Contents

Acknowledgments

Thanks if you did anything.

Part I

Methods

Part II

Topics

Chapter 1

Enumerative Combinatorics on Words

Dominique Perrin and Antonio Restivo

Université Paris-Est, Marne-la-Vallée and University of Palermo

CONTENTS

Contents

1.1 Introduction

Combinatorics on words is a field which has both historical roots and a substantial growth. Its roots are to be found in the early results of Axel Thue on square free words and the development of combinatorial group theory (see [4] for an introduction to the early developments of combinatorics on words). The present interest in the field is pushed by its links with several connexions with other topics external to pure mathematics, notably bioinformatics.

Enumerative combinatorics on words is itself a branch of enumerative combinatorics, centered on the simplest structure constructor since words are the same as finite sequences.

In this chapter, we have tried to cover a variety of aspects of enumerative combinatorics on words. We have focused on the problems of enumeration connected with conjugacy classes. This includes many interesting combinatorial aspects of words like Lyndon words and de Bruijn cycles. One of the highlights of the chapter is the connexion between both of these concepts via the theorem of Fredericksen and Maiorana.

We have put aside some important aspects of enumerative combinatorics on words which would deserve another complete chapter. This includes the enumeration of various families of words subject to a restriction. For example, the enumeration of square-free words is an important problem for which only asymptotic results are known. It is known for example that the number s_n of ternary square-free words of length n satisfies $\lim_{n\to\infty} s_n^{1/n} = 1.302\ldots$ (see [39] or [16]). Other examples of interest include unbordered words or words avoiding more general patterns (on this notion, see [31]).

The chapter is organized as follows.

In Section 1.2, we introduce some basic definitions concerning words used in the sequel. We also introduce basic notions concerning generating series and automata. Both are powerful tools for the enumeration of words.

In Section 1.3, we introduce the notion of conjugacy and the correlated notions of necklaces or circular codes. These notions play a role in almost all the remaining sections of the chapter. We review some classical formulas such as Witt's Formula or Manning's Formula for the zeta function of a set of words.

In Section 1.4, we introduce Lyndon words and prove the important Factorization Theorem (Theorem 1.4.1). We also discuss the problem of generating Lyndon words and present algorithms for generating them in alphabetic order.

In Section 1.5 we introduce the notion of de Bruijn cycle and their relation with Eulerian graphs. We prove the so-called BEST Theorem enumerating the spanning trees in an Eulerian graph and apply it to the enumeration of de Bruijn cycles. We finally present the Theorem of Fredericksen and Maiorana [17] which beautifully connects Lyndon words and de Bruijn cycles (Theorem 1.5.6).

In Section 1.6, we introduce unavoidable sets. We prove that, on any alphabet, there exist unavoidable sets of words of length n which are a set of representatives of the conjugacy classes of words of length n (Theorem 1.6.1).

In Section 1.7, we introduce a transformation on words, known as the Burrows-Wheeler transformation. This transformation is used in text compression. It is closely related with conjugacy.

We show in Section 1.8 that the Burrows-Wheeler transformation is closely related with a well-known bijection on words, known as the Gessel-Reutenauer bijection. We also prove some results due to Higgins [23] which generalize the theorem of Fredericksen and Maiorana (Theorem 1.8.5).

In Section 1.9, we show that the Burrows-Wheeler is also related to a well-kown concept in string processing, the so-called suffix arrays. We end the section with several results due to Schurman and Stoye [38] concerning the enumeration of suffix arrays

Acknowledgments The authors wish to thank Nicolas Auger, Maxime Crochemore, Francesco Dolce, Gregory Kucherov, Eduardo Moreno, Giovanna Rosone and Christophe Reutenauer who have read the manuscript and made corrections. They also thank the referee who has helped to substantially improve the presentation. The support of ANR project Eqinocs is acknowledged by the first author.

1.2 Preliminaries

We briefly introduce the basic terminology on words. Let A be a finite set usually called the *alphabet*. The elements of A are called *letters*.

A word w on the alphabet A is denoted $w = a_1a_2\cdots a_n$ with $a_i \in A$. The integer n is the length of w. We denote as usual by A^* the set of words over A and by ε the empty word. For a word w, we denote by $|w|$ the length of w. We use the notation $A^+ = A^* - \{\varepsilon\}$. The set A^* is a monoid. Indeed, the concatenation of words is associative, and the empty word is a neutral element for concatenation. The set A^+ is sometimes called the *free semigroup* over A, while A^* is called the *free monoid*.

A word w is called a *factor* (resp. a *prefix*, resp. a *suffix*) of a word u if there exist words x, y such that $u = xwy$ (resp. $u = wy$, resp. $u = xw$). The factor (resp. the prefix,

resp. the suffix) is *proper* if $xy \neq \varepsilon$ (resp. $y \neq \varepsilon$, resp. $x \neq \varepsilon$). The prefix of length k of a word w is also denoted by $w[0..k-1]$.

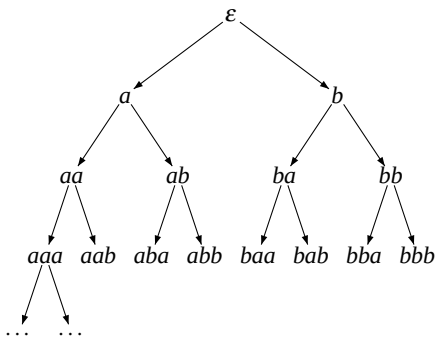

Figure 1.2.1
The tree of the free monoid on two letters.

The set of words over a finite alphabet A can be conveniently seen as a tree. Figure 1.2.1 represents the set $\{a,b\}^*$ as a binary tree. The vertices are the elements of A^*. The root is the empty word ε. The sons of a node x are the words xa for $a \in A$. Every word x can also be viewed as the path leading from the root to the node x. A word x is a prefix of a word y if it is an ancestor in the tree. Given two words x and y, the longest common prefix of x and y is the nearest common ancestor of x and y in the tree.

The set of factors of a word x is denoted $F(x)$. We denote by $F(X)$ the set of factors of words in a set $X \subset A^*$.

The *lexicographic order*, also called *alphabetic order*, is defined as follows. Given two words x,y, we have $x < y$ if x is a proper prefix of y or if there exist factorizations $x = uax'$ and $y = uby'$ with a,b letters and $a < b$. This is the usual order in a dictionary. Note that $x < y$ in the radix order if $|x| < |y|$ or if $|x| = |y|$ and $x < y$ in the lexicographic order.

A *border* of a word w is a nonempty word which is both a prefix and a suffix of w. A word w is *unbordered* if its only border is w itself. For example, a is a border of aba and $aabab$ is unbordered.

1.2.1 Generating series

For a set X of words, we denote by $f_X(z) = \sum_{n\geq 0} \mathrm{Card}(X \cap A^n) z^n$ the *generating series* of X.

Operations on sets can be transferred to their generating series. First, if X,Y are disjoint, then

$$f_{X\cup Y}(z) = f_X(z) + f_Y(z). \tag{1.2.1}$$

Next, the product XY of two sets X,Y is defined by $XY = \{xy \mid x \in X, y \in Y\}$. We say the the product is *unambiguous* if $xy = x'y'$ for $x,x' \in X$ and $y,y' \in Y$ implies $x = x'$ and $y = y'$. Then if the product of X,Y is unambiguous

$$f_{XY}(z) = f_X(z) f_Y(z). \tag{1.2.2}$$

A set $X \subset A^+$ is a *code* if the factorization of a word in words of X is unique. Formally, X is a code if $x_1x_2\cdots x_n = y_1y_2\cdots y_m$ with $x_i,y_j \in X$ and $n,m \geq 1$ implies $n = m$ and $x_i = y_i$ for $1 \leq i \leq n$.

As a particular case, a *prefix code* is a set which does not contain any proper prefix of one of its elements. The submonoid generated by a prefix code X is right unitary, that is to say that $u,uv \in X^*$ implies $v \in X^*$. Conversely, any right unitary submonoid is generated by a prefix code.

If X is a code, then

$$f_{X^*}(z) = \frac{1}{1 - f_X(z)} \tag{1.2.3}$$

In fact, since the sets X^n, X^m are disjoint for $n \neq m$, we have $f_{X^*}(z) = \sum_{n\geq 0} f_{X^n}(z)$. By unique decomposition, we also have $f_{X^n}(z) = (f_X(z))^n$. Thus $f_{X^*}(z) = \sum_{n\geq 0} f_X(z)^n$ whence the result.

Example 1 *Let $X = \{a, ba\}$. The set X is a prefix code. We have* $\mathrm{Card}(X^k \cap A^n) = \binom{k}{n-k}$. *Indeed, a word in $X^k \cap A^n$ is a product of $n-k$ words ba and $2k-n$ words a. It is determined by the choice of the positions of the $n-k$ words ba among k possible ones.*

On the other hand, $\mathrm{Card}(X^* \cap A^n) = F_{n+1}$ *where F_n is the* Fibonacci sequence *defined by $F_0 = 0$, $F_1 = 1$ and $F_{n+1} = F_n + F_{n-1}$ for $n \geq 1$ (the first values are given in Table 1.2.1). This is a consequence of the fact that $f_{X^*}(z) = \frac{1}{1-z-z^2}$ by Equation* (1.2.3)*. Since $f_{X^*}(z) = \sum_{k\geq 0} f_{X^k}(z)$ we obtain the well-known identity relating Fibonacci numbers and binomial coefficients*

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13
F_n	0	1	1	2	3	5	8	13	21	34	55	89	144	233

Table 1.2.1
The first values of the Fibonacci sequence.

$$F_{n+1} = \sum_{k\leq n} \binom{k}{n-k} \tag{1.2.4}$$

which sums binomial coefficients along the parallels to the first diagonal in Pascal's triangle (see Table 1.2.2).

		1						
	⁄	1	1					
1	⁄	1	2	1				
1	⁄	1	3	3	1			
2	⁄	1	4	6	4	1		
3	⁄	1	5	10	10	5	1	
5		1	6	15	20	15	6	1

Table 1.2.2
Pascal's triangle.

