Nyldon words

Émilie Charlier
Department of Mathematics
University of Liège
Allée de la Découverte 12
4000 Liège, Belgium
echarlier@uliege.be

Manon Philibert
Laboratoire d'Informatique et Systèmes
Aix-Marseille University
52 Av. Escadrille Normandie Niemen
13397 Marseille, France
manon.philibert@lis-lab.fr

Manon Stipulanti*
Department of Mathematics
University of Liège
Allée de la Découverte 12
4000 Liège, Belgium
m.stipulanti@uliege.be

January 28, 2019

Abstract

The Chen-Fox-Lyndon theorem states that every finite word over a fixed alphabet can be uniquely factorized as a lexicographically nonincreasing sequence of Lyndon words. This theorem can be used to define the family of Lyndon words in a recursive way. If the lexicographic order is reversed in this definition, we obtain a new family of words, which are called the Nyldon words. In this paper, we show that every finite word can be uniquely factorized into a lexicographically nondecreasing sequence of Nyldon words. Otherwise stated, Nyldon words form a complete factorization of the free monoid with respect to the decreasing lexicographic order. Then we investigate this new family of words. In particular, we show that Nyldon words form a right Lazard set.

2010 *Mathematics Subject Classification*: 68R15, 94A45.
Keywords: Lyndon words, Nyldon words, complete factorization of the free monoid, Lazard factorization, Hall set, comma-free code

---

1 Introduction

The Chen-Fox-Lyndon theorem states that every finite word w can be uniquely factorized as $w = \ell_1 \ell_2 \cdots \ell_k$ where $\ell_1, \ldots, \ell_k$ are Lyndon words such that $\ell_1 \geq_{\text{lex}} \cdots \geq_{\text{lex}} \ell_k$. This theorem can be used to define the family of Lyndon words over some totally ordered

*Corresponding author.

alphabet in a recursive way: the letters are Lyndon; a finite word of length greater than one is Lyndon if and only if it cannot be factorized into a nonincreasing sequence of shorter Lyndon words. In a Mathoverflow post dating from November 2014, Grinberg defines a variant of Lyndon words, which he calls Nyldon words, by reversing the lexicographic order in the previous recursive definition: the letters are Nyldon; a finite word of length greater than one is Nyldon if and only if it cannot be factorized into a *nondecreasing* sequence of shorter Nyldon words [Gri14]. The class of words so obtained is not, as one might first think, the class of maximal words in their conjugacy classes. Grinberg asks three questions:

1. How many Nyldon words of each length are there?
2. Is there an equivalent to the Chen-Fox-Lyndon theorem for Nyldon words? More precisely, is it true that for every finite word w, there exists a unique sequence $(n_1, \dots, n_k)$ of Nyldon words satisfying $w = n_1 \cdots n_k$ and $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$?
3. Is it true that every primitive word admits exactly one Nyldon word in its conjugacy class, whereas any non primitive word has no such conjugate?

In this paper, we discuss the properties of this new class of words in the more general context of the complete factorizations of the free monoid as introduced by Schützenberger in [Sch65] and later on extensively studied in [BPR10, Lot97, Mél92, Reu93, Vie78]. In particular, we show that each of Grinberg's questions has a positive answer.

Another variant of Lyndon words has been recently studied in [BDFZZ18]: the *inverse Lyndon words*. Similarly to the Nyldon words, which are studied in the present paper, the inverse Lyndon words are built from the decreasing lexicographic order: they are the words that are greater than any of their proper suffixes. Although it is true that Nyldon words are greater than all their Nyldon proper suffixes (this is Theorem 13), the families of Nyldon words and inverse Lyndon words do not coincide. For example, the family of inverse Lyndon words is prefix-closed whereas the family of Nyldon words is not. Another major difference between those two families of words is that the inverse Lyndon factorization of a word is not unique in general, while the Nyldon factorization is always unique (this is Theorem 16).

This paper has the following organization. Section 2 contains the necessary background and definitions. In Section 3, we briefly discuss the prefixes of Nyldon words. Then, in Section 4, we show that the unicity of the Nyldon factorization implies that there are equally many Nyldon and Lyndon words of each length. In Section 5, we focus on suffixes of Nyldon words. In particular, we show that Nyldon words are greater than all their Nyldon proper suffixes. This allows us to prove in Section 6 that the Nyldon factorization of a word is unique. Then, in Section 7, we provide algorithms for computing the Nyldon factorization of a word and for generating the Nyldon words up to any given length. Next, in Section 8, we show how to obtain a standard factorization of the Nyldon words, similarly to the standard factorization of the Lyndon words. In Section 9, we recall Schützenberger's theorem on factorizations of the free monoid and we show that this result combined with the unicity of the Nyldon factorization implies the primitivity of Nyldon words, as well as the fact that each primitive conjugacy class contains exactly one Nyldon word. Then, in Section 10, we realize an in-depth comparison between the families of Nyldon and Lyndon words. In particular, we emphasize how the exclusive knowledge of the recursive definition of the Lyndon words allows us to recover some known properties of Lyndon words, in a similar way to what we do for Nyldon words. However, we note that the proofs in the

Lyndon case are not simple translations of the proofs from the Nyldon case. We also show that the standard factorization of Nyldon words introduced in Section 8 can be understood as the analogue of the right standard factorization of the Lyndon words. In Section 11, we show that the Nyldon words form a right Lazard factorization, but not a left Lazard factorization. Finally, in Section 12, we apply Melançon's algorithm in order to compute the Nyldon conjugate of any primitive word. Along the way, we mention seven open problems.

2 Preliminaries

Throughout the text, A designates an arbitrary finite alphabet of cardinality at least 2 and endowed with a total order $<$. Furthermore, when we need to specify the letters in A, we use the notation $A = \{\mathtt{0}, \mathtt{1}, \dots, \mathtt{m}\}$, with $\mathtt{0} < \mathtt{1} < \cdots < \mathtt{m}$. We use the usual definitions (prefix, suffix, factor, etc.) and notation of combinatorics on words; for example, see [Lot97]. In particular, A^* is the set of all finite words over the alphabet A and A^+ is the set of all nonempty finite words over the alphabet A. The empty word is denoted by ε and the length of a finite word w is denoted by $|w|$. A prefix (resp., suffix, factor) u of a word w is *proper* if $u \neq w$. Moreover, we say that a sequence $(w_1, \dots, w_k)$ of nonempty words over A is a *factorization* of a word w if $w = w_1 \cdots w_k$. We also say that k is the *length* of this factorization. Note that, with the previous notation, we have $w = \varepsilon$ if and only if $k = 0$. In particular, if w is a nonempty word then $k \geq 1$. Recall that the *lexicographic order on* A^*, which we denote by $<_{\text{lex}}$, is a total order on A^* induced by the order $<$ on the letters and defined as follows: $u <_{\text{lex}} v$ either if u is a proper prefix of v or if there exist $\mathtt{i}, \mathtt{j} \in A$ and $p \in A^*$ such that $\mathtt{i} < \mathtt{j}$, $p\mathtt{i}$ is a prefix of u and $p\mathtt{j}$ is a prefix of v. As usual, we write $u \leq_{\text{lex}} v$ if $u <_{\text{lex}} v$ or $u = v$. Also recall that two words x and y are *conjugate* if they are circular permutations of each other: $x = uv$ and $y = vu$ for some words u, v. A word x is a *power* if $x = u^k$ for some word u and some integer $k \geq 2$. In particular, the empty word ε is considered to be a power. Finally, a word which is not a power is said to be *primitive*.

Definition 1. A finite word w over A is *Lyndon* if it is primitive and lexicographically minimal among its conjugates. We let $\mathcal{L}$ denote the set of all Lyndon words over A.

It is easily verified that a word w is Lyndon if and only if $w \neq \varepsilon$ and, for all $u, v \in A^+$ such that $w = uv$, we have $w <_{\text{lex}} vu$.

The following theorem is usually referred to as the Chen-Fox-Lyndon factorization theorem [CFL58], although these authors never formulated the result in this way. It was simultaneously obtained by Širšov [Šir58]. Also see [BCL13, BP07] for a discussion on the origins of this theorem.

Theorem 2 ([CFL58, Šir58])**.** *For every finite word w over A, there exists a unique factorization $(\ell_1, \dots, \ell_k)$ of w into Lyndon words over A such that $\ell_1 \geq_{\text{lex}} \cdots \geq_{\text{lex}} \ell_k$.*

This result allows us to (re)define the set of Lyndon words recursively.

Corollary 3. *Let L be the set of words over A recursively defined as follows: letters are in L; a word of length at least two belongs to L if and only if it cannot be factorized into a lexicographically nonincreasing sequence of shorter words of L. Then $L = \mathcal{L}$.*

Proof. An easy induction on n shows that $L \cap A^n = \mathcal{L} \cap A^n$ for all $n \geq 1$. ◻

Now we consider the following class of words, which were baptized the *Nyldon words* by Grinberg [Gri14].

Definition 4. Let $\mathcal{N}$ be the set of words over A recursively defined as follows: letters are in $\mathcal{N}$; a word of length at least two belongs to $\mathcal{N}$ if and only if it cannot be factorized into a (lexicographically) nondecreasing sequence of shorter words of $\mathcal{N}$. Otherwise stated, a word w is in $\mathcal{N}$ either if it is a letter, or if there does not exist any factorization $(n_1, \ldots, n_k)$ of w into words in $\mathcal{N}$ such that $\max\{|n_1|, \ldots, |n_k|\} < |w|$ and $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$. The words of $\mathcal{N}$ are called the *Nyldon words*. Moreover, any factorization $(n_1, \ldots, n_k)$ of w into Nyldon words such that $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$ is called a *Nyldon factorization* of w.

With this new vocabulary, we can say that a word w is Nyldon if and only if its only Nyldon factorization is (w). Otherwise stated, $k = 1$ and $n_1 = w$. On the opposite, a nonempty word is not Nyldon if and only if it admits a Nyldon factorization of length at least 2. The fact that every finite word admits at least one Nyldon factorization is immediate from the definition of Nyldon words, but still worth being emphasized.

Proposition 5. *For every finite word w over A, there exists a factorization $(n_1, \ldots, n_k)$ of w into Nyldon words over A such that $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$.*

We illustrate the definition of Nyldon words in the case where $A = \{\mathtt{0}, \mathtt{1}\}$. In what follows, we will refer to the Nyldon words over this alphabet as the *binary Nyldon words*.

Example 6. By Definition 4, the letters `0` and `1` are Nyldon. The word `00` (resp., `01`, `11`) of length 2 admits the Nyldon factorization $(\mathtt{0}, \mathtt{0})$ (resp., $(\mathtt{0}, \mathtt{1})$, $(\mathtt{1}, \mathtt{1})$), and hence is not a Nyldon word. Consequently, the only binary Nyldon word of length 2 is `10`. The binary Lyndon and Nyldon words up to length 7 are stored in Table 1.

Lyndon	Nyldon	Lyndon	Nyldon	Lyndon	Nyldon
0	0	000001	100000	0001011	1000110
1	1	000011	100001	0001101	1000111
01	10	000101	100010	0001111	1001010
001	100	000111	100011	0010011	1001100
011	101	001011	100110	0010101	1001110
0001	1000	001101	100111	0010111	1001111
0011	1001	001111	101100	0011011	1011000
0111	1011	010111	101110	0011101	1011001
00001	10000	011111	101111	0011111	1011010
00011	10001	0000001	1000000	0101011	1011100
00101	10010	0000011	1000001	0101111	1011101
00111	10011	0000101	1000010	0110111	1011110
01011	10110	0000111	1000011	0111111	1011111
01111	10111	0001001	1000100		

Table 1: The binary Lyndon and Nyldon words up to length 7.

Note that Nyldon words are not lexicographically extremal among their conjugates since, for instance, `101` is a binary Nyldon.

3 Prefixes of Nyldon words

Apart from the letters, the binary Nyldon words of Table 1 all start with the prefix 10. This fact is true for longer lengths as well as all alphabet sizes, as shown by the following proposition. However, note that both 100 and 101 are prefixes of Nyldon words (see Table 1).

