Discrete Mathematics 341 (2018) 232–243

Contents lists available at ScienceDirect

Discrete Mathematics

journal homepage: www.elsevier.com/locate/disc

# Hall sets, Lazard sets and comma-free codes

Dominique Perrin [a,*], Christophe Reutenauer [b]

[a] Université Paris Est, LIGM, France
[b] Université du Québec à Montréal, LaCIM, Canada

ARTICLE INFO

*Article history:*
Received 17 January 2017
Received in revised form 15 July 2017
Accepted 23 August 2017
Available online 27 September 2017

*Keywords:*
Comma-free codes
Free Lie algebras

ABSTRACT

We investigate the relationship between two constructions of maximal comma-free codes described, respectively, by Eastman and by Scholtz and the notions of Hall sets and Lazard sets introduced in connection with factorizations of free monoids and bases of free Lie algebras.



## 1. Introduction

The notion of comma-free code has been introduced by Golomb, Gordon and Welsh [5] after the mention of their possible role in molecular genetics [3]. These codes are defined by a property of nonoverlap which makes the decoding very simple. The definition implies that the words of a comma-free code are primitive and that the code cannot contain two distinct conjugate words. Thus the cardinality of a comma-free code formed of words of length $n$ on $k$ letters is bounded by the number $p(n, k)$ of conjugacy classes of primitive words of length $n$ on $k$ letters.

It was conjectured in [5] that for every odd integer $n$ there exists a comma-free code which is a system of representatives of the conjugacy classes of words of length $n$ and thus that for odd length the maximal possible value $p(n, k)$ is reached. The result is false for even length and no formula is even conjectured for the maximal cardinality of a comma-free code of even length $n$ on $k$ letters (see [6]).

This conjecture was proved by Eastman [4] and later Scholtz [10] gave a different construction. In [1] and in [9], the construction of Scholtz was described using concepts related with factorizations of free monoids and bases of free Lie algebras, namely Hall sets and Lazard sets. These concepts introduced initially by Schützenberger form a remarkable interaction of notions from classical algebra, such as free Lie algebras and notions from information theory such as comma-free codes. They were studied extensively by Viennot [11] who introduced the terms of Lazard and Hall sets.

Recently Knuth [7] has incorporated the problem of constructing comma-free codes as an example involving techniques important for backtrack programming. He gives in particular a simplified description of Eastman's construction in Exercise 32.

The aim of this article is to show that Eastman's construction is also related to Hall and Lazard sets. We prove that there exists a Lazard set of words such that for any odd integer $n$, its words of length $n$ form the comma-free codes obtained by Eastman's construction (Theorem 21).

Eastman's construction has an advantage over Scholtz construction: for a comma-free code $X$ constructed by his method, it gives an algorithm to find the conjugate of a given primitive word which belongs to $X$. The algorithm is in polynomial time with respect to the length of the word. We will show here that this is also the case for Scholtz construction using an algorithm due to Melançon to find the conjugate of a primitive word which belongs to a given Hall set.

* Corresponding author.
*E-mail address:* perrin@univ-mlv.fr (D. Perrin).

http://dx.doi.org/10.1016/j.disc.2017.08.034
0012-365X/

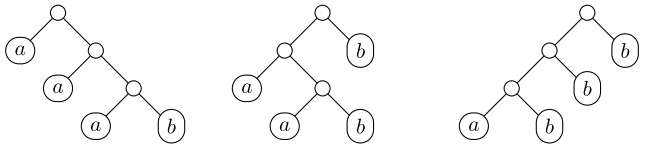


**Fig. 1.** The 3 Lyndon trees of degree 4.

The paper is organized as follows. In Section 2 we recall the definition and basic properties of Hall sets of trees and words and of Lazard sets of words.

In Section 3, we recall some definitions and properties of codes. We define circular codes and the subfamily of comma-free codes.

In Section 4.1, we give an account of the notions of dips and superdips introduced by Knuth [7] to describe Eastman's algorithm.

In Section 4.2, we describe Eastman's algorithm in terms of the notions introduced in the preceding section and prove that it gives a maximal comma-free code of length $n$ for each odd integer $n$ (Proposition 17).

In the last section, we describe the algorithm of Melançon (see [9]) which allows to find the conjugate of a primitive word which belongs to a Lazard set $Z$. It can be applied to find the conjugate of a primitive word which belongs to a comma-free code of the form $Z \cap A^n$. In particular it gives such an algorithm for the code obtained by Scholtz construction.

## 2. Hall sets and Lazard sets

We begin by recalling the notions of Hall and Lazard sets. These sets were introduced in connexion with the description of bases of free Lie algebras (for some historical background, see [9] p. 103 and [2] p. 18).

### 2.1. Hall sets

The free magma $M(A)$ on the alphabet $A$ is the set of all terms containing the letters and closed under the binary operation $x, y \mapsto (x, y)$. It can be identified with the set of complete binary trees with leaves labeled by $A$.

A totally ordered subset $H$ of $M(A)$ containing $A$ is called a *Hall set* (*of trees*) if for any $x, y \in H$, one has $(x, y) \in H$ if and only if $x < y$ and either $y \in A$ or $y = (z, t)$ with $z \leq x$. Moreover, in this case $x < (x, y)$. An element of $H$ is called a *Hall tree*.

We warn the reader that we follow here the notation of Viennot in [11]. The notation used in [9], following Lothaire [8], is symmetrical and a Hall set $H$ in [9] is such that for any $x, y \in H$, one has $(x, y) \in H$ if and only $x < y$ and either $x \in A$ or $x = (v, w)$ with $y \leq w$. Moreover, in this case $(x, y) < y$. To recover the above definition, one has to reverse the order and to take the mirror image.

We denote by $A^*$ the set of words on the alphabet $A$. The *foliage* of an element $z$ of $M(A)$ is the word $f(z) \in A^*$ defined by $f(a) = a$ if $a \in A$ and $f(x, y) = f(x)f(y)$. Thus the foliage of $z$ is obtained by erasing the parentheses when $z$ is viewed as a term and by following the frontier of the tree if $z$ is viewed as a binary tree.

A *Hall set of words* is the foliage of a Hall set of trees. Its elements are called *Hall words*

Fix a Hall set. By [9, Corollary 4.5], each Hall word is the foliage of a unique Hall tree.

Two words $x, y$ are *conjugate* if they have the form $x = uv$ and $y = vu$. Since a conjugate of a word is just cyclic shift, conjugacy is an equivalence on words. A word $x$ is *primitive* if it is not a power of another word. The conjugacy class of a primitive word is made of primitive words and a primitive word has $|x|$ distinct conjugates (see [8] for a more detailed presentation of these properties).

A *Lyndon word* is a primitive word which is minimal in its conjugacy class.

**Example 1.** The set of Lyndon words on $A$ is a Hall set. Indeed, for each Lyndon word $x$ which is not a letter, its *(left) standard factorization* is the pair $(y, z)$ such that $x = yz$ where $y$ is the longest proper prefix of $x$ which is a Lyndon word. We then associate to any Lyndon word $x$ a tree $\pi(x)$ by $\pi(a) = a$ if $a \in A$ and $\pi(x) = (\pi(y), \pi(z))$ if $(y, z)$ is the standard factorization of $x$. One may then verify the condition defining a Hall set (see [11] p. 15 or [1, Exercise 8.1.4]). The set of Lyndon words of length 4 on $A = \{a, b\}$ is $\{aaab, aabb, abbb\}$. The corresponding trees are represented in Fig. 1.