Example 2 *The* Dyck set *is the set of words on the alphabet* $\{a,b\}$ *having an equal number of occurrences of a and b. It is a right unitary submonoid and thus it is generated by a prefix code D called the* Dyck code . *Let* D_a *(resp.* D_b*) be the set of words of D beginning with a (resp. b). We have*

$$D_a = aD_a^*b \quad \text{and} \quad D_b = bD_b^*a. \tag{1.2.5}$$

Let us verify the first one. The second one is symmetrical. Clearly any $d \in D_a$ *ends with b. Set* $d = ayb$*. Then y has the same number of occurrences of a and b and thus* $y \in D^*$*. Set* $y = y_1 \cdots y_n$ *with* $y_i \in D$*. If some* y_i *begins with b, then* $ay_1 \cdots y_{i-1}b$ *is a proper prefix of d which belongs to* D^**, a contradiction with the fact that D is a prefix code. Thus all* y_i *are in* D_a *and* $y \in aD_a^*b$*. Conversely, any word in* aD_a^*b *is clearly in* D_a*.*

Since all products in (1.2.5) *are unambiguous, we obtain* $f_{D_a}(z) = z^2 f_{D_a^*}(z)$*. Since* D_a *is a code, by* (1.2.3)*, this implies* $f_{D_a}(z) = z^2/(1 - f_{D_a}(z))$*. We conclude that* $f_{D_a}(z)$ *is the solution of the equation*

$$y(z)^2 - y(z) + z^2 = 0. \tag{1.2.6}$$

such that $y(0) = 0$*. Thus, we obtain the formula*

$$f_{D_a}(z) = \frac{1 - \sqrt{1-4z^2}}{2} \tag{1.2.7}$$

Finally, since $D = D_a \cup D_b$ *and* $f_{D_a}(z) = f_{D_b}(z)$ *for reasons of symmetry, we obtain*

$$f_D(z) = 1 - \sqrt{1-4z^2} \tag{1.2.8}$$

Using the binomial formula, we obtain $\mathrm{Card}(D \cap A^{2n}) = -(-4)^n \binom{1/2}{n}$*. An elementary computation shows that* $\binom{1/2}{n} = (2(-1)^{n-1}/n4^n)\binom{2n-2}{n-1}$*. Thus*

$$\mathrm{Card}(D \cap A^{2n}) = \frac{2}{n}\binom{2n-2}{n-1} \tag{1.2.9}$$

As a consequence, and since $D_a = aD_a^*b$ *by* (1.2.5)*, we obtain the important and well-known fact that*

$$\mathrm{Card}(D_a^* \cap A^{2n}) = \frac{1}{n+1}\binom{2n}{n} \tag{1.2.10}$$

These numbers are called the Catalan numbers *(see Table 1.2.3).*

n	1	2	3	4	5	6	7	8	9	10
	1	1	2	5	14	42	132	429	1430	4862

Table 1.2.3
The first Catalan numbers.

1.2.2 Automata

An *automaton* on the alphabet A is given by a set Q of *states*, a set $E \subset Q \times A \times Q$ of edges, a set I of *initial* states and a set T of *terminal* states. The automaton is denoted $\mathscr{A} = (Q, E, I, T)$ or (Q, I, T) if E is understood.

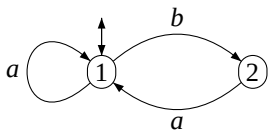

Figure 1.2.2
An automaton

Example 3 *Figure 1.2.2 represents an automaton with two states and three edges. The initial edges are indicated with an incoming edge and the terminal ones with with an outgoing edge. Here state* 1 *is both the unique initial and terminal state.*

A *path* in the automaton is a sequence of consecutive edges (p_i, a_i, p_{i+1}) for $1 \leq i \leq n$. The integer n is the *length* of the path. The word $w = a_1 a_2 \cdots a_n$ is its *label*. We denote $p_1 \xrightarrow{w} p_n$ such a path. A path $i \xrightarrow{w} t$ is *successful* if $i \in I$ and $t \in T$. The set *recognized* by the automaton is the set of labels of successful paths. The automaton is said to be *unambiguous* if for each word w there is at most one successful path labeled w. Thus, an unambiguous automaton defines a bijection between the set of successful paths and the set of their labels. As a particular case, an automaton is *deterministic* if it has at most one initial state and for each state p, at most one edge labeled by a given letter starting at p.

Example 4 *The automaton represented in Figure 1.2.2 recognizes the set* $\{a, ba\}^*$ *of Example 1. It is deterministic and thus unambiguous.*

The *adjacency matrix* of the automaton $\mathscr{A} = (Q, E, I, T)$ is the $Q \times Q$-matrix with integer coefficients defined by

$$M_{p,q} = \text{Card}\{e \in E \mid e = (p, a, q) \text{ for some } a \in A\}.$$

It is clear that for each $n \geq 1$, $M^n_{p,q}$ is the number of paths of length n from p to q. Thus we have the following useful statement.

Proposition 1 *Let $\mathscr{A} = (Q, I, T)$ be an unambiguous automaton, let M be its adjacency matrix and let X be the set recognized by $\mathscr{A}$. For each $n \geq 1$,*

$$\text{Card}(X \cap A^n) = \sum_{i \in I, t \in T} M^n_{i,t}$$

Example 5 *The adjacency matrix of the automaton represented in Figure 1.2.2 is*

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

It is easy to verify that

$$M = \begin{bmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{bmatrix}.$$

Thus, by Proposition 1, we have $\text{Card}(\{a, ba\}^* \cap A^n) = F_{n+1}$*, as already seen in Example 1.*

1.3 Conjugacy

We define necklaces and primitive necklaces. We enumerate first primitive necklaces (Witt's Formula, Proposition 4) and then arbitrary ones (Proposition 6). See [30] for a more detailed presentation. These notions have been extended to more general structures (see in particular the case of partial words in [6]).

1.3.1 Periods

An integer $p \geq 1$ is a *period* of a word $w = a_1 a_2 \cdots a_n$ where $a_i \in A$ if $a_i = a_{i+p}$ for $i = 1, \ldots, n-p$. The smallest period of w is called the *minimal* period of w.

Proposition 2 (Fine, Wilf) *If p, q are periods of a word w of length $\geq p + q - \gcd(p, q)$, then w has period $\gcd(p, q)$.*

Proof. Set $w = a_1 a_2 \cdots a_n$ with $a_i \in A$ and $d = \gcd(p, q)$. We may assume that $p \geq q$. Assume first that $d = 1$. Let us show that $p - q$ is a period of w. Let i be such that $1 \leq i \leq n - p + q$. If $i \leq n - p$, we have $a_i = a_{i+p} = a_{i+p-q}$. Otherwise, we have

$i > n-p$ and thus $i > q-1$. Then $a_i = a_{i-q} = a_{i+p-q}$. Thus w has period $p-q$. Since $\gcd(p,q) = \gcd(p-q,q)$ we obtain by induction on $p+q$ that w has period 1.

In the general case, we consider the alphabet $B = A^d$. On this alphabet w has periods p/d, q/d and length $n/d \geq p/d + q/d$. By the first part, it has period 1 as a word on the alphabet B and thus period d on the alphabet A. ∎

Example 6 *The word $w = abaababaaba$ has periods* 5 *and* 8 *and length* $11 = 5 + 8 - 2$. *By Proposition 2, no word of length* 12 *can have periods* 5 *and* 8 *without having period* 1.

More generally, let x_n be the Fibonacci sequence of words defined by $x_1 = b$, $x_2 = a$ and $x_{n+1} = x_n x_{n-1}$ for $n \geq 2$. For $n \geq 3$, let y_n be the word x_n minus its two last letters. The word y_7 is the word w above. Then, for $n \geq 6$, y_{n+1} has periods F_n and F_{n-1}. Indeed, $y_{n+1} = x_n y_{n-1}$ shows that y_{n+1} has period F_n. Moreover,

$$\begin{aligned} y_{n+1} &= x_n y_{n-1} = x_{n-1}x_{n-2}x_{n-2}y_{n-3} = x_{n-1}x_{n-2}x_{n-3}x_{n-4}y_{n-3} \\ &= x_{n-1}x_{n-1}x_{n-4}y_{n-3} \end{aligned}$$

which shows that F_{n-1} is a period since $x_{n-4}y_{n-3}$ is a prefix of x_{n-3} and thus of x_{n-1}. Since $|y_{n+1}| = F_n + F_{n-1} - 2$, this shows that the bound of Proposition 2 is the best possible.

A word $w \in A^+$ is *primitive* if $w = u^n$ for $u \in A^+$ implies $n = 1$.

Two words x, y are *conjugate* if there exist words u, v such that $x = uv$ and $y = vu$. Thus conjugate words are just cyclic shifts of one another. Conjugacy is thus an equivalence relation. The conjugacy class of a word of length n and period p has p elements if p divides n and has n elements otherwise. In particular, we note the following result.

Proposition 3 *A primitive word of length n has n distinct conjugates.*

1.3.2 Necklaces

A class of conjugacy is often called a *necklace*, represented on a circle (read clockwise, see Figure 1.3.3).

Let $p(n,k)$ be the number of primitive necklaces of length n on k letters. Every word of length n is in a unique way a power of a primitive word of length d with d dividing n and such a word has d distinct conjugates. Thus, for any $n \geq 1$,

$$k^n = \sum_{d|n} d\ p(d,k) \tag{1.3.11}$$

This can be written, using generating series, as a formula called the *Cyclotomic Identity*.

$$\frac{1}{1-kz} = \prod_{n \geq 1} \frac{1}{(1-z^n)^{p(n,k)}}. \tag{1.3.12}$$

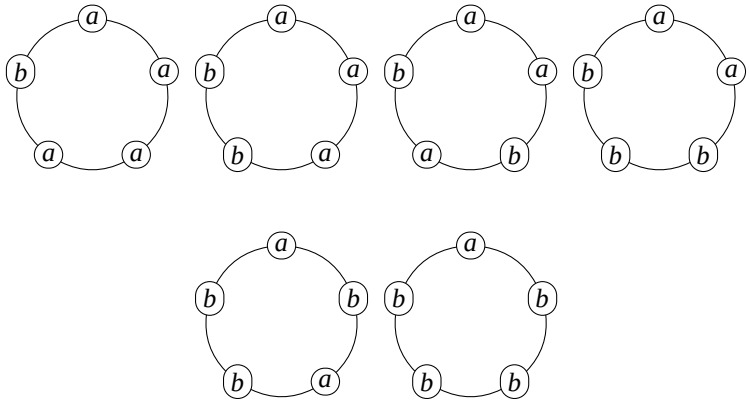

Figure 1.3.3
The six primitive necklaces of length 5 on the alphabet $\{a,b\}$.

Indeed, taking the logarithm of both sides in Equation (1.3.12), we obtain

$$\begin{aligned}\sum_{n\geq 1}\frac{k^n z^n}{n} &= \sum_{n\geq 1} -p(n,k)\log(1-z^n)\\ &= \sum_{n\geq 1} p(n,k)\sum_{m\geq 1}\frac{z^{nm}}{m} = \sum_{n\geq 1}\sum_{n=de} p(d,k)\frac{z^n}{e}\end{aligned}$$

and thus $k^n/n = \sum_{n=de} p(d,k)/e$ whence Formula (1.3.11).

We are going to find a converse giving an expression for the numbers $p(n,k)$. This solution of the system of linear equations (1.3.11) uses the following function.

The *Möbius function* is defined by $\mu(1)=1$ and for $n>1$

$$\mu(n) = \begin{cases}(-1)^i & \text{if } n \text{ is the product of } i \text{ distinct prime numbers}\\ 0 & \text{otherwise}\end{cases}$$

Table 1.3.4 gives the first values of the Möbius function.

n	1	2	3	4	5	6	7	8	9	10
$\mu(n)$	1	-1	-1	0	-1	1	-1	0	0	1

Table 1.3.4
The values of $\mu(n)$ for $n \leq 10$.

Proposition 4 (Witt's Formula) *The number of primitive necklaces of length n on k letters is $p(n,k) = \frac{1}{n}\sum_{d|n}\mu(n/d)k^d$.*

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

Table 1.3.5
The number $p(n,k)$ of primitive necklaces of length n on k letters for $2 \leq k+n \leq 10$.

Table 1.3.5 gives the first values of $p(n,k)$. We prove some properties of the Möbius function before giving the proof of Proposition 4.