Proposition 7. *Each Nyldon word over* $\{\mathtt{0},\mathtt{1},\dots,\mathtt{m}\}$ *of length at least* 2 *starts with* $\mathtt{ij}$ *with* $\mathtt{0} \le \mathtt{j} < \mathtt{i} \le \mathtt{m}$.

Proof. Let $w = \mathtt{ij}u$, with $u \in \{\mathtt{0},\mathtt{1},\dots,\mathtt{m}\}^*$ and $\mathtt{0} \le \mathtt{i} \le \mathtt{j} \le \mathtt{m}$. Our aim is to show that w cannot be Nyldon. Let $(n_1,\dots,n_k)$ be a Nyldon factorization of $\mathtt{j}u$. Then n_1 begins with $\mathtt{j}$. Since $\mathtt{i}$ is Nyldon, $\mathtt{i} \le_{\text{lex}} \mathtt{j} \le_{\text{lex}} n_1$ and $k \ge 1$, we obtain that $(\mathtt{i}, n_1,\dots,n_k)$ is a Nyldon factorization of w of length at least 2, whence w is not Nyldon. □

For instance, all binary Nyldon of length at least 2 start with the prefix 10, and all ternary Nyldon words of length at least 2 start with 10, 20 or 21. Other prefixes than those of Propostion 7 are forbidden in the family of Nyldon words. We introduce the following definition.

Definition 8. We say that a finite word p over A is a *forbidden prefix* if no Nyldon word over A starts with p.

In the following proposition, we exhibit a family of forbidden prefixes by generalizing the argument of Propostion 7.

Proposition 9. *All elements of the set*

$$F = \{p_1p_2p_3 \in A^* \colon p_1 \in \mathcal{N},\ p_1 \le_{\text{lex}} p_2 \ \text{ and, for all } u \in A^* \text{ and for all Nyldon factorizations } (n_1,\dots,n_k) \text{ of } p_2p_3u,\ \text{ one has } |n_1| \ge |p_2|\}$$

are forbidden prefixes.

Proof. Let $p \in F$. Then it can be decomposed as $p = p_1p_2p_3$ where p_1, p_2, p_3 satisfy the properties described in the statement (see Figure 1). We have to show that all words

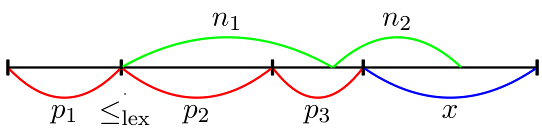

Figure 1: A family of forbidden prefixes.

over A starting with p are not Nyldon. Let $w = pu$, with $u \in A^*$. Let $(n_1,\dots,n_k)$ be a Nyldon factorization of p_2p_3u. By definition of F, $p_1 \le_{\text{lex}} p_2$ and p_2 must be a prefix of n_1, hence $p_1 \le_{\text{lex}} n_1$. Since we also have $p_1 \in \mathcal{N}$, we obtain that $(p_1, n_1,\dots,n_k)$ is a Nyldon factorization of w of length at least 2, whence w is not Nyldon. □

Example 10. We obtain from Proposition 9 that, for all $k \in \mathbb{N}$, the words $\mathtt{10}^k\mathtt{10}^k$, $\mathtt{10}^k\mathtt{1011}$, $\mathtt{101}^{k+1}\mathtt{01}^{k+1}$, $\mathtt{10}^{k+2}\mathtt{110}^{k+1}\mathtt{11}$ are forbidden prefixes. The corresponding factors p_1, p_2, p_3 as in the definition of the set F are given in Table 2.

w	p_1	p_2	p_3
10^k10^k	10^k	10^k	ε
10^k1011	10^k	101	1
$101^{k+1}01^{k+1}$	101^k	101^k	1
$10^{k+2}110^{k+1}11$	$10^{k+2}1$	$10^{k+1}1$	1

Table 2: Some forbidden prefixes.

All the examples of forbidden prefixes that we have come from the family F of Proposition 9. In general, it seems hard to understand the prefixes of Nyldon words, while we are able to understand their suffixes quite well (see Section 5). We leave the following unresolved question as an open problem.

Open Problem 11. Characterize the language of forbidden prefixes. In particular, prove or disprove that there are other forbidden prefixes than those given by the family F of Proposition 9.

As is well known, the nonempty prefixes of Lyndon words are exactly the sesquipowers of Lyndon words distinct of the maximal letter [Duv83, Knu11]. A *sesquipower* of a word x is a word x^kp where k is a positive integer and p is a proper prefix of x. It is not true that all sesquipowers of Nyldon words distinct of the letters can be prefixes of Nyldon words since for example 1010 is a forbidden prefix. Moreover, we know that each prefix of a Lyndon word is a sesquipower of exactly one Lyndon word [Duv83, Knu11]. This property does not hold for Nyldon words either since for example, the word 101101 is the prefix of the Nyldon word 1011011 and is a sesquipower of both the Nyldon words 101 and 10110. However, we do not know whether all prefixes of Nyldon words are sesquipowers of Nyldon words.

Open Problem 12. Prove or disprove that prefixes of Nyldon words must be sesquipowers of Nyldon words.

4 Counting Nyldon words of length n

As already noticed by Grinberg, if we can prove the unicity of the Nyldon factorization (which is actually proved in Section 6), then we know that there are equally many Lyndon and Nyldon words of each length. Indeed, suppose that every finite word admits a unique Nyldon factorization. We proceed by induction on the length n of the words to show that $\#(\mathcal{N} \cap A^n) = \#(\mathcal{L} \cap A^n)$ for all $n \geq 1$. Since all the letters are both Lyndon and Nyldon, the base case is verified. Now suppose that $n \geq 2$ and that, for all $m < n$, there is the same number of Lyndon and Nyldon words of length m. Since we have assumed that the Nyldon factorization of each word is unique, the number of words of length n that are not Nyldon is equal to the number of Nyldon factorizations $(n_1, \ldots, n_k)$ such that $|n_1|, \ldots, |n_k| < n$. Similarly, we know by Theorem 2 that the Lyndon factorization of each word is unique, hence the number of words of length n that are not Lyndon is equal to the number of Lyndon factorizations $(\ell_1, \ldots, \ell_k)$ such that $|\ell_1|, \ldots, |\ell_k| < n$. Then, using the induction hypothesis, we obtain that $\#(\mathcal{N}^c \cap A^n) = \#(\mathcal{L}^c \cap A^n)$, where X^c denotes the complement of the set X. This, of course, implies that $\#(\mathcal{N} \cap A^n) = \#(\mathcal{L} \cap A^n)$ as well.

We illustrate this construction for the binary words of length 4. Table 3 is built as follows. The left column contains the non-Lyndon words of length 4, ordered in decreasing lexicographic order. If the non-Lyndon word of a certain row admits the Lyndon factorization $(\ell_1, \dots, \ell_k)$, then the corresponding non-Nyldon word in the same row has the Nyldon factorization $(n_1, \dots, n_k)$, where $\{|\ell_1|, \dots, |\ell_k|\} = \{|n_1|, \dots, |n_k|\}$ and for each occurrence of a factor ℓ in the Lyndon factorization $(\ell_1, \dots, \ell_k)$, if ℓ the jth word of length $|\ell|$ in the list of Lyndon words of length up to 3 in decreasing lexicographic order, then, in the corresponding Nyldon factorization $(n_1, \dots, n_k)$, there is an occurrence of the jth word of the same length $|\ell|$ in the list of Nyldon words of length up to 3 in increasing lexicographic order.

Lyndon			Nyldon		
$\mathtt{1} >_{\text{lex}} \mathtt{011} >_{\text{lex}} \mathtt{01} >_{\text{lex}} \mathtt{001} >_{\text{lex}} \mathtt{0}$			$\mathtt{0} <_{\text{lex}} \mathtt{1} <_{\text{lex}} \mathtt{10} <_{\text{lex}} \mathtt{100} <_{\text{lex}} \mathtt{101}$		
1111	(1, 1, 1, 1)	$1+1+1+1$	0000	(0, 0, 0, 0)	$1+1+1+1$
1110	(1, 1, 1, 0)	$1+1+1+1$	0001	(0, 0, 0, 1)	$1+1+1+1$
1101	(1, 1, 01)	$1+1+2$	0010	(0, 0, 10)	$1+1+2$
1100	(1, 1, 0, 0)	$1+1+1+1$	0011	(0, 0, 1, 1)	$1+1+1+1$
1011	(1, 011)	$1+3$	0100	(0, 100)	$1+3$
1010	(1, 01, 0)	$1+2+1$	0110	(0, 1, 10)	$1+1+2$
1001	(1, 001)	$1+3$	0101	(1, 101)	$1+3$
1000	(1, 0, 0, 0)	$1+1+1+1$	0111	(0, 1, 1, 1)	$1+1+1+1$
0110	(011, 0)	$3+1$	1100	(1, 100)	$1+3$
0101	(01, 01)	$2+2$	1010	(10, 10)	$2+2$
0100	(01, 0, 0)	$2+1+1$	1110	(1, 1, 10)	$1+1+2$
0010	(001, 0)	$3+1$	1101	(1, 101)	$1+3$
0000	(0, 0, 0, 0)	$1+1+1+1$	1111	(1, 1, 1, 1)	$1+1+1+1$

Table 3: There are equally many non-Nyldon words and non-Lyndon words of length 4.

The sequence

$$2, 1, 2, 3, 6, 9, 18, 30, 56, 99, 186, 335, 630, 1161, \dots$$

that counts the number of binary Lyndon words of length $n \geq 1$ is referred to as A001037 in Sloane's On-Line Encyclopedia of Integer Sequences.

5 Suffixes of Nyldon words

Unlike their prefixes, the suffixes of Nyldon words satisfy nice properties. In this section, we prove two results that will allow us to obtain the unicity of the Nyldon factorization.

Theorem 13. *Let $w \in A^*$ be a Nyldon word. For each Nyldon proper suffix s of w, we have $s <_{\text{lex}} w$.*

Proof. We proceed by induction on $|w|$. The case $|w| = 1$ is obvious. If $|w| = 2$, write $w = \mathtt{ij}$ with $\mathtt{i}, \mathtt{j} \in A$. The only Nyldon proper suffix of w is its last letter $\mathtt{j}$ and $\mathtt{j} <_{\text{lex}} \mathtt{i} <_{\text{lex}} w$ by Proposition 7.

Now, we suppose that $|w| \geq 3$ and that the result is true for all Nyldon words shorter than w. Proceed by contradiction and assume that there exists a Nyldon proper suffix s of w such that

$$s \geq_{\text{lex}} w. \tag{1}$$

Among all such suffixes of w, we choose s to be the longest and we write $w = ps$ with $p \in A^*$. Our goal is to reach the contradiction that this well-chosen suffix s is not Nyldon by exhibiting a Nyldon factorization of s of length at least 2.

First, we show that $p \notin \mathcal{N}$. Indeed, if $p \in \mathcal{N}$ and $p \leq_{\text{lex}} s$, then (p, s) is a Nyldon factorization of w of length 2, contradicting that w is Nyldon. Thus, if p were Nyldon, then $s <_{\text{lex}} p <_{\text{lex}} w$, which would contradict the assumption (1) on s. Hence $p \notin \mathcal{N}$.

Let $(p_1, \ldots, p_k)$ be a Nyldon factorization of p of length $k \geq 2$. Since $w, s \in \mathcal{N}$, we must have

$$p_k >_{\text{lex}} s, \tag{2}$$

for otherwise $(p_1, \ldots, p_{k-1}, p_k, s)$ would be a Nyldon factorization of w of length $k+1 \geq 2$. Moreover, $p_k s \notin \mathcal{N}$ for otherwise $(p_1, \ldots, p_{k-1}, p_k s)$ would be a Nyldon factorization of w of length $k \geq 2$.