### 2.2. Lazard sets

We denote $A^{\leq n} = \varepsilon \cup A \cup \cdots \cup A^n$ the set of words of length at most $n$.

A totally ordered set $Z$ of words on the alphabet $A$ is called a *Lazard set* if the following holds. For any integer $n \geq 1$, denote the set $Z \cap A^{\leq n} = \{z_1, z_2, \ldots, z_k\}$ with

$$z_1 < z_2 < \cdots < z_k. \tag{1}$$

For $1 \leq i \leq k$, let $Z_i$ be the sequence of sets defined by $Z_1 = A$ and for $1 \leq i \leq k$,

$$Z_{i+1} = z_i^*(Z_i \setminus z_i). \tag{2}$$

Then $z_i \in Z_i$ for $1 \leq i \leq k$ and

$$Z_{k+1} \cap A^{\leq n} = \emptyset. \tag{3}$$

The set $Z_{i+1}$ obtained from $Z_i$ as in Eq. (2) is said to be the result of the *Lazard elimination* of $z_i$ in $Z_i$. The same remark on the choice of right or left made for Hall sets holds for Lazard sets. We follow here the choice of Viennot [11] and of [1]. The choice made by the second author in [9] is symmetrical. The sets $Z_i$ are defined there by $Z_{i+1} = (Z_i \setminus z_i)z_i^*$.

By a result due to Viennot [11] Corollaire 1.1 p. 35, Hall sets of words and Lazard sets coincide (see [9, Theorem 4.18]).

The Hall set of trees corresponding to a given Lazard set $Z$ is obtained via the mapping $\pi : Z \to M(A)$ defined as follows. Let $n \geq 1$, let $Z \cap A^{\leq n} = \{z_1, z_2, \ldots, z_k\}$ with $z_1 < z_2 < \cdots < z_k$ and let $Z_1, \ldots, Z_k$ be the sequence of prefix codes defined by (2). Let $z \in Z \cap A^{\leq n}$. If $z \in A$, we set $\pi(z) = z$. Otherwise, let $i$ be the least index such that $z \in Z_{i+1}$. One has then $z = z_i y$ with $y \in Z$ and we set $\pi(z) = (\pi(z_i), \pi(y))$.

**Example 2.** Let us verify that the set $L$ of Lyndon words satisfies the condition of the definition of Lazard sets with $n = 5$. One has $L \cap A^{[5]} = \{a, aaaab, aaab, aaabb, aab, aabab, aabb, aabbb, ab, ababb, abb, abbb, abbbb, b\}$. The corresponding sequence of prefix codes is

$$\begin{aligned}
Z_1 &= \{a, b\},\\
Z_2 \cap A^{\leq 5} &= \{aaaab, aaab, aab, ab, b\},\\
Z_3 \cap A^{\leq 5} &= \{aaab, aab, ab, b\},\\
Z_4 \cap A^{\leq 5} &= \{aaabb, aab, ab, b\},\\
Z_5 \cap A^{\leq 5} &= \{aab, ab, b\},\\
Z_6 \cap A^{\leq 5} &= \{aabab, aabb, ab, b\},\\
Z_7 \cap A^{\leq 5} &= \{aabb, ab, b\},\\
Z_8 \cap A^{\leq 5} &= \{aabbb, ab, b\},\\
Z_9 \cap A^{\leq 5} &= \{ab, b\},\\
Z_{10} \cap A^{\leq 5} &= \{ababb, abb, b\},\\
Z_{11} \cap A^{\leq 5} &= \{abb, b\},\\
Z_{12} \cap A^{\leq 5} &= \{abbb, b\},\\
Z_{13} \cap A^{\leq 5} &= \{abbbb, b\},\\
Z_{14} \cap A^{\leq 5} &= \{b\}.
\end{aligned}$$

The following result is analogous with Proposition 4.1 in [9] which is stated for Hall sets (and requires a total order on $M(A)$).

**Proposition 3.** *Assume that $A^*$ is totally ordered with an order such that for any $u, v \in A^*$, if $u < v$, then $u < uv$. Then there is a unique Lazard set $Z$ ordered by the restriction to $Z$ of this order.*

**Proof.** 1. Uniqueness: suppose that such a set $Z$ exists. Then by definition of a Lazard set, for any $n \geq 1$, one has $Z \cap A^{\leq n} = \{u_1, \ldots, u_k\}$, with $u_1 < u_2 < \cdots < u_k$, with

$$\begin{aligned}
u_1 &\in X_1 = A,\\
u_2 &\in X_2 = u_1^*(X_1 \setminus u_1),\\
&\cdots\\
u_k &\in X_k = u_{k-1}^*(X_{k-1} \setminus u_{k-1}),
\end{aligned}$$

and

$$X_{k+1} \cap A^{\leq n} = \emptyset. \tag{4}$$

We show that $u_i$ is the smallest element of $X_i \cap A^{\leq n}$. Observe first that $X_i \setminus u_i \subset X_{i+1}$ if $i = 1, \ldots, k$. Suppose that the smallest element $v$ of $X_i$ is not equal to $u_i$; then $v < u_i$; by the previous observation, we have $v \in X_{i+1}$ and by Eq. (4) we must have $v = u_j$ for some $j > i$; thus $u_i < v$, contradiction.

Since $u_i$ is the smallest element of $X_i \cap A^{\leq n}$, it is uniquely defined by the given order $<$, and we deduce recursively that $Z \cap A^{\leq n}$ is unique and finally that so is also $Z$.

2. Existence: let $n \geq 1$. Define $X_1 = A$, and recursively for $i = 1, 2, 3, \ldots$, $u_i = \min(X_i \cap A^{\leq n})$, $X_{i+1} = u_i^*(X_i \setminus u_i)$. One has $u_i < u_{i+1}$: indeed, $u_{i+1} = u_i^p v$, with $v \in X_i \setminus u_i$ and by definition $u_{i+1} \in A^{\leq n}$, hence $v \in X_i \cap A^{\leq n}$; note that

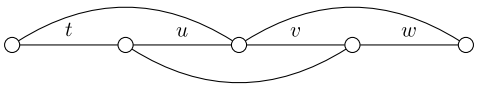


**Fig. 2.** The non-overlap condition.

$u_i = \min(X_i \cap A^{\leq n}) < v$; then either $p = 0$ and $u_i < v = u_{i+1}$; or $p > 0$ and by the property of the order, $u_i < v$, thus $u_i < u_i v$, and recursively $u_i < u_i^p v = u_{i+1}$.

We have therefore $u_1 < u_2 < u_3 < \cdots$. This implies that the words $u_i$ are all distinct and thus, since the words in $A^{\leq n}$ are finitely many, for some $k$ $X_{k+1} = \emptyset$ and the process stops.

Consider the similar construction for $n+1$. We denote by $Y_1 = A, Y_2, \ldots, Y_{l+1} = \emptyset$ and $v_j = \min(Y_j \cap A^{\leq n+1}), j = 1, \ldots, l$, the corresponding sets and words. We claim that for some integers $1 = j_1 < j_2 < \cdots < j_k \leq l$, one has $\forall i = 1, \ldots, k$, $u_i = v_{j_i}$ and that for every $j$ not equal to some $j_i$, one has $v_j \in A^{n+1}$.

The claim implies that when $n$ is replaced by $n+1$, then the set $\{v_j, j = 1, \ldots, l\}$ is equal to the disjoint union of the set $\{u_i, i = 1, \ldots, k\}$ and of a set of words of length $n+1$. This implies that if we define the $Z$ by the condition: $\forall n, Z \cap A^{\leq n} = \{u_i, i = 1, \ldots, k\}$, then $Z$ is well-defined. Moreover, the previous construction shows that $Z$ is a Lazard set.