Proposition 5 *One has*

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{otherwise} \end{cases}$$

Proof. Indeed, for $n \geq 2$, let $n = p_1^{k_1} \cdots p_m^{k_m}$ and $d = p_1^{\ell_1} \cdots p_m^{\ell_m}$ be the prime decompositions of n, d. Then $\mu(d) \neq 0$ if and only if all ℓ_i are $0, 1$ and then $\mu(d) = (-1)^t$ with $t = \sum_{i=1}^m \ell_i$. Moreover, there are $\binom{m}{t}$ possible choices giving the same sum t. Thus

$$\sum_{d|n} \mu(d) = \sum_{t=0}^{m} (-1)^t \binom{m}{t} = 0$$

since, by the binomial identity, the last expression is $(1-1)^m$. ∎

For two functions α, β from $\mathbb{N} \setminus 0$ into a ring R, their *convolution product* is the function $\alpha * \beta : \mathbb{N} \setminus 0 \to R$ defined by

$$\alpha * \beta(n) = \sum_{de=n} \alpha(d)\beta(e).$$

This product is associative with neutral element the function $\underline{1}$ with value 1 on 1 and 0 elsewhere. By Proposition 5 the function $n \mapsto \sum_{d|n} \mu(d)$ is the function $\underline{1}$. This shows that the Möbius function is the inverse for the convolution product of the constant function equal to 1.

Proof of Proposition 4. Set $\alpha(n) = k^n$ and $\beta(n) = np(n,k)$. Since $k^n = \sum_{d|n} dp(d,k)$ by Equation (1.3.11), we have $\alpha = \beta * \gamma$ where γ is the constant function equal to 1. Since $\gamma * \mu = \underline{1}$, the convolution product of both sides by the Möbius function gives $\alpha * \mu = \beta$, that is $np(n,k) = \sum_{n=de} \mu(d)k^e$. ∎

Recall that Euler's *totient function* φ is defined as follows. The value of $\varphi(n)$ for $n \geq 1$ is the number of integers k with $1 \leq k \leq n$ such that $\gcd(n,k) = 1$. In other words, for $n \geq 2$, $\varphi(n)$ is the number of integers k for $1 \leq k < n$ which are prime to n. One has $n = \sum_{d|n} \varphi(d)$. Indeed, for each divisor d of n the set M_d of integers

n	1	2	3	4	5	6	7	8	9	10
$\varphi(n)$	1	1	2	2	4	2	6	4	6	4

Table 1.3.6
The values of the Euler function $\varphi(n)$ for $n \leq 10$.

$m \leq n$ such that $\gcd(n,m) = d$ has $\varphi(n/d)$ elements. Thus $n = \sum_{d|n} \mathrm{Card}(M_d) = \sum_{d|n} \varphi(n/d) = \sum_{d|n} \varphi(d)$.

Let $c(n,k)$ be the number of necklaces of length n on k letters. Table 1.3.7 gives the first values of the numbers $c(n,k)$. The values in Table 1.3.7 can be easily com-

n	1	2	3	4	5	6	7	8	9
$c(n,1)$	1	1	1	1	1	1	1	1	1
$c(n,2)$	2	3	4	6	8	14	20	36	
$c(n,3)$	3	6	11	24	51	130	315		
$c(n,4)$	4	10	24	70	208	700			
$c(n,5)$	5	15	45	165	481				
$c(n,6)$	6	21	36	216					
$c(n,7)$	7	28	34						
$c(n,8)$	8	36							
$c(n,9)$	9								

Table 1.3.7
The values of the number $c(n,k)$ of necklaces of length n on k letters for $2 \leq k+n \leq 10$.

puted from those of Table 1.3.5 using the fact that $c(n,k) = \sum_{d|n} p(d,k)$. The following statement gives a direct way to compute the numbers $c(n,k)$ (see [21], where it is credited to McMahon).

Proposition 6 $c(n,k) = \frac{1}{n} \sum_{d|n} \varphi(n/d) k^d$.

Proof. Consider the multiset formed by the n circular shifts of the words of length n (each word of length n may appear several times). The total number of the shifts in $nc(n,k)$. On the other hand, each word $w = a_0 \cdots a_{n-1}$ of length n appears with a multiplicity which is the number of integers p with $0 \leq p < n$ such

that $w = a_p \cdots a_{n-1} a_0 \cdots a_{p-1}$, that is which are a period of w^2. But p is a period of w^2 if and only if w is a power of a word of length $\gcd(n,p)$. Thus

$$nc(n,k) = \sum_{0 \le p < n} k^{\gcd(n,p)}. \tag{1.3.13}$$

Since there are $\varphi(n/d)$ integers p with $0 \le p < n$ such that $d = \gcd(n,p)$, the result follows. ∎

We illustrate the proof of Proposition 6 in the following example.

Example 7 *Let $A = \{a,b\}$. The multiset of circular shifts of words of length 4 is the multiset of $6 \times 4 = 24$ elements represented below.*

aaaa	*aaaa*	*aaaa*	*aaaa*
aaab	*aaba*	*abaa*	*baaa*
aabb	*abba*	*bbaa*	*baab*
abab	*baba*	*abab*	*baba*
abbb	*babb*	*bbab*	*bbba*
bbbb	*bbbb*	*bbbb*	*bbbb*

The words appearing more than once are $abab, baba$ which appear twice and $aaaa, bbbb$ which appear 4 times.

The following array gives for each value of $p = 1,2,3$ the set of words w of length 4 such that p is a period of w^2 (for $p = 0$ it is the set of all words of length 4).

p		$\gcd(p,4)$
0	**aaaa**, **aaab**, **aaba**, **aabb**, **abaa**, **abab**, **abba**, **abbb**, **baaa**, **baab**, **baba**, **babb**, **bbaa**, **bbab**, **bbba**, **bbbb**	4
1	**a***aaa*, **b***bbb*	1
2	**aa***aa*, **ab***ab*, **ba***ba*, **bb***bb*	2
3	**a***aaa*, **b***bbb*	1

The value of $d = \gcd(p,4)$ is indicated on the right. The corresponding prefix of length d of each word is indicated in boldface. The row indexed p contains 2^d elements coresponding to the binary words of length d in boldface. In this way we have illustrated Equation 1.3.13 since summing the cardinalities of the sets in each row, we obtain $24 = 16+2+4+2$.

1.3.3 Circular codes

A *circular code* is a set of words X on the alphabet A such that any necklace has a unique factorization in words of X. In particular, a circular code is a code.

Formally, X is a circular code if for $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ in X the equality $sx_2 \cdots x_n p = y_1 \cdots y_m$ with $x_1 = ps$ and s nonempty implies $n = m$, $p = 1$ and $x_i = y_i$ for $1 \le i \le n$.

Example 8 *The set $X = \{a, ba\}$ is a circular code. Indeed, there is at most one way to paste every occurrence of b with the a following it.*

Example 9 *The set $X = \{ab, ba\}$ is not a circular code. Indeed, the necklace of ab has two possible factorizations.*

It can be shown that a submonoid M of A^* is generated by a circular code if and only if it satisfies the following condition for any $u, v \in A^*$.

$$uv, vu \in M \Leftrightarrow u, v \in M. \tag{1.3.14}$$

For a proof, see [5, Chapter 7]. Note that (1.3.14) implies for any $u \in M$ and $n \geq 1$

$$u^n \in M \Leftrightarrow u \in M. \tag{1.3.15}$$

Let S be a set of words on the alphabet A and let $s_n = \mathrm{Card}(S \cap A^n)$ in such a way that $f_S(z) = \sum_{n \geq 0} s_n z^n$.

The *zeta function* of S is the series

$$\zeta_S(z) = \exp \sum_{n \geq 1} \frac{s_n}{n} z^n.$$

The following is due to Manning (see [5, Chapter 7]). The proof uses an argument due to [41].

Theorem 1.3.1 *Let X be a circular code and let S be the set of words having a conjugate in X^*. Then*

$$\zeta_S(z) = \frac{1}{1 - f_X(z)}. \tag{1.3.16}$$

or equivalently

$$f_S(z) = \frac{z f_X'(z)}{1 - f_X(z)}. \tag{1.3.17}$$

Proof. For $x \in X$, denote $g_{n,x}$ the number of words of the form $w = syp$ of length n with $y \in X^*$ and $x = ps$ with p nonempty. Since X is circular, the triple (s, y, p) is uniquely determined by w. Conversely, every word of $S \cap A^n$ is of this form for some $x \in X$. Thus $g_{x,n} = |x| \,\mathrm{Card}(X^* \cap A^{n-|x|})$ and $\mathrm{Card}(S \cap A^n) = \sum_{x \in X} g_{n,x}$. We obtain

$$\begin{aligned} \mathrm{Card}(S \cap A^n) &= \sum_{x \in X} g_{n,x} = \sum_{x \in X} |x| \,\mathrm{Card}(X^* \cap A^{n-|x|}) \\ &= \sum_{m=0}^{n} m \,\mathrm{Card}(X \cap A^m) \,\mathrm{Card}(X^* \cap A^{n-m}). \end{aligned}$$

This shows that $f_S(z) = z f_X'(z) f_{X^*}(z)$ whence Formula (1.3.17). Formula (1.3.16) is obtained from (1.3.17) by taking the derivative of the logarithm of each side. ∎

Let $u_n = \mathrm{Card}(X \cap A^n)$ in such a way that $f_X(z) = \sum_{n\geq 0} u_n z^n$. Using Formula (1.3.17), we obtain for any $n \geq 1$ the formula known as *Newton's Formula* in the context of symmetric functions

$$s_n = nu_n + \sum_{1\leq i\leq n-1} s_i u_{n-i}. \tag{1.3.18}$$

Since from Equation (1.3.17) we have $f_S(z) = \frac{z f'_X(z)}{1-f_X(z)}$, we deduce that $f_S(z) = z f'_X(z) + f_S(z) f_X(z)$, whence Formula (1.3.18).

Let now P be the set of primitive necklaces in S and let $p_n = \mathrm{Card}(P \cap A^n)$. Then since a word of S of length n is a power of a primitive word of length d with d dividing n and that this word has d conjugates, we have the following equality, generalizing Equation (1.3.11)

$$s_n = \sum_{d|n} d\ p_d. \tag{1.3.19}$$

Like Equation (1.3.11), Equation (1.3.19) can be written as an equation relating power series and giving a generalization of the Cyclotomic Identity (1.3.12), namely,

$$f_{X^*}(z) = \prod_{n\geq 1} \frac{1}{(1-z^n)^{p_n}}. \tag{1.3.20}$$

Let c_n be the total number of necklaces in S, primitive or not. A word of length n in S is in a unique way a power of a primitive word of S. Thus $c_n = \sum_{d|n} p_d$ We give below two examples of computation of s_n, p_n, c_n.

Example 10 *Let S be the set of representatives of necklaces on $A = \{a,b\}$ without consecutive occurrences of b. Then S is the set of words having a conjugate in X^* where X is the circular code $X = \{a, ba\}$. Thus, by Theorem 1.3.1, we have*

$$\zeta_S(z) = \frac{1}{1-z-z^2}.$$

By Newton's Formula, since $u_1 = u_2 = 1$ and $u_n = 0$ for $n \geq 3$, we have $s_{n+1} = s_n + s_{n-1}$ for $n \geq 2$.