Now let $(n_1, \ldots, n_\ell)$ be a Nyldon factorization of $p_k s$ of length $\ell \geq 2$. We claim that there exist $i \in \{1, \ldots, \ell\}$ and $x, y \in A^+$ such that

$$n_i = xy, \quad p_k = n_1 \cdots n_{i-1} x \quad \text{and} \quad s = y n_{i+1} \cdots n_\ell.$$

Indeed, suppose instead that $p_k = n_1 \cdots n_j$ and $s = n_{j+1} \cdots n_\ell$ with $1 \leq j \leq \ell - 1$ (recall that p_k and s are not empty). But since both p_k and s are Nyldon, this implies $p_k = n_1$ and $s = n_2$ (and hence, $\ell = 2$). But $n_1 \leq_{\text{lex}} n_2$, which contradicts (2).

Next, we must have $i \leq \ell - 1$. Indeed, suppose that $i = \ell$. Then $y = s$ and $n_\ell = xs$. By induction hypothesis, we would get that $s <_{\text{lex}} n_\ell$. Then, by using (1), we obtain that n_ℓ is a Nyldon proper suffix of w longer than s and such that $n_\ell >_{\text{lex}} s \geq_{\text{lex}} w$. This is impossible by maximality of the length of s.

Now let $(y_1, \ldots, y_t)$ be a Nyldon factorization of y (with $t \geq 1$). Then y_t is a Nyldon proper suffix of n_i, which is itself Nyldon. Since $|n_i| < |w|$, the induction hypothesis yields $y_t <_{\text{lex}} n_i$. But then, since $n_i \leq_{\text{lex}} n_{i+1}$, we obtain that $(y_1, \ldots, y_t, n_{i+1}, \ldots, n_\ell)$ is a Nyldon factorization of s of length at least 2, contradicting that s is Nyldon as announced. ☐

Theorem 14. *Let $w \in A^+$ and let $(n_1, \ldots, n_k)$ be a Nyldon factorization of w. Then n_k is the longest Nyldon suffix of w.*

Proof. Let s denote the longest Nyldon suffix of w. If w is Nyldon, then $k = 1$ and $s = w = n_k$. Suppose now that w is not Nyldon. Then $k \geq 2$ and s is a proper suffix of w. Write $w = ps$. By choice of s and since n_k is Nyldon, we have $|n_k| \leq |s|$. Let us show that $|n_k| = |s|$. Proceed by contradiction and suppose that $|n_k| < |s|$. Since s is Nyldon, we cannot have $s = n_i \cdots n_k$ with $i < k$. Therefore, there must exist $i \in \{1, \ldots, k-1\}$ and $x, y \in A^+$ such that $n_i = xy$, $p = n_1 \cdots n_{i-1} x$ and $s = y n_{i+1} \cdots n_k$. Let $(y_1, \ldots, y_t)$ be a Nyldon factorization of y (with $t \geq 1$). From Theorem 13, we deduce that $y_t <_{\text{lex}} n_i$ since y_t is a proper suffix of n_i. But then, since $n_i \leq_{\text{lex}} n_{i+1}$, we obtain that $(y_1, \ldots, y_t, n_{i+1}, \ldots, n_k)$ is a Nyldon factorization of s of length at least 2, contradicting that s is Nyldon. Consequently, $|n_k| = |s|$, which in turn implies $n_k = s$. ☐

However, it is not true that the longest Nyldon prefix of w is the first factor of any Nyldon factorization, as illustrated below.

Example 15. Let $w = \texttt{10100}$. Then $(\texttt{10}, \texttt{100})$ is a Nyldon factorization of w although its longest Nyldon prefix is $\texttt{101}$.

6 Unicity of the Nyldon factorization

Using Theorem 14, we obtain that there can be only one Nyldon factorization of each word, which positively answers the second question of Grinberg mentioned in the introduction (and hence the first as well in view of Section 4). In particular, the Nyldon words form a complete factorization of the free monoid A^*, which we make explicit below.

Theorem 16. *For every finite word w over A, there exists a unique sequence $(n_1, \ldots, n_k)$ of Nyldon words such that $w = n_1 \cdots n_k$ and $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$.*

Proof. The existence of the Nyldon factorization is known from Proposition 5. Let us show that the unicity follows from Theorem 14. We proceed by induction on the length of the words. If $|w| \leq 1$, then the result is clear. Now suppose that $|w| \geq 2$ and that every finite word shorter than w admits a unique Nyldon factorization. If w is Nyldon, its only possible factorization is (w). Suppose now that w is not Nyldon and let $(n_1, \ldots, n_k)$ be a Nyldon factorization of w. Then $k \geq 2$ and n_k is the longest Nyldon suffix of w by Theorem 14. Since $(n_1, \ldots, n_{k-1})$ is a Nyldon factorization of a word shorter than w, the factors $n_1, \ldots, n_{k-1}$ are also completely determined by w by induction hypothesis. Therefore, there cannot be another Nyldon factorization of w than $(n_1, \ldots, n_k)$. □

From now on, thanks to Theorem 16, we will talk about *the* Nyldon factorization of a finite word w instead of *a* Nyldon factorization (see Definition 4) to refer to the unique sequence $(n_1, \ldots, n_k)$ of Nyldon words such that $w = n_1 \cdots n_k$ and $n_1 \leq_{\text{lex}} \cdots \leq_{\text{lex}} n_k$.

As mentioned in Section 4, Theorem 16 implies that there are equally many Lyndon and Nyldon words of each length. This result, however, does not provide us with a natural bijection between Lyndon and Nyldon words of the same length. Such a bijection will be given by Theorem 26 below in Section 9. Although Table 3 describes a length-preserving bijection between non-Lyndon and non-Nyldon words, it is not fully satisfying, in the sense that we need to precompute all Lyndon and Nyldon words up to length $n-1$ in order to compute the image of a non-Lyndon word of length n under this bijection.

7 A faster algorithm for computing the Nyldon factorization

The recursive definition of the Nyldon words comes with a natural algorithm for generating the Nyldon words up to any given length. Unfortunately, the complexity of this first algorithm is clearly prohibitive. In this section, we show how to rapidly compute the Nyldon factorization of a word, and in turn, rapidly generate the Nyldon words up to any given length.

Proposition 17. *Let $w \in A^+$, $w = ps$ where s is the longest Nyldon proper suffix of w and let $(p_1, \ldots, p_k)$ be the Nyldon factorization of p. Then w is Nyldon if and only if $p_k >_{\text{lex}} s$.*

Proof. Let us show, equivalently, that w is not Nyldon if and only if $p_k \leq_{\text{lex}} s$. If $p_k \leq_{\text{lex}} s$, then w is not Nyldon since its Nyldon factorization is $(p_1, \ldots, p_k, s)$, which is of length at least 2. In order to prove the other direction, we now suppose that w is not Nyldon. Let $(w_1, \ldots, w_\ell)$ be the Nyldon factorization of w. Then $\ell \geq 2$. By Theorem 14, we obtain that $w_\ell = s$. But then $(p_1, \ldots, p_k)$ and $(w_1, \ldots, w_{\ell-1})$ are both Nyldon factorizations of p. By unicity of the Nyldon factorization, we must have $p_k = w_{\ell-1}$. Since $w_{\ell-1} \leq_{\text{lex}} w_\ell$, we obtain $p_k = w_{\ell-1} \leq_{\text{lex}} w_\ell = s$, as desired. □

The previous result combined with Theorem 14 provides us with an algorithm that rapidly computes the Nyldon factorization of a nonempty finite word. In Algorithm 1, $w[i]$ denotes the ith letter of w and $w[i,j]$ denotes the factor $w[i]\cdots w[j]$ of w. Further, $\mathrm{NylF}(i)$ designates the ith element of the list NylF while $\mathrm{NylF}(-1)$ denotes its last element.

Algorithm 1 Compute the Nyldon factorization.

Require: $w \in A^+$
Ensure: NylF is the Nyldon factorization of w

```
n ← |w|, NylF ← (w[n])
for i = 1 to n − 1 do
   NylF ← (w[n − i], NylF)
   while length(NylF) ≥ 2 and NylF(1) >_lex NylF(2) do
      NylF ← (NylF(1) · NylF(2), NylF(3), ..., NylF(−1))
   end while
end for
return NylF
```

Proposition 18. *Algorithm 1 halts for every input $w \in A^+$ and outputs the Nyldon factorization of w. Moreover, the worst case complexity of Algorithm 1 is $O\big(\frac{n(n-1)}{2}\big)$ where n is the length of the input w.*

Proof. Clearly, Algorithm 1 halts for every input $w \in A^+$ since it goes exactly $n-1$ times into the for loop and for each value of i in the $n-1$ steps of the for loop, it goes at most i times into the while loop. Then, since each lexicographic comparison between two words and basic manipulations of lists can be realized in constant time, the worst case complexity of Algorithm 1 is $O(1+2+\cdots+n-1) = O\big(\frac{n(n-1)}{2}\big)$.

Let us now prove that Algorithm 1 is correct. We claim that, for each step i of the for loop, if the variable NylF contains the Nyldon factorization of the suffix $w[n-i+1,n]$ of length i of w when it enters the for loop, then it exits with the Nyldon factorization of the suffix $w[n-i,n]$ of length $i+1$ of w. Since NylF is initialized with $(w[n])$, which is the Nyldon factorization of the suffix of length 1 of w, the claim implies that, after the last iteration of the for loop, that is for $i = n-1$, NylF contains the Nyldon factorization of w.

In order to prove the claim, we suppose that $1 \le i < n$ and that, before the ith step of the for loop, NylF contains the Nyldon factorization of the suffix $w[n-i+1,n]$ of length i of w, which we denote by $(u_1,\ldots,u_k)$. First, NylF is updated to $(w[n-i],u_1,\ldots,u_k)$, before entering the while loop.

By Theorem 14, we know that, for each $j \in \{1,\ldots,k\}$, u_j is the longest Nyldon suffix of $u_1\cdots u_j$. Then, since $w[n-i]$ is just a letter, u_j is also the longest Nyldon proper suffix of $w[n-i]u_1\cdots u_j$. From Proposition 17, we successively obtain that, for every $j \in \{1,\ldots,k\}$, the word $w[n-i]u_1\cdots u_j$ is Nyldon if and only if $w[n-i]u_1\cdots u_{j-1} >_{\mathrm{lex}} u_j$. Two cases are now possible.

Either there is some $j \in \{1, \ldots, k-1\}$ such that

$$\begin{aligned} w[n-i] &>_{\text{lex}} u_1 \\ w[n-i]u_1 &>_{\text{lex}} u_2 \\ &\vdots \\ w[n-i]u_1u_2\cdots u_{j-1} &>_{\text{lex}} u_j \\ w[n-i]u_1\cdots u_j &\leq_{\text{lex}} u_{j+1}, \end{aligned}$$

in which case the words $w[n-i]u_1\cdots u_{j'}$ are Nyldon for all $j' \in \{1, \ldots, j\}$ by Proposition 17. In this case, Algorithm 1 tells us to successively update NylF from $(w[n-i], u_1, \ldots, u_k)$ to

$$(w[n-i]u_1, u_2, \ldots, u_k),\ (w[n-i]u_1u_2, u_3, \ldots, u_k),\ \ldots,\ (w[n-i]u_1\cdots u_j, u_{j+1}, \ldots, u_k).$$

The last value of NylF, which is $(w[n-i]u_1\cdots u_j, u_{j+1}, \ldots, u_k)$, is thus the Nyldon factorization of the suffix $w[n-i, n]$ of length $i+1$ of w.