In order to prove the claim, observe that $X_1 = Y_1$ and that if for some $i \in \{1, \ldots, k\}$ and some $j \in \{1, \ldots, l\}$, one has $X_i \cap A^{\leq n} = Y_j \cap A^{\leq n}$, then one may have two cases:

- $u_i = v_j$ (which means that $v_j \in A^{\leq n}$): then $X_{i+1} \cap A^{\leq n} = Y_{j+1} \cap A^{\leq n}$, since $X_{i+1} = u_i^*(X_i \setminus u_i\}$ and $Y_{j+1} = v_j^*(Y_j \setminus v_j)$.
- $u_i \neq v_j$ (which means that $v_j \in A^{n+1}$): then $X_i \cap A^{\leq n} = Y_{j+1} \cap A^{\leq n}$, since $Y_{j+1} \cap A^{\leq n} = v_j^*(Y_j \setminus v_j) \cap A^{\leq n} = Y_j \cap A^{\leq n}$.

This observation implies the existence of the numbers $j_i$ satisfying the claim, together with the property $X_1 = Y_1$ and: $\forall i = 1, \ldots, k$, $Y_j \cap A^{\leq n} = X_{i+1} \cap A^{\leq n}$ if $j = j_i + 1, \ldots, j_{i+1}$. ■

As the proof shows, uniqueness is true under no special assumption on the order. However, existence requires the extra property, as shows the example of an order on the free monoid $\{a, b\}^*$ satisfying $ab < a < b$: there is no Lazard set $Z$, since for $n = 2$ one should have $X_1 = A, u_1 = a, X_2 = a^*b, u_2 = ab$ and then one has not $u_1 < u_2$.

## 3. Circular and comma-free codes

In this section, we introduce some basic notions concerning codes (see [1] for a more detailed presentation).

### 3.1. Codes

Let $A$ be an alphabet. We denote by $A^+$ the set of nonempty words on $A$. A set $X \subset A^+$ is a *code* if every word in $A^*$ has at most one factorization in words of $X$. Formally, for any $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ in $X$ one has

$$x_1 \cdots x_n = y_1 \cdots y_m$$

only if $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

A *prefix code* is a set $X \subset A^+$ such that no word of $X$ is a prefix of another word of $X$. Symmetrically, a *suffix code* is a set $X \subset A^+$ such that no word of $X$ is a suffix of another word of $X$.

A code (resp. a prefix code) $X$ is *maximal* if for any code (resp. prefix code) $Y$ such that $X \subset Y \subset A^*$, one has $X = Y$ or $Y = A$.

A *circular code* is a set $X \subset A^+$ such that any word written on a circle has at most one factorization in words of $X$. Formally, for any $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ in $X$ and $p \in A^*, s \in A^+$, one has

$$sx_2 \cdots x_n p = y_1 \cdots y_m,\ x_1 = ps$$

only if $n = m$, $p = \varepsilon$, and $x_i = y_i$ for $i = 1, \ldots, n$.

It can be shown that a code $X$ is circular if and only if the submonoid $X^*$ satisfies the following condition. For any $u, v \in A^*$, one has

$$uv, vu \in X^* \Rightarrow u, v \in X^*.$$

### 3.2. Comma-free codes

Let $n \geq 1$. A set $X \subset A^n$ is a *comma-free code* if no word of $X$ overlaps nontrivially a product of two words of $X$. More precisely, for any $t, u, v, w \in A^*$, if $tu, uv, vw \in X$, then $u = w = \varepsilon$ or $t = v = \varepsilon$ (see Fig. 2).

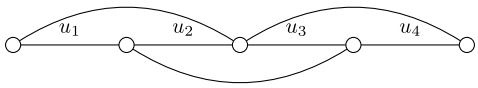


**Fig. 3.** The more general non-overlap condition.

**Table 1**
The number $\ell_n(k)$ of conjugacy classes of primitive words of length $n$ on $k$ letters.

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $\ell_n(1)$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $\ell_n(2)$ | 2 | 1 | 2 | 3 | 6 | 9 | 18 | 30 | |
| $\ell_n(3)$ | 3 | 3 | 8 | 18 | 48 | 116 | 312 | | |
| $\ell_n(4)$ | 4 | 6 | 20 | 60 | 204 | 670 | | | |
| $\ell_n(5)$ | 5 | 10 | 40 | 150 | 476 | | | | |
| $\ell_n(6)$ | 6 | 15 | 30 | 195 | | | | | |
| $\ell_n(7)$ | 7 | 21 | 27 | | | | | | |
| $\ell_n(8)$ | 8 | 28 | | | | | | | |
| $\ell_n(9)$ | 9 | | | | | | | | |

**Table 2**
Binary comma free codes with $\ell_n(2)$ elements.

| $n$ | |
|---|---|
| 2 | $ab$ |
| 3 | $aab$, $bab$ |
| 4 | $aaab$, $baab$, $bbab$ |
| 5 | $aaaab$, $aabab$, $baaab$, $babab$, $bbaab$, $bbbab$ |

One may verify that $X \subset A^n$ is comma-free if and only if the submonoid $X^*$ satisfies the following more general non-overlap condition (see Fig. 3). For any $u_1, u_2, u_3, u_4 \in A^*$, one has

$$u_1u_2, u_2u_3, u_3u_4 \in X^* \Rightarrow u_1, u_2, u_3, u_4 \in X^*. \tag{5}$$

Indeed, assume that $X$ is comma-free. Since $u_2u_3 \in X^*$, we have $u_2 = u_2'u_2''$ and $u_3 = u_3'u_3''$ with $u_2' \in X^*$, $u_2''u_3' \in X$ and $u_3'' \in X^*$. Then, by the property of non-overlap for words of $X$, either $u_2'' = \varepsilon$ or $u_3' = \varepsilon$. In both cases, $u_2, u_3 \in X^*$ and thus also $u_1, u_2 \in X^*$. The converse is obvious.

It is clear that a comma-free code is circular. In particular a comma-free code of length $n$ contains only primitive words and at most one element of each conjugacy class of primitive words of length $n$. Denote by $\ell_n(k)$ the number of conjugacy classes of primitive words of length $n$ on an alphabet with $k$ elements. It is well known that one has by the Witt Formula

$$\sum_{d|n} d\ell_d(k) = k^n \text{ and } \ell_n(k) = \frac{1}{n}\sum_{d|n} \mu(d)k^{n/d}$$

where the sums run on the divisors $d$ of $n$ and where $\mu$ is the Möbius function. The first values of the numbers $\ell_n(k)$ are tabulated in Table 1.

**Theorem 4** (*Eastman [4]*). *For any alphabet $A$ with $k$ letters and for any odd integer $n \geq 1$, there is a comma-free code $X \subset A^n$ with $\ell_n(k)$ elements.*

Comma-free codes on $A = \{a, b\}$ with $\ell_n(2)$ words of length $n$ are displayed in Table 2 for $2 \leq n \leq 5$. Let $\gamma_n(k)$ be the maximal number of elements of a comma-free code of words of length $n$ on $k$ letters. It can be shown that $\gamma_n(k) < \ell_n(k)$ for each $n = 2i$ provided $k > 2^i + i$ [6]. In particular $\gamma_2(4) = 5$, which can be verified directly (see [1] p. 292).