We obtain the values indicated in Table 1.3.8. The 3 necklaces of length 5 without bb (in agreement with $c_5 = 3$) are represented in Figure 1.3.4.

n	1	2	3	4	5	6	7	8	9	10	11	12	13
s_n	1	3	4	7	11	18	29	47	76	123	199	322	521
p_n	1	1	1	1	2	2	4	5	8	11	18	25	40
c_n	1	2	2	3	3	5	5	8	10	15	19	31	41

Table 1.3.8
The values of s_n, p_n, c_n for $n \leq 13$.

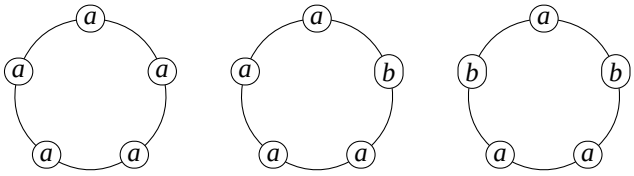

Figure 1.3.4
The 3 necklaces of length 5 on the alphabet $\{a,b\}$ without *bb*.

Example 11 *Let next S be the set of representatives of necklaces on $A = \{a,b\}$ without occurrence of bbb. Then S is the set of words having a conjugate in X^* where X is the circular code $X = \{a, ba, bba\}$. Thus*

$$\zeta_S(z) = \frac{1}{1-z-z^2-z^3}$$

and $s_{n+1} = s_n + s_{n-1} + s_{n-2}$ for $n \geq 3$. We obtain the following values. The 5 *necklaces*

n	1	2	3	4	5	6	7	8	9	10	11	12	13
s_n	1	3	7	11	21	39	71	131	241	443	2757		
p_n	1	1	2	2	4	5	10	15	26	42	74	121	212
c_n	1	2	3	4	5	9	11	19	29	48	75	132	213

Table 1.3.9
The values of s_n, p_n, c_n for the set of necklaces without *bbb*.

of length 5 *without bbb (in agreement with $c_5 = 5$) are represented in Figure 1.3.5.*

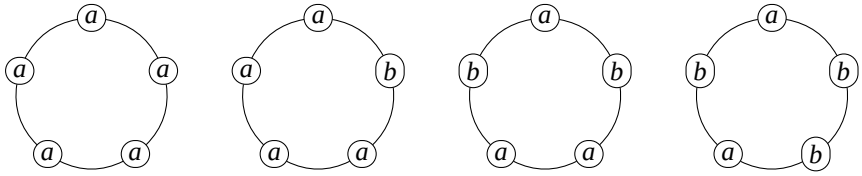

Figure 1.3.5
The 5 necklaces of length 5 on the alphabet $\{a,b\}$ without *bbb*.

The formulae of this section generalize those of the previous one. MacMahon's Iden-

tity (1.3.13) also generalizes to

$$c_n = \frac{1}{n}\sum_{d|n} \varphi(n/d) s_d$$

where φ denotes Euler totient function. This allows a direct computation of the c_n.

1.4 Lyndon words

A *Lyndon word* is a primitive word which is less than all its conjugates in the alphabetic order. We denote by L the set of Lyndon words.

The first Lyndon words on $\{a,b\}$ are

$$a, b$$
$$ab$$
$$aab, abb$$
$$aaab, aabb, abbb$$
$$aaaab, aaabb, aabab, aabbb, ababb, abbbb$$

We first give the following equivalent definition.

Proposition 7 *A word is a Lyndon word if and only if it is strictly smaller than any of its proper suffixes.*

Proof. The condition is sufficient. Indeed, let $w = uv$ with u, v nonempty. Since $w < v$, we have $w < vu$.

It is also necessary. For $w \in L$ let $w = uv$ with u, v nonempty. Assume first that v is a prefix of w and thus that $w = vt$. Since w is a Lyndon word, $w < tv$. But $uv < tv$ implies $u < t$ and thus $vu < vt$, a contradiction. Thus v is not a prefix of w. But then $v < w$ implies that $vu < w$, a contradiction. We conclude that $w < v$. ∎

Note that, as a consequence, a Lyndon word is unbordered. Indeed, if u is both a nonempty suffix and prefix of w, then $u \leq w$ and thus $u = w$ by Proposition 7.

The next statement gives a recursive way to build Lyndon words.

Proposition 8 *If $\ell, m \in L$ with $\ell < m$, then ℓm is a Lyndon word.*

Proof. Let us first show that $\ell m < m$. If ℓ is a prefix of m, then $m = \ell m'$. Then $m < m'$ implies $\ell m < \ell m' = m$. Otherwise, $\ell < m$ implies $\ell m < m$.

Let v be a nonempty proper suffix of ℓm. If v is a suffix of m, then by Proposition 7, $m < v$ and thus $\ell m < m < v$. Otherwise, we have $v = v'm$. Then $\ell < v'$ and thus $\ell m < v'm = v$. By Proposition 7, we conclude that $\ell m \in L$. ∎

For example, we have $aab, ab \in L$ with $aab < ab$ and consequently $aabab \in L$.

1.4.1 The Factorization Theorem

The following result is due to Lyndon (see [30] for more references). It motivated Knuth to call Lyndon words *prime* words in [26].

Theorem 1.4.1 *Any word factorizes uniquely as a nonincreasing product of Lyndon words.*

The proof uses the following result.

Lemma 1 *Let $\ell_1, \ldots, \ell_m$ be a nonincreasing sequence of Lyndon words and let $w = \ell_1 \cdots \ell_m$. Then ℓ_1 is the longest prefix of w which is a Lyndon word and ℓ_m is the minimal nonempty suffix of w.*

Proof. Assume that $\ell \in L$ is a prefix of w longer than ℓ_1. We have $\ell = \ell_1 \cdots \ell_i u$ with $i \geq 1$ and u a nonempty prefix of ℓ_{i+1}. Then $\ell < u \leq \ell_{i+1} \leq \ell_1 < \ell$, a contradiction.

Next, let v be the minimal suffix of w. Then v is in L by Proposition 7. There is an index j, a nonempty suffix s of ℓ_j and a word t such that $v = st$. Then $\ell_m \leq \ell_j \leq s \leq st = v \leq \ell_m$ which implies $v = \ell_m$.

■

Proof of Theorem 1.4.1. We have to show that any word w can be written in a unique way $w = \ell_1 \cdots \ell_m$ with $\ell_1, \ldots, \ell_m \in L$ and $\ell_1 \geq \ldots \geq \ell_m$.

Existence: Since the letters are in L, any word has a factorization in Lyndon words. Consider a factorization $w = \ell_1 \cdots \ell_m$ with m minimal. If $\ell_i < \ell_{i+1}$ for some i, then $w = \ell_1 \cdots \ell_{i-1}(\ell_i \ell_{i+1}) \cdots \ell_m$ is a factorization in Lyndon words since $\ell_i \ell_{i+1} \in L$.

Uniqueness: Assume that $\ell_1 \cdots \ell_m = \ell'_1 \cdots \ell'_{m'}$ with $\ell_i, \ell'_i \in L$, $\ell_1 \geq \ldots \geq \ell_m$ and $\ell'_1 \geq \ldots \geq \ell'_{m'}$. By Lemma 1, we have $\ell_1 = \ell'_1$, which gives the conclusion by induction on m. ■

We illustrate Theorem 1.4.1 by giving below the factorization of the word *abracadabra*.

$$(abracad)(abr)(a)$$

Let P be the set of prefixes of Lyndon words, also called *preprime* words in [26].

We call a word *minimal* if it is minimal for the lexicographic order in its conjugacy class. Clearly, a word is minimal if and only if it is a power of a Lyndon word.

A *sesquipower* of a word x is a word $w = x^n p$ with $n \geq 1$ and p a proper prefix of x. Set $m = |w|$. The word w is determined by x and m. It is called the *m-extension* of x.

The following result appears in Duval [13].

Proposition 9 *The set P is the set of sesquipowers of Lyndon words distinct of the maximal letter.*

The proof uses the following lemma.

Lemma 2 *For any word p and letter a such that pa is a prefix of a minimal word and for any letter b such that $a < b$, the word pb is in L.*

Proof. Let x be a Lyndon word such that pa is a prefix of x^n for some $n \geq 1$. Then $p = x^{n-1}q$ and $x = qar$.

We first show that if $a < b$, then $qb \in L$. Indeed, this is true if q is empty. Otherwise, let t be a proper suffix of q. Then tar is a proper suffix of x. By Proposition 7, this implies $x < tar$ and therefore $q < t$. Thus $pb < tb$. Since any proper suffix of pb is of this form, this shows that $pb \in L$ by Proposition 7 again.

Now, since $x < qb$, we have $x^m qb \in L$ for any $m \geq 1$ by Proposition 8. ∎

Proof of Proposition 9. Let x be a Lyndon word distinct of the maximal letter. Any sesquipower w of x is a prefix of a power x^n of x. By hypothesis, we can write $x = paq$ with a not the maximal letter. Then, by Lemma 2, for any letter $b > a$, we have $x^n pb \in L$ and thus w is in P.

Conversely, we use an induction on the length of $w \in P$. If $|w| = 1$, then $w \in L$. Assume $|w| > 1$. Set $w = va$ with $a \in A$. By induction hypothesis, $v = y^n p$ with $y \in L$, $n \geq 1$ and p proper prefix of y. Set $y = pbu$ with $b \in A$. Since w is a prefix of a Lyndon word, we have $pb \leq pa$ and thus $b \leq a$. If $a = b$, then w is is a sesquipower of y.

Finally if $b < a$, w is a Lyndon word by Lemma 2. ∎

Observe that the Lyndon word x such that w is a sesquipower of x is unique. Indeed, assume that w is a sesquipower of $x, x' \in L$. Assuming that $|x| < |x'|$, we have $x' = x^k p$ with p nonempty prefix of x. Then $p \leq x < x' < p$, a contradiction.

1.4.2 Generating Lyndon words

Proposition 9 can be used to generate Lyndon words of a given length in alphabetic order (this algorithm is due to Fredericksen and Maiorana [17], and independently to Duval [14], see [26]). The idea is to generate all preprime words of this length. This generation problem has been considered in several contexts (see [37], [34] or [26] in particular).

The algorithm SESQUIPOWERS is represented below. We use the alphabet $\{0, \ldots, k-1\}$. This algorithm visits all preprime words $a_1 \cdots a_n$ of length n with an index j such that $a_1 \cdots a_n$ is an extension of $a_1 \cdots a_j$ (we say equivalently that the algorithm visits $a_1 a_2 \cdots a_n$ with index j or that the algorithm visits $a_1 a_2 \cdots a_j$).

```
SESQUIPOWERS(n,k)
 1  for i ← 1 to n do
 2        a_i ← 0
 3  j ← 1
 4  while true do
 5        ▷ Visit a_1 ··· a_n with index j
 6        j ← n
 7        while a_j = k − 1 do
 8              j ← j − 1
 9        if j = 0 then
10              return
11        a_j ← a_j + 1
12        ▷ Now a_1 ··· a_j ∈ L
13        for i ← j + 1 to n do
14              a_i ← a_{i−j}
15              ▷ Make n-extension
```

The assignment at line 11 makes $a_1 \cdots a_j$ a Lyndon word (by Lemma 2). The loop at lines 12-15 realizes the n extension of the word $a_1 \cdots a_j$.