The alternative is that we have

$$\begin{aligned} w[n-i] &>_{\text{lex}} u_1 \\ w[n-i]u_1 &>_{\text{lex}} u_2 \\ &\vdots \\ w[n-i]u_1u_2\cdots u_{k-1} &>_{\text{lex}} u_k. \end{aligned}$$

In this case, Algorithm 1 tells us to successively update NylF from $(w[n-i], u_1, \ldots, u_k)$ to

$$(w[n-i]u_1, u_2, \ldots, u_k),\ (w[n-i]u_1u_2, u_3, \ldots, u_k), \ldots,\ (w[n-i]u_1\cdots u_k).$$

As before, by Proposition 17, we obtain that $w[n-i, n] = w[n-i]u_1\cdots u_k$ is Nyldon, hence the last value of NylF corresponds to the Nyldon factorization of the suffix of length $i+1$ of w. □

We then obtain an algorithm generating all Nyldon words over any alphabet A. Here we suppose that we have access to an already implemented function TestNyldon which returns True if the word is Nyldon and False otherwise. Such a function is easily implemented by combining Algorithm 1 and the fact that a nonempty finite word w is not Nyldon if and only if its Nyldon factorization is of length at least two.

Algorithm 2 Generate the Nyldon words.

Require: $\ell \geq 1$ and an alphabet A
Ensure: Nyl is the list of all Nyldon words over A up to length ℓ

```
Nyl ← A
for i = 2 to ℓ do
  Testlist ← list of all words of length i over A
  Select the elements in Testlist that pass TestNyldon and join the selected elements to
  Nyl
end for
return Nyl
```

Using the fact that the prefixes of Lyndon words are precisely sesquipowers of Lyndon words, Duval obtained a linear algorithm computing the Lyndon factorization [Duv83]. Also see Open Problem 11.

Open Problem 19. Obtain a linear or pseudo-linear algorithm computing the Nyldon factorization.

8 Standard factorization

Another important consequence of Theorem 16 is that it allows us to define a standard factorization of Nyldon words, similarly to the case of Lyndon words. Also see Section 10 where we more specifically discuss the standard factorization of the Nyldon words in relationship with the standard factorization of Lyndon words.

Theorem 20. *Let $w \in A^+$ and let $w = ps$ where s is the longest Nyldon proper suffix of w. Then w is Nyldon if and only if p is Nyldon and $p >_{\text{lex}} s$.*

Proof. The sufficient condition directly follows from Proposition 17. In order to prove the necessary condition, we suppose that w is Nyldon. If p is also Nyldon, then clearly $p >_{\text{lex}} s$ for otherwise the Nyldon factorization of w would be (p, s). So we only have to prove that p is Nyldon. Suppose to the contrary that p is not Nyldon. Then its Nyldon factorization $(p_1, \ldots, p_k)$ is of length $k \geq 2$. Since w is Nyldon, we must have $p_k >_{\text{lex}} s$. Observe that s is also the longest Nyldon proper suffix of $p_k s$. Therefore, we obtain from Proposition 17 that $p_k s$ is Nyldon. But then the Nyldon factorization of w would be $(p_1, \ldots, p_k s)$, which is of length $k \geq 2$. This contradicts that w is Nyldon. □

Definition 21. Let w be a Nyldon word over A and let s be its longest Nyldon proper suffix. Then the *standard factorization* of w is defined to be the pair (p, s), where p is such that $w = ps$.

By Theorem 20, if (p, s) is the standard factorization of a Nyldon word w, then $p >_{\text{lex}} s$. Observe that a Nyldon word w may have several factorizations of the form (u, v) such that

$$u, v \in \mathcal{N} \quad \text{and} \quad u >_{\text{lex}} v. \tag{3}$$

The standard factorization is thus a distinguished such factorization.

Example 22. The word $w = \texttt{1011101}$ is a binary Nyldon word. The two pairs $(\texttt{1011}, \texttt{101})$ and $(\texttt{1011101}, \texttt{1})$ provide factorizations of w satisfying (3). The first one is the standard factorization.

9 Complete factorizations of the free monoid and primitivity

Theorem 16 implies that the class of Nyldon words is a particular case of the complete factorizations of the free monoid A^*, as introduced by Schützenberger in [Sch65].

Definition 23. Let F be a subset of A^+ endowed with some total order $\prec$. An *F-factorization* of a word w over A is a factorization $(f_1, \ldots, f_k)$ of w into words in F such that $f_1 \succeq \cdots \succeq f_k$. Further, such a set F is said to be a *complete factorization of A^** if each $w \in A^*$ admits a unique F-factorization.

Example 24. Thanks to Theorem 2, the Lyndon words over A ordered by $<_{\text{lex}}$ form a complete factorization of A^*. Similarly, thanks to Theorem 16, the Nyldon words over A ordered by $>_{\text{lex}}$ also form a complete factorization of A^*.

In [Sch65], Schützenberger proved a beautiful and deep theorem about factorizations of the free monoid. For our needs, we only state here a particular case of Schützenberger's result.

Theorem 25 (Schützenberger)**.** *Let F be a subset of A^+ endowed with some total order $\prec$. Then any two of the following three conditions imply the third.*

(i) Each word over A admits at least one F-factorization.

(ii) Each word over A admits at most one F-factorization.

(iii) All elements of F are primitive and each primitive conjugacy class of A^+ contains exactly one element of F.

The following result is a consequence of both Theorems 16 and 25. It positively answers the third question raised by Grinberg and mentioned in the introduction.

Theorem 26. *All Nyldon words are primitive and every primitive word admits exactly one Nyldon word in its conjugacy class.*

Proof. The first two conditions of Theorem 25 are satisfied by Proposition 5 and Theorem 16 if we choose F to be the set $\mathcal{N}$ of Nyldon words and the order $\prec$ to be $>_{\text{lex}}$. Consequently, all Nyldon words are primitive and every primitive word admits exactly one Nyldon word in its conjugacy class by Theorem 25. □

A natural length-preserving bijection from Nyldon words to Lyndon words is now easy to describe: with each Nyldon word, we associate the unique Lyndon word in its conjugacy class. The Lyndon conjugate of a primitive word is easy to compute since it is lexicographically minimal among all conjugates. However, the reciprocal map which associates with a Lyndon word its unique Nyldon conjugate is much more difficult to understand. An effective construction of the Nyldon conjugate of any primitive word will be given by Algorithm 3 in Section 12.

10 Nyldon words versus Lyndon words

Lyndon words have many strong properties [Duv83, Lot97]. Some of them have analogues in terms of Nyldon words, while some of them (many in fact) do not. This section is dedicated to an in-depth comparison between Lyndon and Nyldon words. On the one hand, whenever the Nyldon and Lyndon words share some property, then we show that this property in the Lyndon case can also be obtained by only using their recursive definition. This is the case of Theorem 27 and Proposition 36. Interestingly, these new proofs of classical properties of Lyndon words are surprisingly more complicated than their analogues for Nyldon words. This highlights that the inversion of the order is not only cosmetic, but really twists the main arguments of the proofs. On the other hand, whenever some property of the Lyndon words does not stand in the Nyldon case, then we provide explicit counter-examples.

10.1 Sticking to the recursive definition

The following property of Lyndon words (Theorem 27) is well known. It is the analogue of Theorem 13 for Nyldon words. The usual proof of this result uses the classical definition of

Lyndon words: they are primitive and lexicographically minimal among their conjugates. We propose a new proof only using the recursive definition of the Lyndon words. In particular, this also provides us with another proof of Theorem 2, in the same vein as we proved the unicity of the Nyldon factorization.

Theorem 27. *Let w be a Lyndon word over A. For each Lyndon proper suffix s of w, we have $s >_{\text{lex}} w$.*

Proof. Let us stress that, inside this proof, being Lyndon means belonging to the set L defined in Corollary 3. In particular, we suppose that we know nothing else about Lyndon words than this recursive definition (such as primitivity, minimality or unicity of the Lyndon factorization).

We proceed by induction on $|w|$. If $|w| = 1$, the result is obvious. If $|w| = 2$, write $w = \mathtt{ij}$ with $\mathtt{i}, \mathtt{j} \in A$. The only Lyndon proper suffix of w is its last letter $\mathtt{j}$. Since $w, \mathtt{i}, \mathtt{j}$ are Lyndon, we must have $\mathtt{j} >_{\text{lex}} \mathtt{i}$, for otherwise $(\mathtt{i}, \mathtt{j})$ would be a Lyndon factorization of w of length at least 2. Thus, $\mathtt{j} >_{\text{lex}} w = \mathtt{ij}$.

Now, we suppose that $|w| \geq 3$ and that the result is true for all Lyndon words shorter than w. Proceed by contradiction and assume that there exists a Lyndon proper suffix s of w such that

$$s <_{\text{lex}} w. \tag{4}$$

Note that $s \neq w$ since s is a proper suffix. Among all suffixes of w satisfying (4), we choose s to be the longest and we write $w = ps$ with $p \in A^+$.

First, let us prove that p is not Lyndon. We proceed by contradiction and we suppose that p is Lyndon. Then $p <_{\text{lex}} s$, for otherwise (p, s) would be a Lyndon factorization of w of length 2. Using (4), we obtain $p <_{\text{lex}} s <_{\text{lex}} w = ps$. Therefore s must start with p. Write $s = ps_1$ with $s_1 \neq \varepsilon$. Now, since $ps_1 = s <_{\text{lex}} w = ps$, we obtain that $s_1 <_{\text{lex}} s$. Then, since $|s| < |w|$, we obtain by induction hypothesis that s_1 is not Lyndon. Therefore, there exists a Lyndon factorization $(\ell_1, \ldots, \ell_k)$ of s_1 of length $k \geq 2$. Since s is Lyndon, we must have $p <_{\text{lex}} \ell_1$, for otherwise $(p, \ell_1, \ldots, \ell_k)$ would be a Lyndon factorization of s of length at least 2. But then we have $p <_{\text{lex}} s_1 <_{\text{lex}} s = ps_1$. Therefore, p must be a prefix of s_1. Hence we obtain that p^2 is a prefix of s. Again, write $s_1 = ps_2$ with $s_2 \neq \varepsilon$. Similarly we obtain that $s_2 <_{\text{lex}} s_1$, and by using the induction hypothesis, that s_2 is not Lyndon. Then, similarly, p must be a prefix of s_2, and hence p^3 must be a prefix of s. Iterating this process, we obtain that p^n must be a prefix of s for any n, which is impossible.

Now that we know that p is not Lyndon, there must exist a Lyndon factorization $(p_1, \ldots, p_k)$ of p of length $k \geq 2$. Then we must have $p_k <_{\text{lex}} s$ for otherwise $(p_1, \ldots, p_k, s)$ would be a Lyndon factorization of w of length at least 2, contradicting that w is Lyndon.

Let us show that $p_k s$ cannot be Lyndon. Indeed, if $p_k s$ were Lyndon, then by definition of s, we would have $p_k s >_{\text{lex}} w$. But then, since $|p_k s| < |w|$, we would get that $s >_{\text{lex}} p_k s$ by the induction hypothesis. Gathering the last two inequalities, we would finally obtain $s >_{\text{lex}} w$, contradicting (4).

Now let $(\ell_1, \ldots, \ell_r)$ be a Lyndon factorization of $p_k s$ of length $r \geq 2$. As in the proof of Theorem 16, we can show that there exist $i \in \{1, \ldots, r-1\}$ and $x, y \in A^+$ such that

$$\ell_i = xy, \quad p_k = \ell_1 \cdots \ell_{i-1} x \quad \text{and} \quad s = y \ell_{i+1} \cdots \ell_r.$$

Now let $(y_1, \ldots, y_t)$ be a Lyndon factorization of y. Then y_t is a Lyndon proper suffix of ℓ_i. Since ℓ_i is Lyndon and $|\ell_i| < |w|$, the induction hypothesis yields $y_t >_{\text{lex}} \ell_i$. But then

$y_t >_{\text{lex}} \ell_i \geq_{\text{lex}} \ell_{i+1}$, hence $(y_1, \ldots, y_t, \ell_{i+1}, \ldots, \ell_r)$ is a Lyndon factorization of s of length at least 2, a contradiction with the fact that s is Lyndon. □

The converse of Theorem 27 is actually true, hence we can formulate the following theorem.

Theorem 28. *Let $w \in A^+$. Then the following assertions are equivalent.*

(i) w is Lyndon.