### 3.3. Scholtz construction

We reproduce here the construction of Scholtz [10], which implies Theorem 4 (see [1] or [9] for more details).

To describe Scholtz construction, we build a sequence $(x_i, X_i)_{i\geq 1}$ of pairs of a word $x_i$ and a prefix code $X_i$ containing $x_i$. Set $X_1 = A$ where $A$ is an alphabet with $k$ letters. For each $i \geq 1$, assuming $x_1, \ldots, x_{i-1}$ and $X_i$ already build, we choose $x_i$ in $X_i$ as a word of minimal odd length in $X_i$ and we define $X_{i+1}$ by

$$X_{i+1} = x_i^*(X_i \setminus x_i). \tag{6}$$

The following statement implies Theorem 4.

**Theorem 5** (*Scholtz [10]*). *For every odd integer $n$, the set of words of length $n$ in the union of the $X_i$ is a comma-free code with $\ell_n(k)$ elements.*

As an example, for $n = 5$ and $A = \{a, b\}$ with $a < b$, we obtain the comma-free code of Table 2.

We now relate Scholtz construction with Lazard sets. Let $U = \bigcup_{n\geq 1} X_n$. Recall that the *radix order* on $A^*$ is defined by $u < v$ if $|u| < |v|$ or if $|u| = |v|$ and $u$ precedes $v$ in lexicographic order.

**Proposition 6.** *The set $U$ is a Lazard set for the order defined by $u < v$ if $|u|$ is odd and $|v|$ even or if $|u|$, $|v|$ have the same parity and $u < v$ for the radix order.*

**Proof.** It is easy to verify that the order defined is such that if $u < v$ then $u < uv$. Indeed, if $v$ has even length, the lengths of $u$, $uv$ have the same parity and $u$ is shorter than $uv$. And if $v$ has odd length, then $u$ has also odd length and $uv$ has even length, whence the conclusion. Thus the result follows by Proposition 3. ■

## 4. Eastman's construction

We now come to the description of Eastman's construction, as presented in [7, Exercise 32].

### 4.1. Dips and superdips

Let $A$ be a totally ordered alphabet with at least two elements. Consider the set

$$D(A) = \{a_1 a_2 \cdots a_n \mid a_i \in A, n \geq 2, a_1 \geq a_2 \geq \cdots \geq a_{n-1} < a_n\}.$$

The elements of $D(A)$ are called *dips*.

**Example 7.** For $A = \{a, b, c\}$ and $a < b < c$, we have $D(A) = c^*b^*a^*(ab + ac) + c^*b^*bc$.

Let $S(A)$ be the words formed of a dip of odd length followed by a (possibly empty) sequence of dips of even length. The elements of $S(A)$ are called *superdips*.

Note that, on a finite alphabet $A$, the set $S(A)$ is included in the set $U = \cup_{n\geq 1} X_n$ obtained by Lazard eliminations of the shortest word of odd length, as in Scholtz algorithm. However, the iteration of Scholtz algorithm produces words which are not in $S(A)$ as shown in the following example.

**Example 8.** The word $w = a^{10}ca^2ca^2ba^2ca^{12}b$ is in the comma-free code obtained by Scholtz algorithm with the factorization $(a^{10}c, ((a^2c, (a^2b, a^2c)), a^{12}c))$. However $w$ is a product of dips of odd length, and thus is not in $S_1(A)$ (see the continuation of this example in Example 18).

Let $X \subset A^+$ be a maximal prefix code. A word $x \in X^*$ is *synchronizing* if for any $u \in A^*$, one has $ux \in X^*$.

**Proposition 9.** *The set $D(A)$ is a maximal prefix code such that each word of $D(A)$ of length at least 3 is synchronizing.*

**Proof.** It is clear that $D(A)$ is a prefix code. To show that it is maximal, assume that $w$ has no prefix in $D(A)$. Then $w = a_1a_2\cdots a_k$ with $a_1 \geq \cdots \geq a_k$. If $a_k$ is not the largest letter, then $wb \in D(A)$ for a letter $b > a_k$. Otherwise, $wab$ is in $D(A)$ for $a_k > a < b$ ($a$, $b$ exist since $A$ has at least two elements).

Let $u \in A^*$ and let $x \in D(A)$ be of length at least 3. We show that $ux \in D(A)^*$, which will imply that $x$ is a synchronizing word. Set $x = pbac$ with $p \in A^*$ and $a, b, c \in A$ and $b \geq a < c$. Set $ux = yq$ with $y \in D(A)^*$ and $q$ a proper prefix of $D(A)$. Since $a < c$, the word $ac$ is not an internal factor of $D(A)$ nor a proper prefix of $D(A)$. This implies that $q = c$ or $q = 1$. The first case is not possible because $y$ would end with $ba$ which is not a suffix of $D(A)$. Thus $ux \in D(A)^*$. ■

Note that a word of $D(A)$ of length 2 may be not synchronizing. Indeed, for $a < b < c$, we have $bc \in D(A)$ although $aabc \notin D(A)^*$ and thus $bc$ is not synchronizing.

**Proposition 10.** *The set $S(A)$ is a code on the alphabet $A$.*

**Proof.** The set $S(A)$ is a suffix code on the alphabet $D(A)$ and $D(A)$ is a prefix code on the alphabet $A$. ■

Recall that a code $X \subset A^+$ is *thin* if there exists a word $w \in A^*$ which is not a factor of $X$ and that $X$ is *complete* if every word in $X^*$ is a factor of $X^*$.

Note that $S(A)$ is actually a maximal code on the alphabet $A$. Indeed, $D(A)$ is a thin maximal prefix code on the alphabet $A$ and $S(A)$ is a thin maximal suffix code on the alphabet $D(A)$. Thus $S(A)$ is a thin and complete code on the alphabet $A$ by [1, Proposition 2.6.13].

**Proposition 11.** *Any primitive word $w$ of odd length $m \geq 3$ has a conjugate in $D(A)^*$ and thus in $S(A)^*$. More precisely, let $u$ be the factor of $w^3 = puv$ such that $p$ is the shortest prefix of $w^3$ which ends with a dip of length at least 3. Then $u$ is a conjugate of $w$ which is in $D(A)^*$.*

**Proof.** We first show that $w$ has a conjugate in $D(A)^*$. Let $p$ be the shortest prefix of $w^2$ which ends with a dip of length at least 3.

To show the existence of $p$, we consider a factorization $w = uav$ with $a$ the largest letter among the letters occurring in $w$. The word $vua$ has at least some ascent (that is, a factor $bc$ with $b < c$), since $a$ is its largest letter, and since $vua$ is of length at least 3 and not a power of $a$. Hence $vua$ has a prefix $q$ which is a dip of length at least 2. Thus $w^2$ has the prefix $p = uaq$ which ends with a dip of length at least 3.

Assume first that $p$ is a prefix of $w$. Set $w = ps$. Since $p$ is synchronizing, we have $sp \in D(A)^*$, whence the conclusion. Assume next that $w$ is a prefix of $p$. Set $p = wr$ and $w = rs$. Since $p$ is synchronizing, we have $p, wp \in D(A)^*$. Since $wp = wwr = wrsr = psr$, we have $sr \in D(A)^*$.

We may now assume that $w \in D(A)^*$. Since $w$ has odd length, at least one of the dips forming $w$ has odd length. The conjugate starting before this dip is in $S(A)^*$. ■

**Example 12.** We illustrate Proposition 11 by finding the conjugate of the word $w = abracadabra$ which is in $S(A)^*$.