In particular, the sequence of words $a_1 a_2 \cdots a_j$ visited by the algorithm is the sequence of Lyndon words of length at most n in increasing order and the sequence of words $a_1 a_2 \cdots a_n$ visited with index n is the sequence of Lyndon words of length n in increasing order.

We illustrate this on an example. Consider the list in alphabetic order of the words in P of length 5 (we read the list from top to bottom and then from left to right). The letter in boldface is at index j.

a*aaaa*	*aaba***b**	*ab***b***ab*
*aaaa***b**	*aab***b***a*	*abb***b***a*
*aaa***b***a*	*aabb***b**	*abbb***b**
*aaab***b**	*a***b***aba*	**b***bbbb*
*aa***b***aa*	*abab***b**	

The 6 Lyndon words of length 5 are those with the marked letter at the last position.

A possible variant of this algorithm enumerates preprime words in decreasing order.

```
SESQUIPOWERSBIS(n,k)
 1  for i ← 1 to n do
 2      a_i ← k − 1
 3  a_{n+1} ← −1
 4  j ← 1
 5  while true do
 6      ▷ Visit a_1,...,a_n with index j
 7      if a_j = 0 then
 8          return
 9      a_j ← a_j − 1
10      for h ← j + 1 to n do
11          a_h ← k − 1
12      j ← 1
13      h ← 2
14      while a_{h−j} ≤ a_h do
15          ▷ Now a_1 ··· a_{h−1} is the (h − 1)-extension of a_1 ··· a_j
16          if a_{h−j} < a_h then
17              j ← h
18          h ← h + 1
```

At line 8, the assignement realizes the inverse of the operation at line 11 of SESQUIPOWERS. The loop at lines 13-17 implements the computation of the index j such that $a_1 \cdots a_n$ is a sesquipower of $a_1 \cdots a_j$. It is guaranteed to always end by the assignment of line 3.

Recently, Kociumaka, Radoszewski and Rytter have presented a polynomial time algorithm to compute the k-th Lyndon word [27].

1.5 Eulerian graphs and de Bruijn cycles

A *de Bruijn cycle* of order n on k letters is a necklace of length k^n such that every word of length n on k letters appears exactly once as a factor. For example

$$aabb$$
$$aaababbb$$
$$aaaabaabbababbbb$$
$$aaaaabaaabbaabababaabbbbababbabbbbb$$

are de Bruijn cycles of order $2,3,4,5$.

The *de Bruijn graph* of order n on an alphabet A is the following labeled graph. It has A^{n-1} as set of vertices. Its edges are the pairs (u,v) such that $u = aw$, $v = wb$ with $a,b \in A$. Such an edge is labeled b. The de Bruijn graph of orders $3,4$ on the alphabet $\{a,b\}$ are represented in Figure 1.5.6 and Figure 1.5.7. A cycle in a graph is an *Euler*

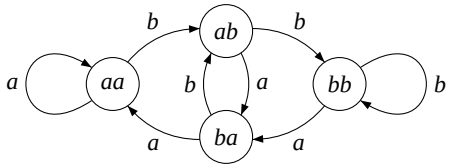

Figure 1.5.6
The de Bruijn graph of order $n = 3$.

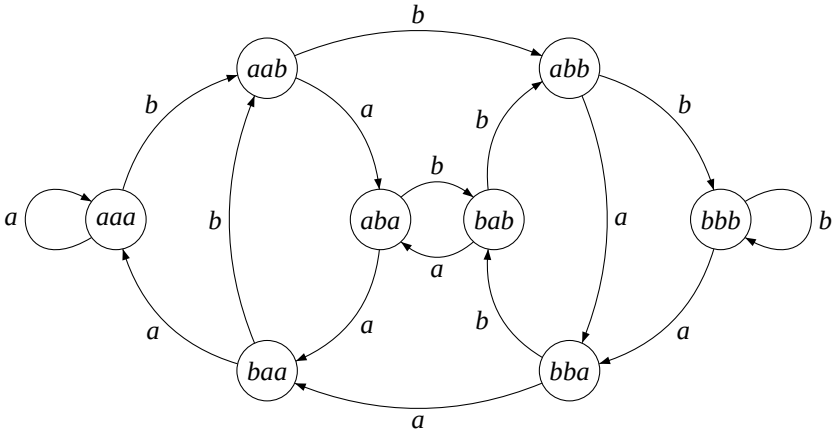

Figure 1.5.7
The de Bruin graph of order $n = 4$

cycle if it uses each edge of the graph exactly once. A finite graph is *Eulerian* if it has an Euler cycle.

It is easy to verify that the de Bruijn cycles of order n are the labels of Euler cycles in the de Bruijn graph of order n. The following result shows the existence of de Bruijn cycles of any order.

Theorem 1.5.1 *A strongly connected finite graph is Eulerian if and only if each vertex has an indegree equal to its outdegree.*

Proof. The condition is necessary since an Euler cycle enters each vertex as many times as it comes out of it.

Conversely, we use an induction on the number of edges of the graph G. If there are no edges, the property is true. Let C be a cycle with the maximal possible number of edges not using twice the same edge. Assume that C is not an Euler cycle. Then, since G is strongly connected, there is a vertex x which is on C and in a non-trivial strongly connected component H of $G \setminus C$. Every vertex of H has an indegree equal to its outdegree. So, by induction hypothesis, H contains an Eulerian cycle D. The cycles C and D have a vertex in common and thus can be combined to form a cycle larger than C, a contradiction. ∎

We denote by $d^-(v)$ the indegree of v (which is the number of edges entering v) and by $d^-(v)$ its outdegree (which is the number of edges coming out of v).

A variant of an Euler cycle is that of *Euler path*. It is a path using all the edges exactly once. It is easy to deduce from Theorem 1.5.1 that a graph has an Euler path from x to y if and only if $d^+(x)-d^-(x)=d^-(y)-d^+(y)=1$ and $d^+(z)=d^-(z)$ for all other vertices.

The computation of an Euler cycle along the lines of the proof of Theorem 1.5.1 is an interesting exercise in recursive programming. It is realized by the following function EULER.

```
EULER(s,t)
 1  if there exists an edge e = (s,x) still unmarked then
 2       MARK(e)
 3       c ← (e, EULER(x,t))
 4       return (EULER(s,s), c)
 5  else return empty
```

The proof of correctness of this algorithm uses the following steps. The function computes an Eulerian path from s (the source) to t (the target). It uses marks on the edges of the graph which are initially all unmarked.

It chooses an edge $e=(s,x)$ leaving s.

If there is an Euler path from s to t beginning with e, the solution is

$$(e, Euler(x,p)).$$

Else the solution is

$$(Euler(s,s), e, Euler(x,p)).$$

The following result is due to van Aarden-Ehrenfest and De Bruijn [1]. We are going to see a derivation of it using linear algebra.

Theorem 1.5.2 *The number of de Bruijn cycles of order n on an alphabet with k letters is*

$$N(n,k)=k^{-n}(k!)^{k^{n-1}}. \tag{1.5.21}$$

In particular, for $k=2$, there are $2^{2^{n-1}-n}$ de Bruijn cycles of order n. Table 1.5.10 lists some values of the numbers $N(n,k)$. The result for $k=2$ was obtained as early as 1894 by Fly Sainte-Marie (see [4] for a historical survey).

Observe that $N(1,k)=(k-1)!$. This is in agreement with the fact that de Bruijn cycles of order 1 are the circular permutations of the k letters.

1.5.1 The BEST Theorem

The following result, known as the BEST Theorem, is due to van Aarden-Ehrenfest and de Bruin [1], and also to Smith and Tutte [40]. For a graph G on a set V of vertices, denote $\pi(G)=\prod_{v\in V}(d^+(v)-1)!$. A spanning tree of G oriented towards a

n	1	2	3	4	5
$N(n,2)$	1	1	2	16	512
$N(n,3)$	2	24	13824		
$N(n,4)$	6	331776			
$N(n,5)$	24				

Table 1.5.10
Some values of the number $N(n,k)$ of de Bruijn cycles of order n on k letters

vertex v is a set of edges T such that, for any vertex w, there is a unique path from w to v using the edges in T.

Theorem 1.5.3 *Let G be an Eulerian graph. Let v be a vertex of G and let $t(G)$ be the number of spanning trees oriented towards v. The number of Euler cycles of G is $t(G)\pi(G)$.*

Proof. Let $\mathscr{E}$ be the set of Euler cycles and let $\mathscr{E}_v$ be the set of Euler paths from vertex v to itself. Since each Euler cycle passes $d^+(v)$ times through v, we have $\mathrm{Card}(\mathscr{E}_v) = d^+(v)\,\mathrm{Card}(\mathscr{E})$.

Let $\mathscr{T}_v$ be the set of spanning trees of G oriented towards v. We define a map $\varphi_v : \mathscr{E}_v \to \mathscr{T}_v$ as follows. Let P be an Euler path from v to v. We define $T = \varphi(P)$ as the set of edges of G used in P to leave a vertex $w \neq v$ for the last time. Let us verify that T is a spanning tree oriented towards v.

Indeed, for each $w \neq v$, there is a unique edge in T going out of w. Continuing in this way, we reach v in a finite number of steps. Thus there is a unique path from w to v.

Conversely, starting from a spanning tree T oriented towards v, we build an Euler path P from v to v as follows. We first use any edge going out of v. Next, from a vertex w, we use any edge previously unused and distinct from the edge in T, as long as such edge exists. There results an Euler path P from v to v which is such that $\varphi(P) = T$. This shows that $\mathrm{Card}(\varphi^{-1}(T)) = d^+(v)!\prod_{w\neq v}(d^+(w)-1)!$. Consequently

$$\mathrm{Card}(\mathscr{E}) = \mathrm{Card}(\mathscr{E}_v)/d^+(v) = t(v)\pi(v).$$

■

We illustrate Theorem 1.5.3 on the example of the de Bruijn graph of order 3 (Figure 1.5.6).

Example 12 *Figure 1.5.8 represents the two possible spanning trees oriented towards bb in the de Bruijn graph of order 3. Following the Eulerian path in the de Bruijn graph of order* 3 *(see Figure 1.5.6), using in turn each of these spanning trees, starting and ending at the root, we obtain the two possible de Bruijn words*

$$aaababbb,\ abaaabbb.$$

Figure 1.5.8
The two spanning trees of de Bruijn graph of order $n = 3$ oriented towards bb.

1.5.2 The Matrix-tree Theorem

Let G be a multigraph on a set V of vertices. Let M be its adjacency matrix defined by $M_{vw} = \mathrm{Card}(E_{vw})$ with E_{vw} the set of edges from v to w. Let D be the diagonal matrix defined by $D_{vv} = \sum_{w \in V} M_{vw}$ and let $L = D - M$ be the *Laplacian matrix* of G. Note that the sum of the elements of each row of L is 0. We denote by $K_v(G)$ the determinant of the matrix C_v obtained by suppressing the row and the column of index v in the matrix L.

The following result is due to Borchardt [8].

Theorem 1.5.4 (Matrix-Tree Theorem) *For any $v \in V$ the number of spanning trees of G oriented towards v is $K_v(G)$*

Proof. Denote by $N_v(G)$ the number of spanning trees oriented towards v.

We use an induction on the number of edges of G. The result holds if there are no edges. Indeed, if there is no edge leading to v, then $N_v(G) = 0$. On the other hand, since the sum of each row of C_v is 0, we have $K_v(G) = 0$. Thus $N_v(G) = K_v(G)$.