(ii) w is lexicographically smaller than all its nonempty proper suffixes.

(iii) w is lexicographically smaller than all its Lyndon proper suffixes.

Proof. It is well known that (i) is equivalent to (ii); see [Lot97] for example. Clearly (ii) implies (iii). We now show that (iii) implies (i), i.e., that the converse of Theorem 27 is true. We proceed by induction on the length of the words. Since the letters are Lyndon, the base case is trivially verified. Now suppose that $|w| \geq 2$, that w is lexicographically smaller than all its Lyndon proper suffixes and that, for each word z shorter than w, if z has the property to be lexicographically smaller than all its Lyndon proper suffixes, then z is Lyndon. Let us write $w = uv$ with $u, v \in A^+$. Our aim is to show that $w <_{\text{lex}} vu$. If v is Lyndon, then $w <_{\text{lex}} v$ by hypothesis, hence $w <_{\text{lex}} vu$. Now suppose that v is not Lyndon. By applying the induction hypothesis to v, we obtain that there exists a Lyndon proper suffix s of v such that $v >_{\text{lex}} s$. Since s is also a Lyndon proper suffix of w, we also have $w <_{\text{lex}} s$ by hypothesis. Consequently, we get $w <_{\text{lex}} s <_{\text{lex}} v <_{\text{lex}} vu$. □

Note that the proof that we give for the converse of Theorem 27, which corresponds to the implication $(iii) \implies (i)$, makes use of the usual definition of Lyndon words. We do not know how to prove this result by using the recursive definition of Lyndon words only.

Open Problem 29. Prove the implication $(iii) \implies (i)$ of Theorem 28 by only using the recursive definition of Lyndon words.

Interestingly, the converse of Theorem 13, which is the Nyldon analogue of Theorem 27, is *not* true, as illustrated in Example 30.

Example 30. Take $w = \texttt{1011011}$. The Nyldon factorization of w is $(\texttt{101}, \texttt{1011})$, hence w is not Nyldon. The only Nyldon proper suffixes of w are $\texttt{1}$ and $\texttt{1011}$, which are both lexicographically smaller than w.

10.2 Standard and Širšov factorizations

The following very useful characterization of Lyndon words fails in the case of Nyldon words. For a proof of this result, see for example [Lot97].

Proposition 31. *A finite word w over A is Lyndon if and only if w is a letter or there exists a factorization (u, v) of w into Lyndon words such that $u <_{\text{lex}} v$.*

By Theorem 20, we know that if a word is Nyldon, then it admits at least one factorization (u, v) into Nyldon words such that $u >_{\text{lex}} v$. However, the converse is not true as illustrated in the following example.

Example 32. Take $w = \texttt{1001010010}$. Then $(u, v) = (\texttt{10010100}, \texttt{10})$ is a factorization of w into Nyldon words such that $u >_{\text{lex}} v$. However, w is not Nyldon since its Nyldon factorization is $(\texttt{10010}, \texttt{10010})$.

We have already seen in Section 8 that, similarly to the Lyndon words, the Nyldon words admit a standard factorization. Recall that the *standard factorization* of a Lyndon word w is the factorization (u, v) of w where the right factor v is chosen to be the longest Lyndon proper suffix of w. Analogously to Theorem 20, it is well known that if (u, v) is the standard factorization of a Lyndon word w, then the left factor u is Lyndon and $u <_{\text{lex}} v$ [Lot97]. The following property of the right factors of the standard factorizations of Lyndon words does not stand in the case of Nyldon words. A proof of this result can be found in [Duv83].

Proposition 33. *The longest Lyndon proper suffix of a Lyndon word coincides with its lexicographically smallest nonempty proper suffix.*

Example 34. Let $w = \texttt{100}$. It is a binary Nyldon word. The longest Nyldon proper suffix of w is $v = \texttt{0}$. Hence the standard factorization of w is $(\texttt{10}, \texttt{0})$. However, the lexicographically greatest proper suffix of w is $\texttt{00}$.

There is another distinguished factorization of Lyndon words: the *Širšov factorization* of a Lyndon word w is the factorization (u, v) of w where the left factor u is chosen to be the longest Lyndon proper prefix of w [Šir62]. Symmetrically to the standard factorization, the Širšov factorization (u, v) has the property that its right factor v must be Lyndon and such that $u <_{\text{lex}} v$ [Vie78]. The Širšov factorization can be seen as a left standard factorization whereas the usual standard factorization can be seen as a right standard factorization. Therefore, the Lyndon words are left and right privileged words. Nyldon words, however, are not left privileged. Indeed, most properties of Lyndon prefixes do not have analogues in terms of Nyldon words, whereas Lyndon and Nyldon words share many properties concerning their suffixes (though not all of them – see for example Proposition 33 and Example 34, as well as Proposition 7).

In the next example, we illustrate that there cannot be any analogue to the Širšov factorization for Nyldon words.

Example 35. Consider the Nyldon word $w = \texttt{10010100100}$. The longest Nyldon proper prefix of w is $\texttt{100101001}$, while the corresponding suffix $\texttt{00}$ is not Nyldon.

10.3 Factorizing powers

Lyndon words are primitive by definition. But, similarly to the proof we gave of Theorem 27, even if we suppose that all we know about Lyndon words is their recursive definition, then we can easily deduce that Lyndon word are primitive and that each primitive conjugacy class contains exactly one Lyndon word from their recursive definition. Otherwise stated, we do not need the full power of Schützenberger's theorem in order to obtain the primitivity of Lyndon words. We give a new proof of the following well-known result, where we again assume that being Lyndon only means belonging to the set L defined in Corollary 3. Again, the point is to highlight the major differences in the behaviors of Lyndon and Nyldon words. In particular, the proof shows that the Lyndon factorizations of powers are straightforwardly obtained from the Lyndon conjugate of their primitive roots. In view of Example 37 below, we see that the same reasoning does not work for Nyldon words.

Proposition 36. *Lyndon words are primitive and each primitive conjugacy class contains exactly one Lyndon word.*

Proof. We proceed by induction on the length n of the words. Since all letters are Lyndon, the base case $n = 1$ is verified. Now suppose that $n > 1$ and that all Lyndon words of length less than n are primitive and that there is exactly one element in each primitive conjugacy class which is Lyndon. Let w be a power of length n. Then $w = x^m$, for some primitive word x and $m \geq 2$. By induction hypothesis, we know that x possesses a Lyndon conjugate: $y = vu$ is Lyndon and $x = uv$. Then $w = x^m = (uv)^m = u(vu)^{m-1}v = uy^{m-1}v$. Let $(u_1, \ldots, u_k)$ and $(v_1, \ldots, v_\ell)$ be the Lyndon factorizations of u and v respectively. Then $u_k \geq_{\text{lex}} y \geq_{\text{lex}} v_1$ by Theorem 27. Therefore $w = u_1 \cdots u_k \cdot y^{m-1} \cdot v_1 \cdots v_\ell$ is not Lyndon. So far, we have obtained that all Lyndon words of length n are primitive.

Now suppose that there exist distinct Lyndon words x, y of length n in the same conjugacy class. Let $x = uv$ and $y = vu$ with $u, v \neq \varepsilon$. Then, in the same way as in the previous paragraph, we obtain that x^2 has two distinct Lyndon factorizations: (x, x) and $(u_1, \ldots, u_k, y, v_1, \ldots, v_\ell)$. But Theorem 27 implies the unicity of the Lyndon factorization (similarly as Theorem 13 implies the unicity of the Nyldon factorization), hence we have reached a contradiction. □

The fact that all Nyldon are primitive is a consequence of Schützenberger's theorem; see Theorems 25 and 26. It would thus be interesting to be able to understand the Nyldon factorizations of powers. Indeed, if we could effectively deduce the Nyldon factorization of a power $w = u^k$ from the Nyldon conjugate of its primitive root u, then we would obtain a much simpler proof of the fact that no power can be Nyldon. In the following example, we show some surprising Nyldon factorizations of successive powers of some primitive word.

Example 37. The primitive word

$$u = \texttt{01111011011111011110111}$$

is not Nyldon since its Nyldon factorization is $(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, \texttt{1011111011110111})$. The Nyldon conjugate of u is

$$n = \texttt{10111101101111101111011} = \texttt{1}u\texttt{1}^{-1}.$$

We write $u = ps$, with $p = \texttt{0111101}$ and $s = \texttt{1011111011110111}$. Thus, the Nyldon factorization of u is given by $(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, s)$. Next, the Nyldon factorization of u^2 is given by $(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, x, yp, s)$, where $x = \texttt{101111}$ and $xy = s$. Table 4 stores the Nyldon factorizations of the powers of u.

k	Nyldon factorization of u^k
1	$(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, s)$
2	$(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, x, yp, s)$
3	$(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, x, y, px, yp, s)$
≥ 4	$(\texttt{0}, \texttt{1}, \texttt{1}, \texttt{1}, \texttt{101}, x, y\texttt{1}^{-1}, n^{k-4}, n\texttt{1}px, yp, s)$

Table 4: The Nyldon factorizations of the powers of $u = \texttt{01111011011111011110111}$.

In fact, it seems surprisingly difficult to understand the Nyldon factorizations of powers. We leave this concern for future work and state the following open problem.

Open Problem 38. Given a primitive word u, is it true that there exists some positive integer K such that, for all $k \geq K$, the Nyldon factorization of u^k is of the form $(p_1, \ldots, p_m, v^{k-K}, s_1, \ldots, s_n)$ where v is the Nyldon conjugate of u? If yes, characterize the smallest such k. More generally, describe the Nyldon factorizations of powers in terms of the Nyldon conjugates of their primitive roots.

10.4 About codes

We end our comparison between Nyldon and Lyndon words by a discussion on circular codes and comma-free codes. Recall that a subset F of A^* is a *code* if for any $x_1, \ldots, x_m, y_1, \ldots, y_n$ in F, we have $x_1 \cdots x_m = y_1 \cdots y_n$ if only if $m = n$ and $x_i = y_i$ for all $i \in \{1, \ldots, m\}$.

Definition 39. Let F be a code.

- F is said to be a *circular code* if for any $u, v \in A^*$, we have $uv, vu \in F^* \implies u, v \in F^*$.
- F is said to be a *comma-free code* if for any $w \in F^+$ and $u, v \in A^*$, we have $uwv \in F^* \implies u, v \in F^*$.

The Lyndon words of length n over a k-letter alphabet that form a comma-free code are completely characterized in terms of n and k, see for example [BPR10].

Theorem 40. *Let A be an alphabet of size k and let $n \geq 1$. Then the set $\mathcal{L} \cap A^n$ of Lyndon words of length n over A is a comma-free code if and only if $n = 1$, or $n = 2$ and $k \in \{2, 3\}$, or $n \in \{3, 4\}$ and $k = 2$.*

We provide an analogous result for Nyldon words. In particular, note that the Nyldon and Lyndon words of length n over some alphabet A do not necessarily form a comma-free code simultaneously. Surprisingly enough, Nyldon words more often form a comma-free code than Lyndon words. We cut the proof of Theorem 45 in several technical lemmas.

Lemma 41. *The set $\mathcal{N} \cap A$ is a comma-free code for any alphabet A.*

Proof. This is immediate since $\mathcal{N} \cap A = A$. □

Lemma 42. *Let A be an alphabet of size k. Then $\mathcal{N} \cap A^2$ is a comma-free code if and only if $k \in \{2, 3\}$.*

Proof. If $k = 2$, then $A = \{\mathtt{0}, \mathtt{1}\}$ and $\mathcal{N} \cap A^2 = \{\mathtt{10}\}$. In this case, $\mathcal{N} \cap A^2$ is clearly a comma-free code.

