;= a^k b\,u \;\cdot\; a^k b\,v \quad \text{(second kind)}.$$

This means that the left factor of a Lyndon word w can only begin by $a^{k+1}b$ or $a^k b$ when we know that the right factor begin by $a^k b$ (otherwise w cannot be in $\mathcal{L}_n \setminus a\mathcal{L}_{n-1}$). Let us fix a integer parameter $\lambda \in Z^+$. Then the words u, v of $\mathcal{X}_k^*$ can be uniquely written as $u = u'u''$ and $v = v'v''$ where u' and v' are the smallest prefixes of u and v of length greater than λ and ending by a b (there is always such a symbol b if these words are not empty since then u and v end with a b). When $|u|, |v| \geq \lambda$ we define $\varphi(w)$ for a word $w = a^k b\,u \;\cdot\; a^k b\,v$ (resp. $w = a^{k+1}b\,u \;\cdot\; a^k b\,v$) by

$$\varphi(w) = a^k b\,u'v''a^k b\,v'u'' \qquad \text{(resp. } a^{k+1}b\,u'v''a^k b\,v'u''\text{)}.$$

For example, considering $w = a^k b\,abb \;\cdot\; a^k b\,baab\;a^k b\,bbb$, if we choose $\lambda = 2$ and so $u' = ab, u'' = b, v' = baab, v'' = a^k b\,bbb$ then we get $\varphi(w) = a^k b\,aaba^k b\,bbb \cdot a^k b\,ab \in \mathcal{L}$. Here $|w| = |\varphi(w)| = 3k + 13$ and the length of the right standard factor are $2k + 9$ and $k + 3$ respectively.

Some words give hints of what we must be careful about if we want $\varphi(w)$ to be a Lyndon word.

- If we want the application φ to be well defined, the parameter λ must be greater or equal to 1. So the longest runs of a's have to be separated by non-empty words. If $w = a^k b \;\cdot\; a^k b\,b$, then $u = \varepsilon$ is the empty word. The application exchanging u and v gives a words which is no longer a Lyndon word.
- If $w = a^k b\,ab \;\cdot\; a^k b\,abb$, then $u = ab$ and it is a prefix of v. For any choice of λ, $\varphi(w)$ is not a Lyndon word. So the longest runs of a's have to be separated by words having distinct prefixes to ensure that $\varphi(w)$ is a Lyndon word.
- If $w = a^k b\,bab \;\cdot\; a^k b\,bbab$, then if we choose $\lambda = 1$, we get $\varphi(w) = a^k b\,bbaba^k b\,bab \notin \mathcal{L}$ (since $u' = v'$ and $u'v'' = bbab > v'u'' = bab$). Thus we have to take care, when we apply the transformation that $\varphi(w)$ is still smaller than its proper suffixes (this is ensured if $u' \neq v'$).

The application φ and set $\mathcal{D}_n$ are dependent and to suit our needs they are implicitly determined by the following constraints

1. The function φ is an involution on $\mathcal{D}_n$: $\varphi(\varphi(w)) = w$.
2. The standard factorization of $\varphi(w)$ for $w \in \mathcal{D}_n$ is

$$\varphi(w) = a^{k+1}b\,u'v'' \;\cdot\; a^k b\,v'u'' \text{ (first kind)}$$
$$\varphi(w) \;= a^k b\,u'u'' \;\cdot\; a^k b\,v'v'' \quad \text{(second kind)}.$$

3. The lengths of right factors of w and $\varphi(w)$ satisfy

$$|\texttt{right}(w)| + |\texttt{right}(\varphi(w))| = |w| \ \ (1 + o(|w|))\,.$$

4. The set $\mathcal{D}_n$ "captures" most of the set $\mathcal{L}'_n$ in an asymptotic way when n grows to ∞, that is

$$\frac{\mathrm{Card}(\mathcal{D}_n)}{\mathrm{Card}(\mathcal{L}'_n)} = 1 - o(1).$$

Most of these conditions are related to the properties of the longest runs of a's. Hence, in the following parts, we study some combinatorial properties of the longest runs of a's in Lyndon words to characterize φ and $\mathcal{D}_n$ precisely.