The shortest prefix of $w^2$ which ends with a dip of length at least 3 is $p = abrac$. The conjugate of $w$ starting after $p$ is *ad ab raab rac* which factorizes in dips as indicated. Its conjugate *rac ad ab raab* is in $S(A)$

**Proposition 13.** *No word of $S(A)$ overlaps nontrivially the product of two words in $S(A)$, in the sense that for $x, y, z \in S(A)$, if $x = x_1x_2$ with $y, x_1$ (resp. $z, x_2$) comparable for the suffix (resp. prefix) order, then $x_1$ or $x_2$ is empty.*

**Proof.** Since both $y$ and $z$ are superdips, each one is in a unique way a product of dips. Let $a_1a_2 \cdots a_k$ be the last dip of $y$ in this factorization and let $b_1b_2 \cdots b_\ell$ be the first dip of $z$. Note that since $z \in S(A)$, $\ell$ is odd. We assume that $x_1, x_2$ are nonempty.

Assume first that $x_1$ is a letter. Then $x_1 = a_k$. If $a_k < b_1$, the first dip of $x$ is $a_kb_1$ and has even length, which is impossible. Otherwise, the first dip of $x$ is $a_kb_1b_2 \cdots b_\ell$ and thus has also even length, a contradiction.

Assume next that $x$ has length 2. Then $x_1 = a_{k-1}a_k$. But then the first dip of $x$ has length 2, which is again impossible.

Thus $x_1$ has length at least 3. We distinguish two cases.

**Case 1.** Assume that $x_1$ is shorter than $y$. Set $y = ux_1$ (see Fig. 4 on the left). Since $x_1$ ends with $a_{k-1}a_k$ which is not an internal factor of $D(A)$, the first dip of $x$ is a prefix of $x_1$. Set $x_1 = ts$ where $t$ is the first dip of $x$. Since $t$ has odd length, its length is at least 3. Since a dip of length at least 3 is synchronizing, and since $y = ux_1 = uts$, we have $ut, s \in D(A)^*$ and thus also $x_1 \in D(A)^*$ since $t, s \in D(A)^*$. This implies that $x_2 \in D(A)^*$ a contradiction, since the first dip of $x_2$ is equal to the first dip of $z$ which has odd length and since $x$ is a superdip.

**Case 2.** Assume now that $x_1$ is longer than $y$. Set $x_1 = vy$ (see Fig. 4 on the right). Let $t$ be the first dip of $y$. Since $t$ is synchronizing as in Case 1, we have $t, vt \in D(A)^*$. Set $y = ts$. Since $y, t \in D(A)^*$, we have $s \in D(A)^*$. Thus $x_2 \in D(A)^*$ a contradiction, as in Case 1. ■

The following corollary shows that the submonoid $S(A)^*$ satisfies the non-overlap Condition (5).

**Corollary 14.** *For any $u_1, u_2, u_3, u_4 \in A^*$ such that $u_1u_2, u_2u_3, u_3u_4 \in S(A)^*$, one has $u_2, u_3 \in S(A)^*$.*

**Proof.** We first note that the statement holds if $u_2$ or $u_3$ is empty. We thus assume the contrary and set $u_2u_3 = x_1x_2 \cdots x_k$ with $k \geq 1$ and $x_i \in S(A)$ for $1 \leq i \leq k$. Similarly, set $u_3u_4 = z_1 \cdots z_m$ with $m \geq 1$ and $z_i \in S(A)$ for $1 \leq i \leq m$. Similarly, set $u_1u_2 = y_1 \cdots y_\ell$ with $\ell \geq 1$ and $y_i \in S(A)$ for $1 \leq i \leq \ell$.

We may assume that $u_1$ is a prefix of $y_1$ since otherwise we may simplify by $y_1$ on the left. Then there is a unique index $i$ with $1 \leq i \leq k$ such that in the prefix order (see Fig. 5)

$$u_1x_1 \cdots x_{i-1} \leq y_1 < u_1x_1 \cdots x_i.$$

Thus $x_i$ overlaps $y_1y_2$ (or $y_1z_1$ if $\ell = 1$). By Proposition 13, we have $y_1 = u_1x_1 \cdots x_{i-1}$. This implies that $u_2 = x_1 \cdots x_{i-1}y_2 \cdots y_\ell$ and thus $u_2 \in S(A)^*$. We have consequently $u_3 \in D(A)^*$ since $D(A)$ is a prefix code and finally $u_3 \in S(A)^*$ since its first dip is also the first dip of $z_1$ and has odd length. ■

### 4.2. Eastman algorithm

Let $(S_n(A))_{n\geq 0}$ be the sequence of maximal codes on the alphabet $A$ obtained as follows. We start with $S_0(A) = A$. For $n \geq 1$, let $D_n(A) = D(S_{n-1}(A))$ and $S_n(A) = S(S_{n-1}(A))$ where $D, S$ are defined in the previous exercise with $S_{n-1}(A)$ instead of $A$ as an alphabet, using the order on $S_{n-1}(A)$ induced by the radix order on $A^*$.

Note that each $S_n(A)$ is formed of words of odd length. Indeed, this is true for $n = 0$ and assuming that it is true for $n - 1$, we obtain the property for $D_n(A)$ using the following lemma whose proof is left to the reader. A *graded alphabet* is a set $A$ with a map $\deg : A \to \mathbb{N}$ associating to each letter its degree. The degree of a word is the sum of the degrees of its letters.

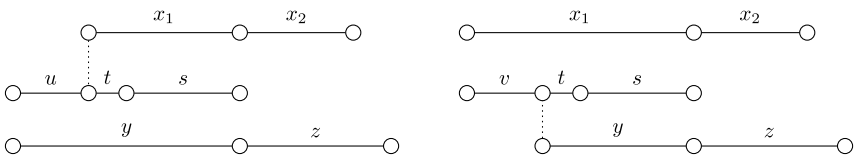


**Fig. 4.** The two cases.

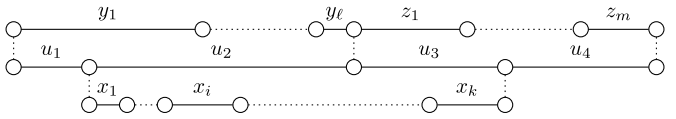


**Fig. 5.** The proof of Corollary 14.

**Lemma 15.** *Let $A$ be a graded alphabet such that* $\deg(a)$ *is odd for every $a \in A$. Then every word of odd length has odd degree.*

The elements of $D_n(A)$ are called $n$-dips and those of $S_n(A)$ are called $n$-superdips. Recognizing if a word $w$ is in $D_n(A)^*$ or $S_n(A)^*$ can be done operating for increasing values of $i = 1, \dots, n$. Assume that $w \in S_{i-1}(A)^*$. Then one may use a left to right scan of $w$ to write $w = xp$ with $x \in D_i(A)^*$ and $p$ a proper prefix of $D_i(A)$. Set $x = x_1 \cdots x_k$ with $x_j \in D_i(A)$. Selecting the blocks beginning with an odd $i$-dip, we obtain $x = qy_1 \cdots y_\ell$ with $y_j \in S_i(A)$. Then $w \in S_i(A)^*$ if and only if $p = q = \varepsilon$.

Consider the following algorithm to compute a conjugate of a primitive word $w \in A^*$ of odd length $m \geq 3$ which is in $S_n(A)$ for some $n$. For successive values of $n \geq 1$, we perform the following steps for a word $w \in S_{n-1}(A)^*$ with length at least 3 on the alphabet $S_{n-1}(A)$.