Suppose that $k = 3$. Then $A = \{\mathtt{0}, \mathtt{1}, \mathtt{2}\}$ and $\mathcal{N} \cap A^2 = \{\mathtt{10}, \mathtt{20}, \mathtt{21}\}$. Take $x \in (\mathcal{N} \cap A^2)^+$ and $u, v \in A^*$ such that $uxv \in (\mathcal{N} \cap A^2)^*$. Then there exists $\ell \geq 1$ and words $y_1, \ldots, y_\ell$ in $\mathcal{N} \cap A^2$ such that $uxv = y_1 \ldots y_\ell$. Since words in $\mathcal{N} \cap A^2$ are all of the same length 2, in order to show that $u, v \in (\mathcal{N} \cap A^2)^*$, it is enough to prove that x cannot start strictly within a factor y_i. Suppose to the contrary that it does. Then $i \in \{1, \ldots, \ell - 1\}$. If $y_i \in \{\mathtt{10}, \mathtt{20}\}$, then x must start with $\mathtt{0}$, whence $x \notin (\mathcal{N} \cap A^2)^+$. Thus $y_i = \mathtt{21}$. Since $x \in (\mathcal{N} \cap A^2)^+$, this implies that x starts with $\mathtt{10}$ and y_{i+1} starts with $\mathtt{0}$, which is impossible. This proves that $\mathcal{N} \cap A^2$ is a comma-free code in this case as well.

Now, suppose that $k \geq 4$. Then $\{\mathtt{0}, \mathtt{1}, \mathtt{2}, \mathtt{3}\} \subseteq A$ and the words $\mathtt{10}, \mathtt{21}, \mathtt{32}$ belong to $\mathcal{N} \cap A^2$. Take $x = \mathtt{21}$, $u = \mathtt{3}$ and $v = \mathtt{0}$. Then $uxv = \mathtt{3(21)0} = \mathtt{(32)(10)} \in N_2^*$ but $u, v \notin (\mathcal{N} \cap A^2)^*$. Therefore, $\mathcal{N} \cap A^2$ is not a comma-free code if $k \geq 4$. □

Lemma 43. *Let A be an alphabet of size k and $n \in \{3,4,5,6\}$. Then $\mathcal{N} \cap A^n$ is a comma-free code if and only if $k = 2$.*

Proof. First, assume that $k = 2$. Then $A = \{\mathtt{0}, \mathtt{1}\}$ and it is easily verified that $\mathcal{N} \cap A^n$ are comma-free codes for $n \in \{3,4,5,6\}$. Let us give some more details for the case $n = 5$. We have $\mathcal{N} \cap A^5 = \{\mathtt{10000}, \mathtt{10001}, \mathtt{10010}, \mathtt{10011}, \mathtt{10110}, \mathtt{10111}\}$. Take $x \in (\mathcal{N} \cap A^5)^+$ and $u, v \in A^*$, and assume that $uxv \in (\mathcal{N} \cap A^5)^*$. Then there exists $\ell \geq 1$ and words $y_1, \ldots, y_\ell$ in $\mathcal{N} \cap A^5$ such that $uxv = y_1 \cdots y_\ell$. Since words in $\mathcal{N} \cap A^5$ are all of the same length 5, in order to show that $u, v \in (\mathcal{N} \cap A^5)^*$, it is enough to prove that x cannot start strictly within a factor y_i. It is easily seen to be the case since all binary Nyldon words of length greater than or equal to 2 start with $\mathtt{10}$ and no Nyldon words start with $\mathtt{1010}$. This proves that $\mathcal{N} \cap A^5$ is a comma-free code. The cases $n \in \{3,4,6\}$ are similar.

Second, we show that $\mathcal{N} \cap A^n$ is not a comma-free code whenever $k \geq 3$ and $n \in \{3,4,5,6\}$. Let $k \geq 3$. Then $\{\mathtt{0}, \mathtt{1}, \mathtt{2}\} \subseteq A$. We start with the case $n = 3$. Since the words $\mathtt{101}$ and $\mathtt{210}$ both belong to $\mathcal{N} \cap A^3$ and since $\mathtt{2}(\mathtt{101})\mathtt{01} = (\mathtt{210})(\mathtt{101})$, we get that $\mathcal{N} \cap A^3$ is not a comma-free code. Similarly, since

$$\mathtt{1000}, \mathtt{1001}, \mathtt{2100} \in \mathcal{N} \cap A^4 \text{ and } \mathtt{2}(\mathtt{1001})\mathtt{000} = (\mathtt{2100})(\mathtt{1000}),$$
$$\mathtt{10000}, \mathtt{10001}, \mathtt{21000} \in \mathcal{N} \cap A^5 \text{ and } \mathtt{2}(\mathtt{10001})\mathtt{0000} = (\mathtt{21000})(\mathtt{10000}),$$
$$\mathtt{100000}, \mathtt{100001}, \mathtt{210000} \in \mathcal{N} \cap A^6 \text{ and } \mathtt{2}(\mathtt{100001})\mathtt{00000} = (\mathtt{210000})(\mathtt{100000}),$$

we get that the codes $\mathcal{N} \cap A^n$ are not comma-free for $n \in \{3,4,5\}$. □

Lemma 44. *If $n \geq 7$, then $\mathcal{N} \cap A^n$ is not a comma-free code for any alphabet A of size greater than or equal to 2.*

Proof. Suppose that $n \geq 7$ and A is an alphabet containing $\{\mathtt{0}, \mathtt{1}\}$. Then, clearly, the words $\mathtt{1011}\mathtt{0}^{n-4}, \mathtt{100101}\mathtt{0}^{n-6}$ and $\mathtt{1}\mathtt{0}^{n-4}\mathtt{100}$ belong to $\mathcal{N} \cap A^n$. Since $\mathtt{101}(\mathtt{1}\mathtt{0}^{n-4}\mathtt{100})\mathtt{101}\mathtt{0}^{n-6} = (\mathtt{1011}\mathtt{0}^{n-4})(\mathtt{100101}\mathtt{0}^{n-6})$, we obtain that $\mathcal{N} \cap A^n$ is not a comma-free code. □

Thanks to Lemmas 41 to 44, we obtain a characterization of the sets of Nyldon words of length n that form a comma-free code.

Theorem 45. *Let A be an alphabet of size k and let $n \geq 1$. Then the set $\mathcal{N} \cap A^n$ of Nyldon words of length n over A is a comma-free code if and only if $n = 1$, or $n = 2$ and $k \in \{2,3\}$, or $n \in \{3,4,5,6\}$ and $k = 2$.*

Clearly, comma-free codes are always circular codes. Further, circular codes made up with words of the same length n only contain primitive words and at most one element of each primitive conjugacy class of length n [BPR10]. Lyndon words of any fixed length n are known to form a circular code, see for example [BPR10]. However, we do not know whether this property holds true in the case of Nyldon words.

Open Problem 46. Do the Nyldon words of any fixed length n form a circular code?

11 Lazard factorizations

A well-known class of complete factorizations of the free monoid is that of Lazard sets [Vie78]. For our purpose, we need to make a distinction between left Lazard sets and right Lazard sets. In [Vie78], a Lazard factorization corresponds to what is called here a left Lazard set.

11.1 Background

Definition 47. A *left* (resp., *right*) *Lazard set* is a subset F of A^+ endowed with some total order $\prec$ and such that, for any integer $n \geq 1$, if $F \cap A^{\leq n} = \{u_1, \ldots, u_k\}$ with $k \geq 1$ and $u_1 \prec \cdots \prec u_k$ (resp., $u_1 \succ \cdots \succ u_k$), and if we consider the sequence $(Y_i)_{i\geq 1}$ of sets defined by $Y_1 = A$ and, for $i \geq 1$, $Y_{i+1} = u_i^*(Y_i \setminus \{u_i\})$ (resp., $Y_{i+1} = (Y_i \setminus \{u_i\})u_i^*$), then we have

(i) for every $i \in \{1, \ldots, k\}$, $u_i \in Y_i$

(ii) $Y_k \cap A^{\leq n} = \{u_k\}$.

A both left and right Lazard set is called a *Viennot set.*

Observe that the sets Y_i of the left (resp., right) Lazard construction are all prefix-free (resp., suffix-free): no word in Y_i is the prefix (resp., suffix) of another word in Y_i. Originally, Viennot sets were called *regular factorizations* [Vie78]. The Lyndon words form a Viennot set if $\prec$ is chosen to be $<_{\text{lex}}$. In Examples 48 and 49, we illustrate the left and right Lazard constructions of the Lyndon words. For other examples of Lazard sets, see, for instance [Vie78] and, more recently, [PR18].

Example 48. We illustrate the fact that the binary Lyndon words form a left Lazard set. We compute the binary Lyndon words up to length 5 thanks to the left Lazard construction of Definition 47. Let $A_1 = \{\mathtt{0}, \mathtt{1}\}$ and, for $i \geq 1$, define $A_{i+1} = a_i^*(A_i \setminus \{a_i\})$ with $a_i = \min_{<_{\text{lex}}}(A_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$. We see in Table 5 that the procedure from Definition 47 halts after 14 steps. The binary Lyndon words of length at most 5 are exactly $a_1, \ldots, a_{14}$. Note that the left Lazard procedure yields the Lyndon words in the increasing lexicographic order.

i	$A_i \cap \{0,1\}^{\leq 5}$	a_i	$B_i \cap \{0,1\}^{\leq 5}$	b_i
1	{0, 1}	0	{0, 1}	1
2	{1, 01, 001, 0001, 00001}	00001	{0, 01, 011, 0111, 01111}	01111
3	{1, 01, 001, 0001}	0001	{0, 01, 011, 0111}	0111
4	{1, 00011, 01, 001}	00011	{0, 00111, 01, 011}	011
5	{1, 01, 001}	001	{0, 0011, 00111, 01, 01011}	01011
6	{1, 0011, 01, 00101}	00101	{0, 0011, 00111, 01}	01
7	{1, 0011, 01}	0011	{0, 001, 00101, 0011, 00111}	00111
8	{1, 00111, 01}	00111	{0, 001, 00101, 0011}	0011
9	{1, 01}	01	{0, 00011, 001, 00101}	00101
10	{1, 011, 01011}	01011	{0, 00011, 001}	001
11	{1, 011}	011	{0, 0001, 00011}	00011
12	{1, 0111}	0111	{0, 0001}	0001
13	{1, 01111}	01111	{0, 00001}	00001
14	{1}	1	{0}	0

Table 5: The Lyndon words form a Viennot set.

Example 49. Now let us illustrate the fact that the binary Lyndon words also form a right Lazard set. We compute the binary Lyndon words up to length 5 thanks to the right Lazard construction of Definition 47. Let $B_1 = \{\mathtt{0},\mathtt{1}\}$ and, for $i \geq 1$, define $B_{i+1} = (B_i \setminus \{b_i\})b_i^*$ with $b_i = \max_{<_{\text{lex}}}(B_i \cap \{\mathtt{0},\mathtt{1}\}^{\leq 5})$. As before, the procedure from Definition 47 ends after 14 steps; see Table 5. The words $b_1, \ldots, b_{14}$ are exactly the Lyndon words of length at most 5. Observe that, this time, the right Lazard procedure yields the Lyndon words in the decreasing lexicographic order.

Remark 50. In view of Definition 47, there are two ways of thinking of the status of the order of the elements of F. First, we can think of the order on F to be induced by some preexisting order on A^*. In this case, we first fix some total order $\prec$ on A^*, and then, there are two possibilities: either there is a corresponding (left and/or right) Lazard set, or there is not. Otherwise stated, either the Lazard procedure ends for each length $n \geq 1$, meaning that, for each $n \geq 1$, there exists some $k \geq 1$ such that (i) and (ii) hold. In this case, for each length n, the left (resp., right) Lazard procedure computes the words of length n in F one by one by outputting the least (resp., greatest) words in $Y_i \cap A^{\leq n}$ (with respect to the preexisting order $\prec$ on A^*) until it reaches a singleton set $Y_k \cap A^{\leq n}$. This is what we did for the Lyndon words in Examples 48 and 49. We first considered the (increasing) lexicographic order on $\{\mathtt{0},\mathtt{1}\}^*$, and then, at each step of the procedure, we outputted the lexicographically least (resp., greatest) word in $A_i \cap \{\mathtt{0},\mathtt{1}\}^n$ (resp., $B_i \cap \{\mathtt{0},\mathtt{1}\}^n$).