Consider now an edge e from w to v. Let G' be the graph obtained by deleting this edge and G'' the graph obtained by merging v and w.

We have

$$N_v(G) = N_v(G') + N_v(G''). \tag{1.5.22}$$

Indeed, the first term of the right hand side counts the number of spanning trees oriented towards v not containing the edge e and the second one the remaining spanning trees. Similarly, we have

$$K_v(G) = K_v(G') + K_v(G''). \tag{1.5.23}$$

Indeed, assume v, w to be the first and second indices. The Laplacian matrices of the graphs G and G'' have the form

$$L = \left[\begin{array}{c|c|c} a & b & x \\ \hline c & d & y \\ \hline z & t & U \end{array}\right], \quad L'' = \left[\begin{array}{c|c} a+b+c+d & x+y \\ \hline z+t & U \end{array}\right].$$

The Laplacian matrix L' of G' being the same as L with $c+1, d-1$ instead of c, d. Then

$$K_v(G) = \left|\begin{array}{c|c} d & y \\ \hline t & U \end{array}\right|, \; K_v(G') = \left|\begin{array}{c|c} d-1 & y \\ \hline t & U \end{array}\right|, \; K_v(G'') = \det(U),$$

and thus Formula (1.5.23) by the linearity of determinants. By induction hypothesis, we have $K_v(G') = N_v(G')$ and $K_v(G'') = N_v(G'')$ By (1.5.22) and (1.5.23) this shows that $K_v(G) = N_v(G)$. ∎

Example 13 *For the graph G of Figure 1.5.6, we have (the matrix C is obtained from L by suppressing the first row and the first column of L).*

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

One has $\det(C) = 2$ *in agreement with Theorem 1.5.4 since, by Example 1.5.8, the graph G has* 2 *spanning trees oriented towards bb.*

It is possible to deduce the explicit formula for the number of de Bruijn cycles of Theorem 1.5.2 from the matrix-tree Theorem.

We denote by G^* the *edge graph* of a graph G. Its set of vertices is the set E of edges of G and its set of edges is the set of pairs $(e, f) \in E \times E$ such that the end of e is the origin of f. It is easy to verify that the edge graph of the de Bruijn graph G_n can be identified with G_{n+1}.

A graph is *regular* of degree k if any vertex has k incoming edges and k outgoing edges. If G is regular, the number $t(G)$ of spanning trees oriented towards a vertex v does not depend on v.

The following result is due to Knuth [24] (see also [25], Exercise 2.3.4.2).

Theorem 1.5.5 *Let G be a regular graph of degree k with m vertices. Then*

$$t(G^*) = k^{m(k-1)-1} t(G).$$

The proof uses the matrix-tree theorem.

It is easy to prove Formula (1.5.21) by induction on n using this result (and the preceding ones). Indeed, by Theorem 1.5.3, and since G_n has k^{n-1} vertices, we have

$$N(n,k) = (k-1)!^{k^{n-1}} t(G_n).$$

Thus (1.5.21) is equivalent to

$$t(G_n) = k^{-n} k^{k^{n-1}}. \tag{1.5.24}$$

Assuming (1.5.24) and using Theorem 1.5.5, we have

$$\begin{aligned} t(G_{n+1}) &= k^{k^{n-1}(k-1)-1}t(G_n) \\ &= k^{k^n-k^{n-1}-1}k^{-n}k^{k^{n-1}} \\ &= k^{-n-1}k^{k^n} \end{aligned}$$

which proves that (1.5.24) holds for $n+1$.

1.5.3 Lyndon words and de Bruijn cycles

The following beautiful result is due to Fredericksen and Maiorana [17].

Theorem 1.5.6 *Let $\ell_1 < \ell_2 < \ldots < \ell_m$ be the increasing sequence of Lyndon words of length dividing n. The word $\ell_1\ell_2\cdots\ell_m$ is a de Bruijn cycle of order n.*

The original statement contains the additional claim that the de Bruijn cycle obtained in this way is lexicographically minimal. We shall obtain this as a consequence of a variant of Theorem 1.5.6 (see Theorem 1.5.7 below).

For example, if $n = 4$ and $A = \{a, b\}$, then

$$aaaabaabbababbbb = a\ aaab\ aabb\ ab\ abbb\ b$$

is a de Bruijn cycle of order 4.

We will use the following lemma.

Lemma 3 *Let w be a prefix of length n of a Lyndon word and let ℓ be its longest prefix in L. Then w is the n-extension of ℓ.*

Proof.

Set $w = \ell s$ and let v be such that $wv \in L$. Set also $r = |\ell|$, $n = |w|$ and $wv = a_1\cdots a_m$ with $a_i \in A$. By Proposition 7, we have $wv < sv$. Thus there is some index t with $1 \le t \le |sv|$ such that $a_j = a_{j+r}$ for $1 \le j \le t-1$ and $a_t < a_{t+r}$. If $t \le n-r$, by Lemma 2, the word $a_1\cdots a_{t+r}$ is a prefix of w which is a Lyndon word longer than ℓ. Thus $a_j = a_{j+r}$ for $1 \le j \le n-r$. This implies that r is a period of w and thus the conclusion. ∎

Proof of Theorem 1.5.6. Since $\ell_1\ell_2\cdots\ell_m$ has length k^n, we only need to prove that any word $w = a_1\cdots a_n$ of length n appears as a factor of $\ell_1\cdots\ell_m\ell_1\ell_2$. We denote by a the first letter of the alphabet and by z the largest one. We consider the following cases.

(a) Assume first that w is primitive and that $w = uv$ with $vu = \ell_k$ and that u is not a power of z. Set $u = pbq$ with $p \in z^*$ and b a letter $b < z$. By Lemma 2, $vpz \in L$. By repeated use of Lemma 8, $vz^{|u|}$ is a Lyndon word. Thus $\ell_{k+1} \le vz^{|u|}$. This implies that v is a prefix of ℓ_{k+1} and thus w is a factor of $\ell_k\ell_{k+1}$.

(b) Assume next that $w = uv$ is primitive, that $u \in z^*$ and that $vu \in L$. We can first rule out the case where $v \in a^*$. Indeed, $z^j a^{n-j}$ is a factor of $\ell_{m-1}\ell_m\ell_1\ell_2$. Let k be the least index such that $v \leq \ell_k$ (the existence of k follows from the fact that $vu = \ell_j$ for some j). Then $\ell_k \leq vu$ and thus v is a prefix of ℓ_k. Let $v' \leq v$ be the Lyndon word such that v is a sesquipower of v'.

(b1) Assume first that $v' \neq \ell_{k-1}$. Let v'' be the word v' with its last letter changed into a. The word visited before v' by Algorithm SESQUIPOWER(n,k) is, in view of Algorithm SESQUIPOWERBIS(n,k), the word $v''z^{n-|v'|}$. Thus ℓ_{k-1} ends with u, ℓ_k begins with v and thus $w = uv$ is a factor of $\ell_{k-1}\ell_k$.

(b2) Otherwise, $v' = \ell_{k-1}$. For the same reason as above, u is a suffix of ℓ_{k-2}. Since v is a sesquipower of v', it is a prefix of $v'v$ and thus also a prefix of $\ell_{k-1}\ell_k$. Thus w is a factor of $\ell_{k-2}\ell_{k-1}\ell_k$.

(c) Assume finally that $w = (uv)^d$ with d dividing n and $vu = \ell_k$.

(c1) If $u \notin z^*$ then $\ell_{k+1} \leq (vu)^{d-1}vz^{|u|}$ since the latter is a Lyndon word. Thus w is a factor of $\ell_k\ell_{k+1}$.

(c2) Otherwise, ℓ_{k-1} ends with at least $(d-1)|w|$ letters z and $\ell_{k+1} \leq (vu)^{d-1}z^{|w|}$. Thus w is a factor of $\ell_{k-1}\ell_k\ell_{k+1}$.

■

We illustrate the cases in the proof for $n = 6$ and $A = \{a,b\}$. Table 1.5.11 gives the sequence ℓ_k.

k	1	2	3	4	5	6	7	8
ℓ_k	a	$aaaaab$	$aaaabb$	$aaabab$	$aaabbb$	aab	$aababb$	$aabbab$

k	9	10	11	12	13	14
ℓ_k	$aabbbb$	ab	$ababbb$	abb	$abbbbb$	b

Table 1.5.11
The Lyndon words of length dividing 6.

(a) Let $w = aabaaa$. Then $u = aab$, $v = aaa$ and $vu = \ell_2$. We find w as a factor of $\ell_2\ell_3$.

(b1) Let $w = baaaab$. Then $u = b$ and $v = aaaab$. We find $k = 3$, $v' = v$ and w is a factor of $\ell_2\ell_3$.

(b2) Let $w = bbabab$. Then $u = bb$, $v = abab$. We find $k = 11$. We have $v' = ab$ and we find w as a factor of $\ell_9\ell_{10}\ell_{11}$.

(c1) Let $w = (aba)^2$. Then $u = a$, $v = ba$ and $k = 6$. We find w as a factor of $\ell_6\ell_7$.

(c2) Let $w = (bab)^2$. Then $u = b$, $v = ab$ and $k = 12$. We find w as a factor of $\ell_{11}\ell_{12}\ell_{13}$.

Let X be a set of words. A de Bruijn cycle of order n relative to X is a necklace such that every word of X of length n appears exactly once as a factor. The usual notion of de Bruijn cycle is relative to $X = A^*$.

Consider for example the set X of words on $\{a,b\}$ which are representatives of necklaces without consecutive occurrences of b (see Example 10). Then *aaab* is a de Bruijn cycle of order 3 relative to X and *aaaabab* of order 4.

The following result, due to Moreno [34], gives a family of sets X for which there are de Bruijn cycles of any order relative to X. Let $\ell_1 < \ell_2 < \ldots < \ell_m$ be the increasing sequence of Lyndon words of length dividing n. For $s < m$, we denote by X_s the set of words such that no factor has a conjugate in $\{\ell_1, \ldots, \ell_s\}$.

Theorem 1.5.7 *For any $s < m$, the sequence $\ell_s\ell_{s+1}\cdots\ell_m$ is a de Bruijn cycle of order n relative to X_s.*

One can deduce from this result the fact that the de Bruijn cycle given by Theorem 1.5.6 is the minimal one for the alphabetic order (see [35]).

As another variant of Theorem 1.5.6, let us quote the following result due to Yu Hin Au [2]: concatenating the Lyndon words of length n in increasing order, one obtains a word which contains cyclically all primitive words of length n exactly once. For example, for $n = 4$ and $A = \{a,b\}$, one obtains the word *aaab aabb abbb* which contains cyclically all 12 twelve primitive words of length 4.

1.6 Unavoidable sets

A word t is said to *avoid* a word p if p is not a factor of t, i.e. if the pattern p does not appear in the text t. For example the word `abracadabra` avoids `baba`. The set of all words avoiding a given set X of words has been of interest in several contexts including the notion of a system of finite type in symbolic dynamics (see [29] for example). This notion has been extended to many other situations (see in particular the case of partial words in [7]).

Let A be a finite alphabet. An *unavoidable* set on A is a set $I \subset A^*$ of words on the alphabet A such that any two-sided infinite word $(a_n)_{n\in Z}$ on the alphabet A admits at least one factor in I. It is of course equivalent to ask that any one-sided infinite word has a factor in I or also, since the alphabet is finite, that the set of words that avoids I is finite (see [31] for an exposition of the properties of unavoidable sets).