1. Let $p$ be the shortest prefix of $w^2$ which ends with an $n$-dip of length at least 3 on the alphabet $S_{n-1}(A)$ (such a prefix exists because the length of $w$ on the alphabet $S_{n-1}(A)$ is at least 3). Replace $w$ by its conjugate starting after $p$. Now $w \in D_n(A)^*$ (see below).
2. Let $q$ be the first dip of odd length of $w$ (it exists because $w$ has odd length). Replace $w$ by its conjugate starting before $q$. Now $w \in S_n(A)^*$.

We stop when $w$ is in $S_n(A)$. It follows from Proposition 11, since $D_n(A) = D(S_{n-1}(A))$, that the conjugate of $w$ chosen as in step 1 of the algorithm is in $D_n(A)^*$.

**Example 16.** We perform the above algorithm on the word *abracadabra* (already considered in Example 12).

Write periodically the word *abracadabra* and factorize it in dips. We obtain

$$abracadabra\ abracadabra \cdots = ab\ rac\ ad\ ab\ raab\ rac\ ad\ ab\ raab \cdots$$

We thus replace *abracadabra* by its conjugate

$$racadabraab = rac\ ad\ ab\ raab.$$

Since it is a superdip (a dip of length 3 followed by dips of lengths 2, 2, 4), the algorithm stops.

We conclude from the above that one has the following result.

**Proposition 17.** *For every odd integer $m \geq 3$, the set of words of length $m$ which belongs to some $S_n(A)$ is a comma-free code which meets all conjugacy classes of primitive words of length $m$.*

**Proof.** Set $U = \cup_{n \geq 0} S_n(A)$. The set $U \cap A^m$ meets every conjugacy class of primitive words of length $m$ because the algorithm above gives this conjugate. To verify that it is comma-free, we prove by induction on $|x| + |y| + |z|$ that a word $x$ in $U$ does not overlap nontrivially a product $yz$ of words $y, z \in U$ in the sense that if $x = x_1x_2$ with $x_1$ a proper suffix of $y$ and $x_2$ a proper prefix of $z$, then $x_1$ or $x_2$ is empty.

The property is true if one of $x, y, z$ is a letter.

Otherwise, we have $x, y, z \in S_1(A)^*$. Set $y = ux_1$ and $z = x_2v$. By Corollary 14, we have $x_1, x_2 \in S_1(A)^*$. Since $S_n(A) = S_{n-1}(S_1(A))$ for all $n \geq 1$, we may apply the induction hypothesis to the words $x', y', z'$ obtained from $x, y, z$ by considering $S_1(A)$ as a new alphabet, obtaining the conclusion. ■

The following example shows that the comma-free codes obtained by Eastman algorithm are not the same as those obtained by Scholtz algorithm.

**Example 18.** Let $A = \{a, b, c\}$, the word $a^{12}ca^{10}ca^2ca^2ba^2c$ is in $S_2(A)$ since all words $a^{12}c$, $a^{10}c$, $a^2c$, $a^2b$ are odd length dips and $a^{12}c \geq a^{10}c \geq a^2c \geq a^2b < a^2c$. Thus it belongs to the comma-free code of length 33 obtained by Eastman algorithm. But it is not in the code obtained by Scholtz algorithm which contains the conjugate $a^{10}ca^2ca^2ba^2ca^{12}c$, as we have seen in Example 8.

## 5. A new Lazard set

The aim of this section is to show that the comma-free code obtained by Eastman algorithm is a Lazard set and thus can be obtained by an elimination method, as in Scholtz construction, although using a different order.

### 5.1. A new order on words

Let $D_n(A)$ be the set of $n$-dips as defined in the previous section and let $P_n(A)$ denote the set of proper prefixes of $D_n(A)$, considered as a set of words on the alphabet $A$. One has

$$A^* \supset D_1(A)^+P_1(A) \supset \cdots \supset D_n(A)^+P_n(A) \supset \cdots$$

For $w \in A^*$, the *index* of $w$, denoted index$(w)$, is the largest integer $n$ such that $w \in D_n(A)^+P_n(A)$. This integer is bounded by the length of $w$ since a word of $D_n(A)$ has length at least 3 on the alphabet $S_{n-1}(A)$ and a fortiori on the alphabet $D_{n-1}(A)$.

**Example 19.** The index of *aabaa* is 1 while the index of *aabaac* is 2.

Note that if $x \in S_n(A)$, then index$(x) = n$. Indeed, $x \in D_n(A)^+$ and $x$ cannot have a prefix in $D_{n+1}(A)$, since the words of $D_{n+1}(A)$ have length at least 2 on the alphabet $S_n(A)$.

We define the following order on $A^*$. For $x, y \in A^*$, we define $x \prec y$ if

(i) $x$ has odd length and $y$ has even length, or
(ii) the lengths of $x$ and $y$ have the same parity and
  (a) index$(x) <$ index$(y)$ or
  (b) index$(x) =$ index$(y)$ and $x < y$ for the radix order.

With the objective of applying Proposition 3, we prove the following property of this order.

**Proposition 20.** *For $x, y \in A^*$, if $x \prec y$, then $x \prec xy$.*

**Proof.** Assume first that $y$ has even length. Then the lengths of $x$, $xy$ have the same parity. Assume that index$(x) = n$. Set $x = zp$ with $z \in D_n(A)^+$ and $p \in P_n(A)$. Since $D_n(A)$ is a maximal prefix code, we have $py \in D_n(A)^*P_n(A)$. Thus $xy \in D_n(A)^+P_n(A)$ showing that index$(xy) \geq n$. If index$(xy) > n$, then $x \prec xy$. Otherwise, $x$ is shorter than $xy$ and thus $x \prec xy$.

Next, if $y$ has odd length, then $x$ is also of odd length and $xy$ has even length, which implies $x \prec xy$. ■

### 5.2. Main result

Let $Z$ be the Lazard set corresponding to the order $\prec$, (which exists and is unique by Propositions 3 and 20).

Let $\Delta = \cup_{n\geq 1}D_n(A)$ and $\Sigma = \cup_{n\geq 0}S_n(A)$, where $D_n(A)$ is the set of $n$-dips on $A$ and $S_n(A)$ the set of $n$-superdips on $A$ (see Section 4.2).

The following result is the main result of this paper.

**Theorem 21.** *For any odd integer $m$, the set of words of $\Sigma$ of length $m$ is equal to the set of words of length $m$ in $Z$.*

### 5.3. Proof of the main result

Let $N \geq 1$ be an integer and set $Z \cap A^{\leq N} = \{z_1, z_2, \ldots, z_M\}$ with $z_1 \prec z_2 \prec \ldots \prec z_M$. Let $Z_1, \ldots, Z_{M+1}$ be the sequence of maximal prefix codes defined, starting with $Z_1 = A$, by $Z_{n+1} = z_n^*(Z_n \setminus z_n)$.

The *standard factorization* of a word $z$ of $Z \cap A^{\leq N}$ which is not a letter is the pair $(z_n, y)$ such that $z = z_ny$ with $1 \leq n < M$ and $y \in z_n^*(Z_n \setminus z_n)$. In this way the Hall tree corresponding to $z$ is $\pi(z) = (\pi(z_n), \pi(y))$.