In fact, we do not need to have a preexisting total order on A^* at our disposal to define a Lazard set F. Instead, we can think of the total order $\prec$ on F to be induced by the Lazard process itself. In this case, the choice of the words u_i that are removed from the sets Y_i at each step is what determines the total order on F: the fact that u_i is outputted before u_{i+1} implies that $u_i \prec u_{i+1}$ (resp., $u_i \succ u_{i+1}$) for a left (resp., right) Lazard process. In particular, since we always have $A = Y_1 \subset F$, this process always induces a total order on the alphabet A. However, there is no reason that the total order $\prec$ on F naturally extends to a total order on the free monoid A^*.

In view of the following result, Lazard sets are sometimes also called *Lazard factorizations.*

Theorem 51. *[Vie78] All left (resp., right) Lazard sets with respect to some total order $\prec$ are complete factorizations of the free monoid with respect to the same total order $\prec$.*

However, it is not true that all complete factorizations of the free monoid can be obtained by the Lazard procedure. The following result of Viennot characterizes the Lazard sets among the complete factorizations of the free monoid.

Theorem 52. *[Vie78] Let F be a complete factorization of A^* with respect to some total order $\prec$. Then F is a left (resp., right) Lazard set with respect to $\prec$ if and only if for all $f, g \in F$, $fg \in F$ implies $f \prec fg$ (resp., $g \succ fg$).*

From Theorem 52, one can deduce a characterization of Viennot sets.

Corollary 53. *Let F be a complete factorization of the free monoid A^* with respect to some total order $\prec$. Then F is a Viennot set if and only if, for all $f, g \in F$, $fg \in F$ implies $f \prec fg \prec g$.*

Example 54. Thanks to Viennot's characterization, we see once again that Lyndon words are a Viennot set with respect to the (increasing) lexicographic order $<_{\text{lex}}$. Indeed, for all

Lyndon words f and g, if fg is also Lyndon, then $f <_{\text{lex}} fg <_{\text{lex}} g$ by Theorem 28. Observe that the first inequality directly follows from the definition of the lexicographic order, and is thus verified by any words f and g. As for the second inequality, it is valid for Lyndon words only.

11.2 Nyldon words form a right Lazard factorization

In the following, we show that the Nyldon words form a right Lazard factorization, but not a left one. Consequently, the set of Nyldon words is not Viennot.

Proposition 55. *The set $\mathcal{N}$ of Nyldon words is not a left Lazard factorization.*

First proof of Proposition 55. Proceed by contradiction and suppose that $\mathcal{N}$ is a left Lazard factorization associated with some order $\prec$. In view of Theorems 16 and 51, the order $\prec$ on $\mathcal{N}$ must coincide with the decreasing lexicographic order $>_{\text{lex}}$. Thanks to Theorem 52, for all $f, g \in \mathcal{N}$ such that $fg \in \mathcal{N}$, we should have $f >_{\text{lex}} fg$, a contradiction. □

Let us now give another proof of Proposition 55, which is not based on Viennot's characterization.

Second proof of Proposition 55. If $\mathcal{N}$ were a left Lazard factorization associated with some order $\prec$, then, similarly to the previous proof, we know that the only possible choice for the order $\prec$ on $\mathcal{N}$ would be $>_{\text{lex}}$. Let us show that the set $\mathcal{N}$ cannot be obtained thanks to the left Lazard procedure with respect to $>_{\text{lex}}$. Without loss of generality, we suppose that $A = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}$, with $\mathtt{0} <_{\text{lex}} \mathtt{1} <_{\text{lex}} \cdots <_{\text{lex}} \mathtt{m}$. Suppose instead that we can produce the Nyldon words in the decreasing lexicographic order by applying the left Lazard procedure. Then, with the notation of Definition 47, for any length n, we must have $Y_1 = A$ and $u_1 = \min_{>_{\text{lex}}} Y_1 = \max_{<_{\text{lex}}} Y_1 = \mathtt{m}$. Then $Y_2 = \mathtt{m}^*\{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}-\mathtt{1}\}$. If $n \geq 3$, then the word $\mathtt{mm0}$ must eventually be outputted by the procedure. But $\mathtt{mm0}$ is not Nyldon by Proposition 7. We have thus reached a contradiction. □

Next, our aim is to show that the Nyldon words form a right Lazard set. This essentially follows from Theorem 16 and Theorem 52.

Theorem 56. *The Nyldon words equipped with the decreasing lexicographic order $>_{\text{lex}}$ form a right Lazard set.*

Proof. On the one hand, we know from Theorem 16 that the Nyldon words form a complete factorization of the free monoid A^* with respect to the decreasing lexicographic order $>_{\text{lex}}$. On the other hand, Theorem 13 tells us that, if w is a Nyldon word and if s is a Nyldon proper suffix of w, then $w >_{\text{lex}} s$. In particular, if f and g are Nyldon words such that fg is also Nyldon, then we have $g <_{\text{lex}} fg$. Thus, we obtain from Theorem 52 that the Nyldon words ordered by the decreasing lexicographic order $>_{\text{lex}}$ form a right Lazard set. □

Example 57. We illustrate that the binary Nyldon words can be obtained as a right Lazard set. Table 6 shows the sets Y_i and the words u_i of Definition 47 corresponding to $n = 5$. Note that the Nyldon words are produced in increasing lexicographic order by the right Lazard procedure, since by definition, they are produced in decreasing order with respect to the decreasing lexicographic order $>_{\text{lex}}$. Thus, at each step we must have $u_i = \max_{>_{\text{lex}}}(Y_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5}) = \min_{<_{\text{lex}}}(Y_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$.

i	$Y_i \cap \{\mathtt{0},\mathtt{1}\}^{\leq 5}$	u_i
1	{0, 1}	0
2	{1, 10, 100, 1000, 10000}	1
3	{10, 101, 1011, 10111, 100, 1001, 10011, 1000, 10001, 10000}	10
4	{101, 10110, 1011, 10111, 100, 10010, 1001, 10011, 1000, 10001, 10000}	100
5	{101, 10110, 1011, 10111, 10010, 1001, 10011, 1000, 10001, 10000}	1000
6	{101, 10110, 1011, 10111, 10010, 1001, 10011, 10001, 10000}	10000
7	{101, 10110, 1011, 10111, 10010, 1001, 10011, 10001}	10001
8	{101, 10110, 1011, 10111, 10010, 1001, 10011}	1001
9	{101, 10110, 1011, 10111, 10010, 10011}	10010
10	{101, 10110, 1011, 10111, 10011}	10011
11	{101, 10110, 1011, 10111}	101
12	{10110, 1011, 10111}	1011
13	{10110, 10111}	10110
14	{10111}	10111

Table 6: The Nyldon words form a right Lazard factorization.

Remark 58. We used the unicity of the Nyldon factorization (that is, Theorem 16) to obtain that the set of Nyldon words is a right Lazard set. Indeed, in order to be able to apply Viennot's characterization (that is, Theorem 52), we need to first know that the set of Nyldon words form a complete factorization of the free monoid.

As is well known, with any (right or left) Lazard set is associated a basis of the free Lie algebra [Vie78, Reu93]. As a consequence of Theorem 56, Nyldon words can be used in order to obtain a new basis of the free Lie algebra.

It is interesting to compare the effectiveness of the Lazard procedure for computing the Lyndon words and the Nyldon words. In fact, it seems that the Nyldon words are produced much faster by the Lazard procedure than the Lyndon words. For example, in order to compute the 14 Lyndon words up to length 5, both the left and right Lazard procedures actually need 13 steps whereas all 14 Nyldon words of length at most 5 are already computed at the fourth step. Indeed, on the one hand, in Table 6, we see that $\mathcal{N} \cap \{\mathtt{0},\mathtt{1}\}^{\leq 5} \subset Y_4$. On the other hand, we see in Table 5 that the word $a_{13} = \mathtt{01111}$ belongs to $A_{13} \setminus A_{12}$ and, similarly, the word $b_{13} = 00001$ belongs to $B_{13} \setminus B_{12}$.

In general, if $k = \#(\mathcal{L} \cap A^{\leq n})$, then both the left and right Lazard procedures need $k-1$ steps in order to compute all the Lyndon words of length up to n. More precisely, the penultimate outputted words by the left and right Lazard procedures are such that $a_{k-1} \in A_{k-1} \setminus \cup_{i=1}^{k-2} A_i$ and $b_{k-1} \in B_{k-1} \setminus \cup_{i=1}^{k-2} B_i$ respectively. Let us give more details in the case of the left Lazard procedure, the right case being symmetric. Since the left Lazard procedure outputs the Lyndon words in increasing lexicographic order, we know that $a_{k-2} = \mathtt{01}^{n-2} \in A_{k-2}$ and $a_{k-1} = \mathtt{01}^{n-1} \in A_{k-1}$. Now, because the sets A_i are prefix-free and a_{k-2} is a prefix of a_{k-1}, we obtain that $a_{k-1} \notin \cup_{i=1}^{k-2} A_i$. However, for a given n, the sets A_i and B_i are in general not of the same size; see, for instance, Table 5.

We leave it as an open problem to characterize the number of steps actually needed by the right Lazard procedure in order to produce all Nyldon words up to length n.

Open Problem 59. Let $n \geq 1$ and let $(Y_i)_{i \geq 1}$ the sequence of sets defined by $Y_1 = A$ and, for $i \geq 1$, $Y_{i+1} = (Y_i \setminus \{u_i\})u_i^*$, where $\mathcal{N} \cap A^{\leq n} = \{u_1, \ldots, u_k\}$ with $u_1 <_{\text{lex}} \cdots <_{\text{lex}} u_k$. Characterize the least integer $j \geq 1$ such that $Y_j \cap A^{\leq n}$ contains all Nyldon words over A up to length n.

Remark 60. For each Lazard set $F \subseteq A^+$ associated with some total order $\prec$, there exists a total order $<_F$ on A^* extending $\prec$ and such that an arbitrary word w over A belongs to F if and only if, for all factorizations (u, v) of w into nonempty words, we have $w <_F vu$ [Mél92]. In other words, the words in a Lazard set F are primitive and minimal among their conjugates with respect to $<_F$. Words in F are also characterized by the fact of being smaller than all their nonempty proper suffixes with respect to $<_F$. We refer to [Mél92] for the formal definition of the order $<_F$. As it happens, in the case of Lyndon words, the order $<_{\mathcal{L}}$ coincide with the lexicographic order $<_{\text{lex}}$ on A^* [Mél92]. However, even though the restriction of $<_{\mathcal{N}}$ to $\mathcal{N}$ coincide with $>_{\text{lex}}$ (as is the case for every Lazard set), the order $<_{\mathcal{N}}$ and $>_{\text{lex}}$ differs on $A^* \setminus \mathcal{N}$ since Nyldon words are not lexicographically maximal among their conjugates.

11.3 Lexicographically extremal choices in the Lazard procedure

We saw in Example 48 that the choice of the lexicographically minimal elements of the sets Y_i at each step of the left Lazard procedure leads to the set of Lyndon words. Symmetrically, the choice of the lexicographically maximal elements of the sets Y_i at each step of the right Lazard procedure also yields the set of Lyndon words; see Example 49. In the case of the Nyldon words, we have seen in Example 57 that by choosing the lexicographically minimal elements of the sets Y_i at each step of the right Lazard procedure produces the Nyldon words. Therefore, it is natural to ask what happens when we choose the lexicographically maximal elements of the sets Y_i at each step of the left Lazard procedure. As it happens, this procedure also leads to the set of Lyndon words. More precisely, we obtain the set $\overline{\mathcal{L}} = \{\overline{w} \in \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}^* \colon w \in \mathcal{L}\}$ where $\overline{\mathtt{i}} = \mathtt{m}{-}\mathtt{i}$ and $\overline{uv} = \overline{u}\,\overline{v}$ for all finite words u, v over $A = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}$. Moreover, we obtain the words of this set in the increasing lexicographic order induced by the total order $\mathtt{0} > \mathtt{1} > \cdots > \mathtt{m}$ on the letters (also called the *inverse lexicographic order* [BDFZZ18]). In fact, $\overline{\mathcal{L}}$ is also a right Lazard set, as shown by Proposition 64 below.