Example 14 *Let $A = \{a,b\}$. The set $U = \{a, b^{10}\}$ is unavoidable since any word of length 10 either has a letter equal to a or is the word b^{10}. On the contrary, the set*

$V = \{aa, b^{10}\}$ is avoidable. Indeed, the infinite word $(ab)^\omega = abababab\ldots$ has no factor in V.

Proposition 10 *On a finite alphabet, any unavoidable set contains a finite unavoidable set.*

Proof. Indeed, let X be an unavoidable set and let S be the set of words avoiding X. Since X is unavoidable, S is finite. Let n be the maximal length of the words of S. Let Z be the set of words of X of length at most $n+1$. Every word of length $n+1$ has a factor in X which is actually in Z. Thus Z is unavoidable. ∎

The following gives an equivalent definition of unavoidable sets which holds for finite sets. It will be used below.

Proposition 11 *Let $I \subset A^*$ be a finite set of words. The following conditions are equivalent.*

(i) *The set I is unavoidable.*

(ii) *Each two-sided infinite periodic word has at least one factor in I.*

Proof. It is enough to show that (ii) $\Rightarrow$ (i). Let $(a_n)_{n\in Z}$ be a two-sided infinite sequence of letters. Let $u \in A^*$ be a word longer than any word in I and having an infinite number of occurrences in the sequence $(a_n)_{n\in\mathscr{Z}}$. This sequence has at least one factor of the form uvu. By the hypothesis, the infinite periodic word $\ldots uvuvuvuv\ldots$ has a factor $w \in I$. The word w is a factor of at least one of the words uv or vu. It is thus also a factor of the sequence $(a_n)_{n\in\mathscr{Z}}$ and thus I is unavoidable. ∎

This statement is false if I is infinite. For example, on a three-letter alphabet, the set of squares is avoidable but every periodic word contains obviously a square.

1.6.1 Algorithms

To check in practice that a given finite set X is unavoidable, there are two possible algorithms.

The first one consists in computing a graph $G = (P,E)$, where P is the set of prefixes of X and E is the set of pairs (p,s) for which there is a letter $a \in A$ such that s is the longest suffix of pa which is in P.

Proposition 12 *A finite set X is unavoidable if and only if every cycle in G contains a vertex in X.*

Proof. For each integer $n \geq 0$, and vertices $u, v \in P$, there is a path of length n from u to v if and only if there exists a word y of length n such that v is the longest suffix of uy in P. This can be proved by induction on n. It follows that there is a path of length n from ε to a vertex $x \in X$ if and only if $AX \cap A^n \neq \emptyset$. ∎

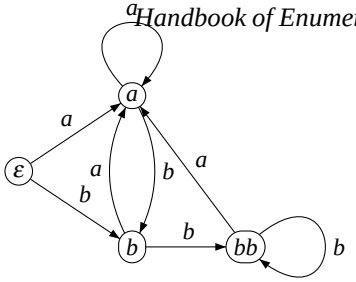

Figure 1.6.9
The graph for $X = \{a, bb\}$.

Example 15 *For $X = \{a, bb\}$, the word graph is given in Figure 1.6.9. By inspection, the set X is unavoidable.*

The second algorithm is sometimes easier to write down by hand. Say that a set Y of words is obtained from a finite set of words X by an *elementary derivation* if

(i) either there exist words $u, v \in X$ such that u is a proper factor of v, and

$$Y = X \setminus v$$

(ii) or there exists a word $x = ya \in X$ with $a \in A$ such that, for each letter $b \in A$ there is a suffix z of y such that $zb \in X$, and

$$Y = (X \cup y) \setminus x$$

A *derivation* is a sequence of elementary derivations. We say that Y is derived from X if Y is obtained from X by a derivation.

Example 16 *Let $X = \{aaa, b\}$. Then we have the derivations*

$$X \to \{aa, b\} \to \{a, b\} \to \{\varepsilon, b\} \to \{\varepsilon\}$$

where the first three arrows follow case (ii) and the last one case (i).

The following result shows in particular that if Y is derived from X, then X is unavoidable if and only if Y is unavoidable. We denote by S_X the set of two-sided infinite words avoiding X.

Proposition 13 *If Y is derived from X, then $S_X = S_Y$.*

Proof. It is enough to consider the case of an elementary derivation. In the first case where $Y = X \setminus v$, where v has a factor in X, then clearly $S_X = S_Y$. In the second case, we clearly have $S_Y \subset S_X$ since Y is obtained by replacing an element of X by one of its factors. Conversely, assume by contradiction the existence of some $s \in S_X \setminus S_Y$. The only possible factor of s in Y is y. Let b be the letter following y in s. Then s has

a factor in X, namely zb where z is the suffix of y such that $zb \in X$ whose existence is granted by the definition of a derivation. This is a contradiction. ∎

The notion of a derivation gives a practical method to check whether a set is unavoidable. We have indeed the following result.

Proposition 14 *A finite set X is unavoidable if and only if there is a derivation from X to the set $\{\varepsilon\}$.*

Proof. Let $X \neq \{\varepsilon\}$ be unavoidable. We prove the existence of a derivation to $\{\varepsilon\}$ by induction on the sum $l(X)$ of the lengths of words in X. If $\varepsilon \in X$, we may derive $\{\varepsilon\}$ from X. Thus assume $\varepsilon \notin X$, and let w be a word of maximal length avoiding X. For each $b \in A$ there is a word $x_b = zb \in X$ which is a suffix of wb. Let $x_a = ya$ be the longest of the words x_b. Then the hypotheses of case (ii) are satisfied and thus there is a derivation from X to a set Y with $l(Y) < l(X)$. The converse is clear by Proposition 13. ∎

In practice, there is a shortcut which is useful to perform derivations. It is described in the following transformation from X to Y.

(iii) there is a word y such that $ya \in X$ for each $a \in A$ and

$$Y = (X \cup y) \setminus \{ya \mid a \in A\}$$

It is clear that such a set Y can be derived from X and thus, we do not change the definition of derivations by adding case (iii) to the definition of elementary derivations. We use this new definition in the following example.

Example 17 *Let $X = \{aaa, aba, abb, bbb\}$. We have the following sequence of derivations (with the symbol a in the word $x = ya$ underlined at each step)*

$$\begin{aligned} \{aaa, ab\underline{a}, ab\underline{b}, bbb\} &\to \{aa\underline{a}, ab, bbb\} \\ &\to \{a\underline{a}, a\underline{b}, bbb\} \to \{a, bb\underline{b}\} \to \{a, b\underline{b}\} \\ &\to \{\underline{a}, \underline{b}\} \to \{\varepsilon\} \end{aligned}$$

Derivations could of course be performed on the left rather than on the right.

1.6.2 Unavoidable sets of constant length

In the sequel, we will be interested in unavoidable sets made of words having all the same length n. The following proposition is easy to prove.

Proposition 15 *Let A be a finite alphabet and let I be an unavoidable set of words of length n on A. The cardinality of I is at least equal to the number of conjugacy classes of words of length n on the alphabet A.*

Proof. Let $u \in A^*$ be a word of length n. The factors of length n of the word u^ω are the elements of the conjugacy class of u. Thus I must contain at least one element of this class. ∎

We are going to prove the following result which shows that the lower bound $c(n,k)$ on the size of unavoidable sets of words of length n on k symbols is reached for all $n,k \geq 1$.

Theorem 1.6.1 *For all $n,k \geq 1$, there exists an unavoidable set formed of $c(n,k)$ words of length n on k symbols.*

This result has been obtained by J. Mykkelveit [36], solving a conjecture of Golomb. His proof uses exponential sums (see below). Later, and independently, it was solved by Champarnaud, Hansel and Perrin [11] using Lyndon words. We shall present this proof here.

It may be convenient for the reader to reformulate the statement in terms of graphs. A *feedback vertex set* in a directed graph G is a set F of vertices containing at least one vertex from every cycle in G. Consider, for $n \geq 1$, the de Bruijn graph G_{n+1} of order $n+1$ on the alphabet A whose vertices are the words of length n on A and the edges are the pairs (au,ub) for all $a,b \in A$ and $u \in A^{n-1}$. It is easy to see that a set of words of length n is unavoidable if the corresponding set of vertices is a feedback vertex set of the graph G_{n+1}. Thus, the problem of determining an unavoidable set of words of length k of minimal size is the same as determining the minimal size of a feedback vertex set in G_{n+1}. The problem is, for general directed graphs, known to be NP-complete (see [18] for example).

As a preparation to a proof of Theorem 1.6.1, we introduce the following notions.

A *division* of a word w is a pair (ℓ^i,u) such that $w = \ell^i u$ where $\ell \in L$, $i \geq 1$ and $u \in A^*$ with $|u| < |\ell|$.

By Proposition 9 each word in P admits at least one division. We say that a Lyndon word $\ell \in L$ *meets* the word w if there is a division of w of the form (ℓ^i,u). It is clear that for any $\ell \in L$ there is at most one such division of w.

The *main division* of $w \in P$ is the division (ℓ^i,u) where ℓ is the shortest Lyndon word which meets w. The word ℓ^i is the *principal part* of w, denoted by $p(w)$, and u is the *remainder*, denoted by $r(w)$.

For example, with $a < b$, the word *aabaabbba* admits two divisions which are $(aabaabbb,a)$ and $(aabaabb,ba)$. The first one corresponds to its decomposition as a sesquipower of a Lyndon word. The second one is its main division.

Let $n \geq 1$ be an integer and let M_n be the set of minimal words of length n. For each $m \in M_n$, let $p(m)$ be its principal part and $r(m)$ its remainder. Let I_n be the set

$$I_n = \{r(m)p(m) | m \in M_n\}$$

We remark that any minimal word which is not primitive appears in I_k.

Example 18 *Table 1.6.12 lists the elements of M_7 and I_7 with the remainder of each word of M_7 in boldface.*

M_7	I_7		
aaaaaaa	*aaaaaaa*	*aabab***ab**	**ab***aabab*
aaaaaab	*aaaaaab*	*aabab***bb**	**bb***aabab*
*aaaaab***b**	**b***aaaaab*	*aabb***abb**	**abb***aabb*
*aaaab***ab**	**ab***aaaab*	*aabb***bab**	**bab***aabb*
*aaaab***bb**	**bb***aaaab*	*aabb***bbb**	**bbb***aabb*
*aaab***aab**	**aab***aaab*	*ababab***b**	**b***ababab*
*aaab***abb**	**abb***aaab*	*ababb***bb**	**bb***ababb*
*aaab***bab**	**bab***aaab*	*abbabb***b**	**b***abbabb*
*aaab***bbb**	**bbb***aaab*	*abbb***bbb**	**bbb***abbb*
*aabaab***b**	**b***aabaab*	*bbbbbbb*	*bbbbbbb*

Table 1.6.12
The sets M_7 and I_7

The object of what follows is to show that I_n is an unavoidable set. By Proposition 15, the number of elements of I_n is the minimal possible number of elements of an unavoidable set of words of length n.

Theorem 1.6.1 will be obtained consequence of the following one, giving a construction of the minimal unavoidable sets.