We need another definition. For some $k \geq 0$, an element $y \in Z$ is *consistently* in $D_{k+1}(A)$ if $y = y_1 \cdots y_s$ with $y_i \in S_k(A) \cap Z$, $y_1 \geq \cdots \geq y_{s-1} < y_s$, and $\pi(y) = (\pi(y_1), \pi(y_2 \cdots y_s))$. Thus, when $y$ is consistently in $D_{k+1}(A)$, its standard factorization is given by its $(k + 1)$-dip structure.

Note that any element of $D_1(A)$ is in $Z$ (for large enough $N$) and is consistently in $D_1(A)$. Indeed, if $y = a_1 \cdots a_k$ with $a_i \in A$ and $a_1 \geq \cdots \geq a_{k-1} < a_k$, then $\pi(y) = (a_1 \ldots (a_{k-1}, a_k))$.

Note also that there can be elements of $Z \cap D_{k+1}(A)$ which are not consistently in $D_{k+1}(A)$, as shown in the example below.

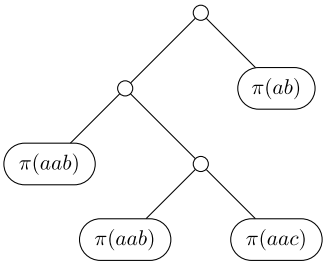


**Fig. 6.** The tree $\pi(y)$.

**Example 22.** The word $z = aabaabaacab$ is in $D_2(A)$ since $aab, aacab \in S_1(A)$ and $aab \geq aab < aacab$. We have also $z \in Z$ with the decomposition $\pi(y) = ((\pi(aab), (\pi(aab), \pi(aac))), \pi(ab))$. But the two decompositions do not coincide (see Fig. 6 where the tree is only partially developed).

We will use two lemmas concerning these notions. The first one describes a situation where the structure of the Lazard set coincides with that of the dip and superdip structure.

**Lemma 23.** *Let $k \geq 0$ and $n \geq 1$ be such that $z_n \in S_k(A)$ and $y \in Z_{n+1}$ with $|y|$ odd. If $y$ is consistently in $D_{k+1}(A)$, then $z_n y$ is consistently in $D_{k+1}(A)$.*

**Proof.** Set $y = y_1 \cdots y_s$ with $y_i \in S_k(A) \cap Z$ and $y_1 \geq \cdots \geq y_{s-1} < y_s$. By the hypothesis, $y_1 = z_m$ with $m < n$ and thus $y_1 \prec z_n$. Thus $z_n \geq y_1 \geq \cdots \geq y_{s-1} < y_s$, which shows that $z_n y$ is in $D_{k+1}(A)$. It is consistently in $D_{k+1}(A)$ since $\pi(z_n y) = (\pi(z_n), \pi(y))$. ■

The next lemma handles a case where the two structures are distinct.

**Lemma 24.** *For $k \geq 0$, if $x \in S_k(A)$ and $y \in D_m(A)$ with $|y|$ even and $1 \leq m \leq k$, then $xy \in S_k(A)$.*

**Proof.** We prove the statement by induction on $k - m$. It is true if $m = k$ by definition of $S_k(A)$. Assume $m < k$ and set $x = u_1 \cdots u_s$ with $u_i \in D_k(A)$. Set $u_s = v_1 \cdots v_t$ with $v_i \in S_{k-1}(A)$ and $v_1 \geq \cdots \geq v_{t-1} < v_t$. We have $v_t y \in S_{k-1}(A)$ by induction hypothesis. Since $v_t < v_t y$ in the radix order, this implies $u_s y = v_1 \cdots v_{t-1}(v_t y) \in D_k(A)$. Since $|y|$ is even, the lengths of $u_s$ and $u_s y$ have the same parity. This implies that $xy = u_1 \cdots u_{s-1}(u_s y)$ is in $S_k(A)$. ■

**Example 25.** Set $x = aabaabaac$ and $y = ab$ as in Example 22. Then, as we have seen, $xy \in D_2(A)$ but the decomposition of $xy$ in 1-dips is $(aab, aab, aacab)$ although $\pi(xy) = ((\pi(aab), (\pi(aab), \pi(aac))), \pi(ab))$.

**Proposition 26.** *The words of odd length in $Z$ are in $\Sigma$.*

**Proof.** For a given integer $k \geq 0$, we define a set $\mathcal{P}_k \subset \Delta \cup \Sigma$ which is formed of the words $y \in Z$ such that the following holds.

(i) If $|y|$ is odd, then $y$ is in $S_k(A)$ or consistently in $D_{k+1}(A)$.
(ii) If $|y|$ is even, then $y$ is consistently in $D_{k+1}(A)$ or in $D_m(A)$ with $m \leq k$.

We prove by induction on $n \geq 1$ that if $n = 1$ or if $z_{n-1}$ has odd length, one has $Z_n \subset \Delta \cup \Sigma$, and more precisely that, for some integer $k \geq 0$, all words of $Z_n$ are in $\mathcal{P}_k$.

This is true for $n = 1$ with $k = 0$. Indeed, $Z_1 = A = S_0(A)$. Thus $Z_1 \subset \mathcal{P}_0$.

Assume now that $n \geq 2$ and that $z_{n-1}$ has odd length. Then by the induction hypothesis $Z_{n-1} \subset \mathcal{P}_{k-1}$ for some $k \geq 1$. Since $z_{n-1}$ has odd length it is either in $S_{k-1}(A)$ or in $D_k(A)$.

**Case 1** Assume first that $z_{n-1} \in S_{k-1}(A)$. We show that in this case $Z_n \subset \mathcal{P}_{k-1}$. We have to prove that for $y \in Z_n \cap \mathcal{P}_{k-1}$, one has $z_{n-1} y \in \mathcal{P}_{k-1}$.

1.1 Suppose that $y$ has odd length. Then $y$ is in $S_{k-1}(A)$ or consistently in $D_k(A)$.

1.1.1 If $y \in S_{k-1}(A)$, then $z_{n-1} y$ is consistently in $D_k(A)$ and thus $z_{n-1} y \in \mathcal{P}_{k-1}$
1.1.2 If $y \in D_k(A)$, then $z_{n-1} y$ is consistently in $D_k(A)$ by Lemma 23 and thus $z_{n-1} y \in \mathcal{P}_{k-1}$ .

1.2 Suppose now that $y$ has even length. Then $y$ is either consistently in $D_k(A)$ or in $D_m(A)$ for $m < k - 1$.

1.2.1 If $y$ is consistently in $D_k(A)$, then $z_{n-1} y$ is consistently in $D_k(A)$ by Lemma 23.
1.2.2 Otherwise, it is in $D_m(A)$ for $m < k - 1$ and thus $z_{n-1} y$ is in $S_{k-1}(A)$ by Lemma 24.

In both cases, $z_{n-1} y \in \mathcal{P}_{k-1}$.

**Case 2** Assume now that $z_{n-1}$ is in $D_k(A)$. We show that in this case $Z_n \subset \mathcal{P}_k$. Since $z_{n-1} \prec z$ for every $z \in Z_{n-1} \setminus z_{n-1}$, the words of odd length in $Z_{n-1}$ cannot be in $S_{k-1}(A)$ since otherwise its index is less than the index of $z_{n-1}$. Since $\mathcal{P}_{k-1} \setminus S_{k-1} \subset \mathcal{P}_k$, we have $Z_{n-1} \subset \mathcal{P}_k$. Thus, we have to prove that for any $y \in Z_n \cap \mathcal{P}_k$, we have $z_{n-1}y \in \mathcal{P}_k$. The proof is the same as in Case 1 with $k$ instead of $k-1$. ■

**Proof of Theorem 21.** By Proposition 26, the set $Z \cap A^m$ is contained in $\Sigma \cap A^m$. Since $Z$ is a Lazard set, by [1, Proposition 8.1.10], the family $\{z \mid z \in Z\}$ is a complete factorization of $A^*$. Thus, by [1, Corollary 8.1.7], it is a set of representatives of the primitive conjugacy classes. Since $\Sigma \cap A^m$ is comma-free by Proposition 17, this forces $Z \cap A^m = \Sigma \cap A^m$. ■

The following example illustrates the proof.