5 Algorithms and experimental results

In this section we give an algorithm to generate random Lyndon words of a given length n and use it to establish some experimental results about the length of the right factor in the standard factorization.

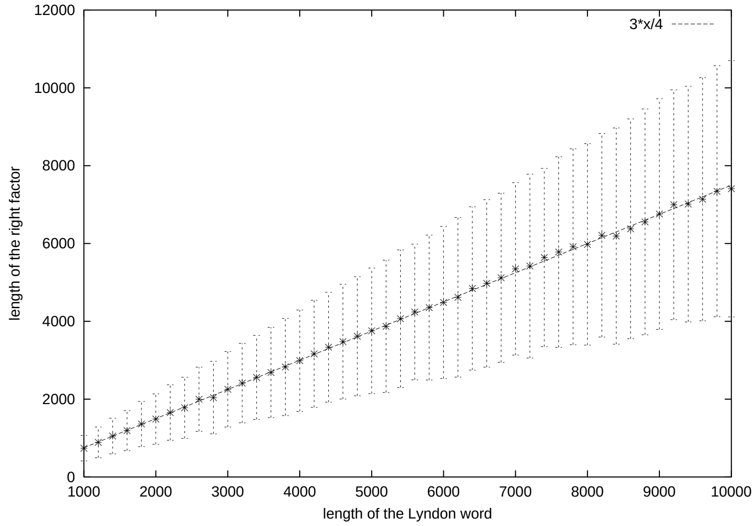

Fig. 1. Average length of the right factor of random Lyndon words with lengths from $1,000$ to $10,000$. Each plot is computed with $1,000$ words. The error bars represents the standard deviation.

Our algorithms use Duval's algorithm [Duv83], which computes in linear time the decomposition of a word into decreasing Lyndon words (see Theorem 1). So we assume that we have a algorithm named `Duval(u)` which produce the Lyndon words $l_1 \geq l_2 \geq \cdots \geq l_k$ such that

$$u = l_1 l_2 \dots l_k.$$

Let the function `Duval(string u, int k, array pos)` be the function which computes the Lyndon decomposition of `u` by storing in an array `pos` of size `k` the positions of the factors.

There exists an algorithm `SmallestConjugate(u)`, proposed by Booth [Lot03,?], that computes 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 2 *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 power 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 3 *Let $l = au$ be a Lyndon word of length greater or equals to 2 starting with a letter a. Let $l_1 \dots l_k$ be the Lyndon factorization of u. The right factor of l in its standard factorization is l_k.*

Proof. By Theorem 1, l_k is the smallest suffix of u, thus it is the smallest proper suffix of l.