Definition 61. Suppose that $A = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}$ and that π is a permutation acting on A. If $\prec$ is a total order on A^* which is induced by the total order $\mathtt{0} < \mathtt{1} < \cdots < \mathtt{m}$ on the letters, then we denote by $\prec^\pi$ the corresponding total order which is induced by the total order $\pi(\mathtt{0}) < \pi(\mathtt{1}) < \cdots < \pi(\mathtt{m})$ on the letters. Moreover, we extend the definition of π on A^+ by setting $\pi(uv) = \pi(u)\pi(v)$ for all $u, v \in A^+$. Then, as usual, if F is a subset of A^+ then $\pi(F)$ designates the set $\{\pi(w) \colon w \in F\}$.

Example 62. If π is the identity, then the order $\prec^\pi$ coincide with the original order $\prec$.

Example 63. With the notation of Definition 61, the set $\overline{\mathcal{L}}$ described above corresponds to $\pi(\mathcal{L})$ where π is the permutation on $\{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}$ defined by $\pi(\mathtt{i}) = \mathtt{m}{-}\mathtt{i}$ for all $\mathtt{i}$. Thus, Proposition 64 below shows in particular that $\overline{\mathcal{L}}$ is a Viennot set ordered by the increasing lexicographic order $<_{\text{lex}}^\pi$ induced by $\mathtt{0} > \mathtt{1} > \cdots > \mathtt{m}$.

Proposition 64. *Suppose that $A = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{m}\}$ and that π is a permutation acting on A. If a subset F of A^+ is a left (resp., right) Lazard set with respect to a total order order $\prec$ which is induced by the total order $\mathtt{0} < \mathtt{1} < \cdots < \mathtt{m}$ on the letters, then the set $\pi(F)$ equipped with the total order $\prec^\pi$ is also a left (resp., right) Lazard set.*

Proof. We do the proof of the left side, the right side case being symmetric. By Theorem 51, the set F is a complete factorization of A^* ordered by $\prec$. Then, since for all $u, v \in A^+$, we have $u \prec v$ if and only if $\pi(u) \prec^\pi \pi(v)$, the set $\pi(F)$ is a complete factorization of A^* ordered by $\prec^\pi$. Thus, we can use Theorem 52 to see that $\pi(F)$ is a left Lazard set ordered by $\prec^\pi$. Let $f, g \in \pi(F)$ such that $fg \in \pi(F)$. Then $f = \pi(u)$ and $g = \pi(v)$ with $u, v \in F$. Since $\pi(uv) = fg \in \pi(F)$, we obtain that $uv \in F$, and hence that $u \prec uv$. This means that $f \prec^\pi fg$. The result now follows from Theorem 52. □

Example 65. We resume Example 63 and we illustrate the fact that $\overline{\mathcal{L}}$ is Viennot in the case of the binary alphabet $\{\mathtt{0}, \mathtt{1}\}$. We compute the words of $\overline{\mathcal{L}}$ up to length 5 thanks to the left and right Lazard constructions. Let $C_1 = D_1 = \{\mathtt{0}, \mathtt{1}\}$ and, for $i \geq 1$, define $C_{i+1} = c_i^*(C_i \setminus \{c_i\})$ (resp., $D_{i+1} = (D_i \setminus \{d_i\})d_i^*$) with $c_i = \min_{<^\pi_{\text{lex}}}(C_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$ (resp., $d_i = \max_{<^\pi_{\text{lex}}}(D_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$). Table 7 shows the successive steps of the left (resp., right) Lazard procedure for the construction of $\overline{\mathcal{L}}$. The words $c_1, \ldots, c_{14}$ (resp., $d_1, \ldots, d_{14}$) are exactly the words in $\overline{\mathcal{L}}$ of length at most 5. Observe that we have $c_1 <^\pi_{\text{lex}} \cdots <^\pi_{\text{lex}} c_{14}$ and $d_1 >^\pi_{\text{lex}} \cdots >^\pi_{\text{lex}} d_{14}$. As previously mentioned, we also have $c_i = \max_{<_{\text{lex}}}(C_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$. This is somewhat surprising since in general it is not true that $\max_{<_{\text{lex}}} S = \min_{<^\pi_{\text{lex}}} S$ if S is any subset of A^*. However, we see that $\min_{<_{\text{lex}}}(D_2 \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5}) = \mathtt{1} \neq \max_{<^\pi_{\text{lex}}}(D_2 \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5}) = d_2 = \mathtt{10000}$. Indeed, if we choose the lexicographically minimal words in the right Lazard procedure, then we obtain the Nyldon words; see Table 6. The equality

$$\max_{<_{\text{lex}}}(C_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5}) = \min_{<^\pi_{\text{lex}}}(C_i \cap \{\mathtt{0}, \mathtt{1}\}^{\leq 5})$$

is due to the fact that the sets C_i are all prefix-free.

i	$C_i \cap \{0,1\}^{\leq 5}$	c_i	$D_i \cap \{0,1\}^{\leq 5}$	d_i
1	{0, 1}	1	{0, 1}	0
2	{0, 10, 110, 1110, 11110}	11110	{1, 10, 100, 1000, 10000}	10000
3	{0, 10, 110, 1110}	1110	{1, 10, 100, 1000}	1000
4	{0, 11100, 10, 110}	11100	{1, 11000, 10, 100}	100
5	{0, 10, 110}	110	{1, 1100, 11000, 10, 10100}	10100
6	{0, 1100, 10, 11010}	11010	{1, 1100, 11000, 10}	10
7	{0, 1100, 10}	1100	{1, 110, 11010, 1100, 11000}	11010
8	{0, 11000, 10}	11000	{1, 110, 1100, 11000}	11000
9	{0, 10}	10	{1, 110}	1100
10	{0, 100, 10100}	10100	{1, 11100, 110}	110
11	{0, 100}	100	{1, 1110, 11100}	11100
12	{0, 1000}	1000	{1, 1110}	1110
13	{0, 10000}	10000	{1, 11110}	11110
14	{0}	0	{1}	1

Table 7: The set $\overline{\mathcal{L}}$ is Viennot.

Table 8 summarizes the possible extremal choices with respect to the lexicographic order of the Lazard constructions; see Examples 48, 49, 57 and 65.

Left Lazard		Right Lazard	
lex min	$\mathcal{L}$	lex min	$\mathcal{N}$
lex max	$\bar{\mathcal{L}}$	lex max	$\mathcal{L}$

Table 8: The sets obtained thanks to the four possible lexicographical extremal choices of the Lazard constructions.

Example 66. Similarly, the set $\overline{\mathcal{N}}$ is a right Lazard set ordered by the decreasing lexicographic order $>_{\text{lex}}^{\pi}$ induced by the order $\mathtt{m} < \cdots < \mathtt{1} < \mathtt{0}$ on the letters, where π is the same permutation as in Examples 63 and 65.

12 Mélançon's algorithm applied to Nyldon words

Left Lazard factorizations correspond to Hall sets, or rather *left Hall sets* [Vie78]. It is easily checked that right Lazard factorizations correspond to *right Hall sets*, as studied in [Mél92]. We refer the reader to [Vie78, Mél92] for the formal definition of Hall sets, and the fact that the notion of left (resp., right) Hall sets and of left (resp., right) Lazard sets are actually equivalent. As a consequence of Theorem 56, the set of Nyldon words is a right Hall set. Therefore, the algorithm of Mélançon can be used to find the unique Nyldon conjugate of any primitive word. In this short section, we recall this algorithm. Let us also mention here that a left version of the algorithm of Mélançon was recently used in [PR18] for computing the factorization corresponding to some new left Lazard set.

In Algorithm 3, $T(i)$ designates the ith element of the list T while $T(-i)$ denotes the $(n-i+1)$th element of T if n is the length of T.

In the same manner as for Algorithm 1, we can easily see that the worst case complexity of Mélançon's algorithm is in $O(n(n-1)/2)$ if n is the length of the input primitive word.

13 Acknowledgment

Manon Stipulanti is supported by the FRIA grant 1.E030.16. We thank Sébastien Labbé for bringing to our attention the Mathoverflow post of Grinberg. We thank Darij Grinberg for his interest in our results and for his comments on a first draft of this text. We are also grateful to the reviewers for their numerous interesting comments.

References

[BCL13] L. A. Bokut, Y. Chen, and Y. Li. Lyndon-Shirshov basis and anti-commutative algebras. *J. Algebra*, 378:173–183, 2013.

[BDFZZ18] P. Bonizzoni, C. De Felice, R. Zaccagnino, and R. Zizza. Inverse Lyndon words and inverse Lyndon factorizations of words. *Adv. in Appl. Math.*, 101:281–319, 2018.

[BP07] J. Berstel and D. Perrin. The origins of combinatorics on words. *European J. Combin.*, 28(3):996–1022, 2007.

Algorithm 3 Compute the Nyldon conjugate of a primitive word.

Require: $w \in A^+$ primitive
Ensure: NylC is the Nyldon conjugate of w

```
NylC ← list of letters of w, T ← list of letters of w
while length(NylC) > 1 do
  if T(1) = min_{<lex} NylC and T(1) <_lex T(−1) then
    T ← (T(2), ..., T(−2), T(−1) · T(1))
  end if
  i ← 2, j ← 2
  while j ≤ length(T) do
    while i ≤ length(T) and T(i) ≠ min_{<lex} NylC do
      i ← i + 1
    end while
    if i ≤ length(T) and T(i) <_lex T(i − 1) then
      T ← (T(1), ..., T(i − 1) · T(i), ..., T(−1))
    end if
    j ← i + 1, i ← i + 1
  end while
  NylC ← T
end while
return NylC
```


[BPR10] J. Berstel, D. Perrin, and C. Reutenauer. *Codes and automata*, volume 129 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010.

[CFL58] K.-T. Chen, R. H. Fox, and R. C. Lyndon. Free differential calculus. IV. The quotient groups of the lower central series. *Ann. of Math. (2)*, 68:81–95, 1958.

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

[Gri14] D. Grinberg. "Nyldon words": understanding a class of words factorizing the free monoid increasingly. `https://mathoverflow.net/questions/187451/`, 2014.

[Knu11] D. E. Knuth. *The art of computer programming. Vol. 4A. Combinatorial algorithms. Part 1.* Addison-Wesley, Upper Saddle River, NJ, 2011.

[Lot97] M. Lothaire. *Combinatorics on words.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997.

[Mél92] G. Mélançon. Combinatorics of Hall trees and Hall words. *J. Combin. Theory Ser. A*, 59(2):285–308, 1992.

[PR18] D. Perrin and C. Reutenauer. Hall sets, Lazard sets and comma-free codes. *Discrete Math.*, 341(1):232–243, 2018.

[Reu93] C. Reutenauer. *Free Lie algebras*, volume 7 of *London Mathematical Society Monographs. New Series.* The Clarendon Press, Oxford University Press, New York, 1993. Oxford Science Publications.

[Sch65] M.-P. Schützenberger. On a factorisation of free monoids. *Proc. Amer. Math. Soc.*, 16:21–24, 1965.

[Šir58] A. I. Širšov. On free Lie rings. *Mat. Sb. N.S.*, 45(87):113–122, 1958.

[Šir62] A. I. Širšov. On the bases of a free Lie algebra. *Algebra i Logika Sem.*, 1(1):14–19, 1962.

[Vie78] G. Viennot. *Algèbres de Lie libres et monoïdes libres*, volume 691 of *Lecture Notes in Mathematics.* Springer, Berlin, 1978. Bases des algèbres de Lie libres et factorisations des monoïdes libres.