**Example 27.** Let $A = \{a, b, c\}$. We have $z_i \in S_0(A) = A$ for $1 \le i \le 3$ and $Z_i \subset \mathcal{P}_0$ for $1 \le i \le 4$ in agreement with Case 1. Then $z_4 = a^2b$ is in $D_1(A)$ and $Z_5 \subset \mathcal{P}_1$ in agreement with Case 2. The smallest element of $Z$ which is in $D_2(A)$ is $a^2ba^2ba^2c$, which is equal to $z_{n-1}$ for some $n \ge 6$. Accordingly, we have $Z_i \subset \mathcal{P}_1$ for $5 \le i \le n-1$ and $Z_n \subset \mathcal{P}_2$.

Consider again the case of $x = a^2ba^2ba^2c$ and $y = ab$ (Example 25). Since $x \in S_2(A)$ and $y \in D_1(A)$, we are in case 1.2.2 with $k = 3$.

### 5.4. The Melançon algorithm

The last part, taken from [9], describes an algorithm due to Melançon to find the conjugate of a primitive word which belongs to a Lazard set. We include it here because it gives an algorithm to compute the conjugate of a word of odd length which belongs to the comma-free code obtained by Scholtz algorithm and an alternative algorithm to obtain the same result for Eastman's comma-free code.

Let $Z$ be a Lazard set. Consider the following algorithm starting with a primitive word $w = a_1a_2 \cdots a_m$ and operating on a sequence $s = (s_1, \ldots, s_n)$ of $n$ elements of $Z$, not all equal. Initially, $s = (a_1, a_2, \ldots, a_m)$. The main step transforms $s$ as follows. Since not all $s_i$ are equal, there is an index $i$ with $1 \le i \le n$ such that $s_i < s_{i+1}$ (taking the indices cyclically) with $s_i$ minimal among $s_1, \ldots, s_n$. If $i < n$, change $s$ into $(s_1, \ldots, s_{i-1}, s_is_{i+1}, s_{i+2}, \ldots, s_n)$. If $i = n$, change $s$ into $(s_ns_1, s_2, \ldots, s_{n-1})$. The algorithm stops when $n = 1$.

**Example 28.** We use the algorithm to find the conjugate of *abracadabra* which is in the code $S_1$, using the order $\prec$.

A first sequence of iterations transforms $s = (a, b, r, a, c, a, d, a, b, r, a)$ into $s = (ab, r, ac, ad, ab, r, a)$. At this step, the minimum is the last one and we obtain $s = (aab, r, ac, ad, ab, r)$. The minimum is now $r$ and thus we obtain in two steps $s = (raab, rac, ad, ab)$. The minimum is now $rac$ and we obtain in two steps $s = (raab, racadab)$. The last one being the minimum, we finally obtain $s = (racadabraab)$.

We claim that the result of the algorithm is the conjugate of $w$ which is in $Z$.

Let $Z \cap A^{\le m} = \{z_1, z_2, \ldots, z_k\}$ and let $Z_i$ be the sequence defined, as in the definition of Lazard sets, by $Z_1 = A$ and $Z_{i+1} = z_i^*(Z_i \setminus z_i)$. For $z \in Z$, we denote $\nu(z) = \min\{i \ge 1 \mid z \in Z_i\} - 1$ and $\delta(z) = \max\{i \ge 1 \mid z \in Z_i\}$. Note that $\delta(z_i) = i$ and that $y < z$ if and only if $\delta(y) < \delta(z)$. Then, for $y, z \in Z$, one has $yz \in Z$ if and only if $\nu(z) \le \delta(y) < \delta(z)$. Moreover, in this case, $\nu(yz) = \delta(y)$.

Let us show that any sequence $s = (s_1, \ldots, s_n)$ obtained during the algorithm is such that all $s_i$ are in $Z$ and for any $s_i$, either $s_i \in A$ or $\nu(s_i) \le \delta(s_1), \ldots, \delta(s_n)$. This is true for the initial value of $s$. Next, if we assume that $s$ has this property and is not constant, let $i$ be such that $s_i < s_{i+1}$ and $s_i = \min\{s_1, \ldots, s_n\}$. Then, since $\nu(s_{i+1}) \le \delta(s_i) < \delta(s_{i+1})$, we have $s_is_{i+1} \in Z$ and $\nu(s_is_{i+1}) = \delta(s_i)$. Since $\delta(s_i) = \min\{\delta(s_1), \ldots, \delta(s_n)\}$, we conclude that $\nu(s_is_{i+1}) \le \delta(s_1), \ldots, \delta(s_n)$. For the other $s_j$, we have $\nu(s_j) \le \delta(s_i)$ by induction hypothesis and thus $\nu(s_j) < \delta(s_is_{i+1})$. When the algorithm stops, we obtain a word in $Z$.

## Acknowledgments

We thank Donald Knuth for fruitful exchanges on this subject which led us to write this paper. We also thank Giuseppina Rindone for reading early versions of the manuscript and the referees for their careful reading and suggestions.

## References

[1] Jean Berstel, Dominique Perrin, Christophe Reutenauer, Codes and Automata, in: Encyclopedia of Mathematics and its Applications, vol. 129, Cambridge University Press, Cambridge, 2010.
[2] L. Bokut, E.S. Chibrikov, Lyndon-Shirshov words, Gröbner-Shirshov bases, and free Lie algebras, in: Non-Associative Algebra and its Applications, in: Lect. Notes Pure Appl. Math., vol. 246, Chapman & Hall/CRC, Boca Raton, FL, 2006, pp. 17–39.
[3] Francis H.C. Crick, John S. Griffith, Leslie E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. USA 43 (1957) 416–421.
[4] Williard L. Eastman, On the construction of comma-free codes, IEEE Trans. Inform. Theory IT-11 (1965) 263–267.
[5] Solomon W. Golomb, Basil Gordon, Lloyd R. Welch, Comma free codes, Canad. J. Math. 10 (1958) 202–209.
[6] B.H. Jiggs, Recent results in comma-free codes, Canad. J. Math. 15 (1963) 178–187.
[7] Donald E. Knuth, The Art of Computer Programming, Vol. 4, Addison-Wesley, 2015, pre-fascicule 5B, Introduction to backtracking.

[8] M. Lothaire, Combinatorics on Words, in: Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997, p. xviii+238. With a foreword by Roger Lyndon and a preface by Dominique Perrin, Corrected reprint of the 1983 original, with a new preface by Perrin.
[9] Christophe Reutenauer, Free Lie Algebras, in: London Mathematical Society Monographs. New Series, vol. 7, The Clarendon Press, Oxford University Press, Oxford Science Publications, New York, 1993.
[10] Robert A. Scholtz, Maximal and variable length comma-free codes, IEEE Trans. Inform. Theory IT-15 (1969) 300–306.
[11] Gérard Viennot, Algèbres de Lie Libres et Monoïdes Libres, in: Lecture Notes in Mathematics, vol. 691, Springer, Berlin, 1978, p. ii+124. Bases des algèbres de Lie libres et factorisations des monoïdes libres.