The algorithm to compute the right factor of a Lyndon word l such that $|l| \geq 2$ is the following:

```
RightFactor(string l[1..n])
 array pos;
 int k;
 pos = Duval(l[2..n], k, pos); // omit the first letter a 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 and 2 present some experimental results obtained with our algorithms.

Open problem The results obtained in this paper are only the first step toward the average case-analysis of the Lyndon tree. The Lyndon tree $T(w)$ of a Lyndon word w is recursively built in the following way

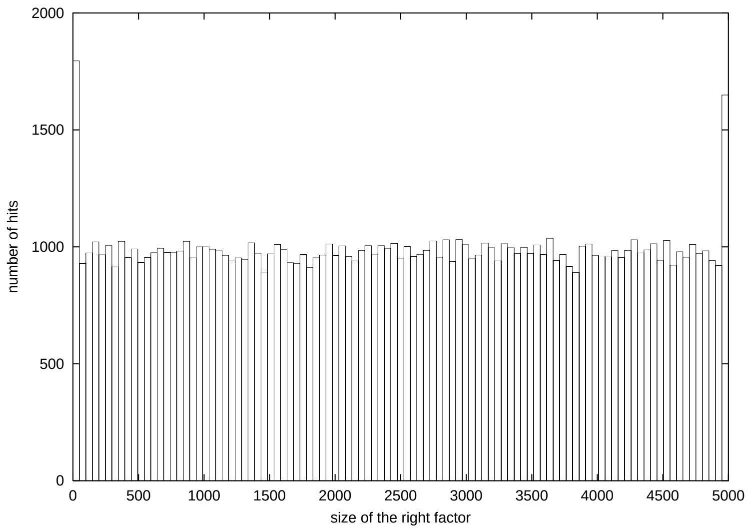

Fig. 2. Distribution of the length of the right factor. We generated $100,000$ random Lyndon words of length $5,000$.

- if w is a letter, then $T(w)$ is a external node labeled by the letter.
- otherwise, $T(w)$ is an internal node having $T(u)$ and $T(v)$ as children where the standard factorization of w is $u \cdot v$.

This structure encodes a nonassociative operation, either a commutator in the free group [CFL58], or a Lie bracketing [Lot83]; both constructions leads to bases of the free Lie algebra.

In order to study the height of the tree obtained from a Lyndon word by successive standard factorizations, it would be very interesting to get more precise informations about the distribution of the right factors of words of $\mathcal{L}'_n$. Fig. 2 hints a very strong equi-repartition property of the length of the right factor over this set. This suggests a very particular subdivision process at each node of the factorization tree which needs further investigations.

References

[BB97] J. Berstel and L. Boasson. The set of lyndon words is not context-free. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, 63:139–140, 1997.

[Boo80] K. S. Booth. Lexicographically least circular substrings. *Inform. Process. Lett.*, 10(4-5):240–242, 1980.

[BP85] J. Berstel and D. Perrin. *Theory of codes*. Academic Press, 1985.

[BP94] J. Berstel and M. Pocchiola. Average cost of Duval's algorithm for generating Lyndon words. *Theoret. Comput. Sci.*, 132(1-2):415–425, 1994.

[Car85] H. Cartan. *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes*. Hermann, 1985.

[CFL58] K.T. Chen, R.H. Fox, and R.C. Lyndon. Free differential calculus IV: The quotient groups of the lower central series. *Ann. Math.*, 58:81–95, 1958.

[Duv83] J.-P. Duval. Factorizing words over an ordered alphabet. *Journal of Algorithms*, 4:363–381, 1983.

[Duv88] J.-P. Duval. Génération d'une section des classes de conjugaison et arbre des mots de Lyndon de longueur bornée. *Theoret. Comput. Sci.*, 4:363–381, 1988.

[FGP01] P. Flajolet, X. Gourdon, and D. Panario. The complete analysis of a polynomial factorization algorithm over finite fields. *Journal of Algorithms*, 40:37–81, 2001.

[FS91] P. Flajolet and M. Soria. The cycle construction. *SIAM J. Disc. Math.*, 4:58–60, 1991.

[FS02] P. Flajolet and R. Sedgewick. Analytic combinatorics–symbolic combinatorics. Book in preparation, 2002. (Individual chapters are available as INRIA Research reports at http://www.algo.inria.fr/flajolet/publist.html).

[Gol69] S. Golomb. Irreducible polynomials, synchronizing codes, primitive necklaces and cyclotomic algebra. In *Proc. Conf Combinatorial Math. and Its Appl.*, pages 358–370, Chapel Hill, 1969. Univ. of North Carolina Press.

[HW38] G. H. Hardy and E. M. Wright. *An Introduction to the Number Theory.* Oxford University Press, 1938.

[Knu78] D. E. Knuth. The average time for carry propagation. *Indagationes Mathematicae*, 40:238–242, 1978.

[Lot83] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of mathematics and its applications.* Addison-Wesley, 1983.

[Lot03] M. Lothaire. *Applied Combinatorics on Words.* 2003. in preparation, chapters available at http://www-igm.univ-mlv.fr/~berstel/Lothaire.

[Lyn54] R. C. Lyndon. On Burnside problem I. *Trans. American Math. Soc.*, 77:202–215, 1954.

[PR01] D. Panario and B. Richmond. Smallest components in decomposable structures: exp-log class. *Algorithmica*, 29:205–226, 2001.

[Reu93] C. Reutenauer. *Free Lie algebras.* Oxford University Press, 1993.

[RSar] F. Ruskey and J. Sawada. Generating Lyndon brackets: a basis for the n-th homogeneous component of the free Lie algebra. *Journal of Algorithms*, (to appear). Available at `http://www.cs.uvic.ca/ fruskey/Publications/`.