Theorem 1.6.2 *Let A be a finite alphabet and let $n \geq 1$. Let M_n be the set of words on the alphabet A of length n and which are minimal in their conjugacy class. For every word $m \in M_n$, let $p(m)$ be the principal part of m and let $r(m)$ be its remainder. Then the set*

$$I_n = \{r(m)p(m) | m \in M_n\}$$

is an unavoidable set.

To prove Theorem 1.6.2, we need some preliminary results.

Proposition 16 *Let ℓ and m be two Lyndon words, with ℓ a prefix of m. Let $s \in A^*$ be a proper suffix of m, with $|s| < |\ell|$. Then for all $i > 0$, the word $w = \ell^i s$ is a Lyndon word.*

Proof. Let t be a proper suffix of w. Three cases may arise.
1. One has $|t| \leq |s|$. Then t is a proper suffix of the Lyndon word m and thus $t > m \geq \ell$ and since $|t| < |\ell|$, we have $t > \ell^n s = w$.
2. One has $|t| > |s|$ and the word t factorizes as $t = \ell^j s$, with $0 \leq j < i$. Since s is a proper suffix of m, we have $s > m \geq \ell$. Consequently $t = \ell^j s > \ell^{j+1}$ and since $|s| < |\ell|$, we have $t > \ell^i s = w$.
3. One has $|t| > |s|$ and the word t factorizes as $t = s't'$, where s' is a proper suffix of ℓ. Since $\ell \in L$, one has $s' > \ell$, and consequently $t = s't' > \ell^i s = w$.

In all cases $t > w$ and thus w is a Lyndon word. ∎

Proposition 17 *Let w be a prefix of a minimal word and let (ℓ^i, u) be its main division. Let $u' \in A^*$ be a word of the same length as u such that the word $w' = \ell^i u'$ is also a prefix of a minimal word. Then the main division of w' is the pair (ℓ^i, u').*

Proof. Let (m^j, v) be the main division of w'. We have $w' = m^j v$ with $|v| < |m|$. Since (ℓ^i, u') is a division of w', the word m is a prefix of ℓ. We are going to show by contradiction that m cannot be a proper prefix of ℓ.

Suppose that m is a proper prefix of ℓ. Since the factorization of a minimal word as a power of a Lyndon word is unique, we cannot have the equality $m^j = \ell^i$. Suppose first that $|m^j| < |\ell^i|$. Since $w' = m^j v = \ell^i u'$, the word m^j is a proper prefix of the word ℓ^i. Thus there exists a non-empty word $x \in A^*$ such that $m^j x = \ell^i$ and $xu' = v$. We thus have

$$w = \ell^i u = m^j xu$$

Since $|xu| = |xu'| = |v| < |m|$, the pair (m^j, xu) is a division of w, which is a contradiction since m is a proper prefix of ℓ and that (ℓ^i, u) is the main division of w.

Let us now suppose that $|m^j| > |\ell^i|$. Since $w' = m^j v = \ell^i u'$, the word ℓ^i is a proper prefix of the word m^j. Since m is a proper prefix of ℓ, there exists an integer $k > 0$ and a prefix m' of m such that $\ell = m^k m'$. Since ℓ is a primitive word, m' is non-empty. As a consequence, ℓ admits m' both as a non-empty prefix and and suffix, which is contradictory since ℓ is a Lyndon word. ∎

The final property needed to prove Theorem 1.6.2 is the following.

Proposition 18 *Let m be a Lyndon word and n a positive integer. Let $N \geq 1$ be the smallest integer such that $|m^N| > n$. Then the word m^{N+1} has a factor in I_n.*

Proof. Let w be the prefix of length n of m^N. Let (ℓ^i, u) be the main division of w. If u is the empty word, then, by construction, $w \in I_n$ and the proposition is true. Suppose that u is not empty.

The word ℓ is a prefix of m since either $|w| < |m|$ or w admits a division of the form (m^j, m'). Let s be the suffix of m having the same length as u. By Proposition 16, the word $\ell^i s$ is a Lyndon word. Thus, by Proposition 17, the main division of $\ell^i s$ is the pair (ℓ^i, s). Consequently, the word $s\ell^i$ belongs to I_n. But this word is a factor of m^{N+1}. Thus m^{N+1} has a factor in I_n. ∎

We are now able to prove Theorem 1.6.2. By Proposition 11, it is enough to show that every periodic two-sided infinite word of the form $\ldots uuuuu \ldots$ has at least one factor in I_n. We may suppose without loss of generality that u is a Lyndon word. Let N be the least integer such that $N|u| > n$. Then, by Proposition 18, the word u^{N+1} has a factor in I_n. Thus I_n is unavoidable.

1.6.3 Conclusion

The proof of J. Mykkeltveit in [36] is based on the following principle, presented in the case of a binary alphabet. Let us associate to a word $w = a_0 a_1 \cdots a_{n-1}$ on the alphabet $\{0,1\}$ the sum $s(w) = \sum a_j \omega^j$ where $\omega = e^{2i\pi/n}$. We denote by $Is(w)$ the

imaginary part of $s(w)$. It can be shown that for each conjugacy class of words, only two cases occur:

(i) either all words w are such that $Is(w) = 0$ (and then, for $n > 2$ one has actually $s(w) = 0$ for each of them)

(ii) or there is, in clockwise order, one block of words w such that $Is(w) > 0$ followed by one block of words w such that $Is(w) < 0$ separated by at most two words w such that $Is(w) = 0$.

Consider the set S_n of words of length n formed of

(i) a representative of each conjugacy class of words w of length n such that $Is(w) = 0$ for all the conjugates.

(ii) the words $w = a_0a_1\cdots a_{n-1}$ of length n such that $Is(w) > 0$ for the first time clockwise.

It is shown in [36] that this set is unavoidable for all $n > 2$. The comparison with the previous family of minimal unavoidable set shows that the families have nothing in common. The sets obtained are indeed different. The sets defined by J. Mykkeltveit have a slight advantage in the sense that the maximal length of words avoiding the set is less. For example, for $n = 20$, there are 256 words of length 2579 that avoid I_n, but none of length 563 that avoid all of S_n (and there is a unique way to avoid S_n with length 562). This computation has been performed using D. Knuth's program UNAVOIDABLE2 (see `http://www-cs-faculty.stanford.edu/~knuth/programs.html`). Our proof has the advantage of using only elementary concepts and in particular no real or complex arithmetic.

Another proof of Theorem 1.6.1 obtained by the first two authors of [11] and presented in [10] is a construction working by stages. To explain these stages, let us consider the case of a binary alphabet $A = \{a,b\}$. Given a set X of two-sided infinite words, we say that a set Y of words is unavoidable in X if every word of X has a factor in Y.

For $i \geq 1$, let X_i be the set of two-sided infinite words on A which avoid a^i. Let $c_i(n,k)$ be the number of conjugacy classes of words x of length n on k symbols such that the words of the form $x^\zeta = \cdots xxx\cdots$ are in X_i. It is thus also equal to the number of orbits of period n in X_i. Table 1.6.13 below gives the values of $c_i(n,2)$ for $1 \leq i \leq 10$ and $1 \leq n \leq 10$. The rows are indexed by i and the columns by n. Thus the second row is the last row of Table 1.3.8 and the third row is the last row of Table 1.3.9. Moreover, substracting 1 to the first 8 entries of the 3 last rows, we obtain the second row of Table 1.3.7 (we have to substract 1 because a^n is missing).

The idea of the step by step construction of a minimal unavoidable set of words of length n is to construct a sequence $Y_1 \subset Y_2 \subset \ldots \subset Y_n$ of sets of words of length n such that for $1 \leq i \leq n$, the set Y_n is unavoidable in X_i with $c_i(n,k)$ elements. This can be stated as the following result.

Theorem 1.6.3 *For each $k \geq 1$, and $n \geq i \geq 1$, there exists a set of $c_i(n,k)$ words of length n on k symbols which is unavoidable in X_i.*

	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	3	3	5	5	8	10	15
3	1	2	3	4	5	9	11	19	29	48
4	1	2	3	5	6	11	15	27	43	59
5	1	2	3	5	7	12	17	31	51	91
6	1	2	3	5	7	13	18	33	55	99
7	1	2	3	5	7	13	19	34	57	103
8	1	2	3	5	7	13	19	35	58	105
9	1	2	3	5	7	13	19	35	59	106
10	1	2	3	5	7	13	19	35	59	107

Table 1.6.13
The values of $c_i(n,2)$.

The alternative proof consists in showing directly that the $c_i(n,k)$ last elements of I_n form a set unavoidable in X_i.

It is interesting to remark that not all minimal unavoidable sets are build in this way. Indeed, there are sets which are minimal unavoidable in X_{n+1} but do not contain a minimal unavoidable set in X_n.

For example, let $Y_1 = \{bbb\}$, $Y_2 = \{bbb, bab\}$, $Y_3 = \{bbb, bab, aab\}$. Then each Y_i for $1 \leq i \leq 3$ is unavoidable in X_i of size $c_i(3,2)$ and $I_3 = Y_3 \cup \{aaa\}$. In particular, the set I_3 contains an unavoidable set in X_2 with 2 elements, namely Y_2. However, the set $J_3 = \{aaa, aba, bba, bbb\}$ obtained from I_3 by exchanging a and b does not contain a two element set unavoidable in X_2.

A set of the form X_i is a particular case of what is called a system of finite type. This is, by definition the set of all two-sided infinite words avoiding a given finite set of words (see [29]). We do not know in general in which systems of finite type it is true that for each n there exists an unavoidable set having no more elements than the number of orbits of period n.

1.7 The Burrows-Wheeler Transform

The Burrows-Wheeler transform is a popular method used for text compression [9]. It produces a permutation of the characters of an input word w in order to obtain a word easier to compress. The presentation given here is close to that of [12].

Suppose w is a *primitive* word over a *totally ordered* alphabet A. Let w_1,w_2,...,w_n be the sequence of conjugates of w in increasing lexicographic order. Let $M(w)$ be the matrix having w_1,w_2,...,w_n as rows. For example, if $w = aabacacb$, the matrix $M(w)$ is

$$M(aabacacb) = \begin{bmatrix} a & a & b & a & c & a & c & b \\ a & b & a & c & a & c & b & a \\ a & c & a & c & b & a & a & b \\ a & c & b & a & a & b & a & c \\ b & a & a & b & a & c & a & c \\ b & a & c & a & c & b & a & a \\ c & a & c & b & a & a & b & a \\ c & b & a & a & b & a & c & a \end{bmatrix}$$

The *Burrows-Wheeler Transform* $T(w)$ of w is the last column of $M(w)$, read from top to bottom. If b_i denotes the last letter of the word w_i, for $i = 1,2,...,n$, then $T(w) = b_1 b_2 ... b_n$. For instance, $T(aabacacb) = babccaaa$.

It is clear that $T(w)$ depends only on the conjugacy class of w, i.e. $T(w) = T(w')$ if w and w' are conjugate. Therefore we may suppose that w is a *Lyndon* word, i.e. $w = w_1$.

The matrix $M(w)$ defines a permutation σ_w (or simply σ when no confusion arises) of $1,2,...,n$:

$$\sigma(i) = j \Longleftrightarrow w_j = a_i \cdots a_n a_1 \cdots a_{i-1} \tag{1.7.25}$$

In other terms, $\sigma(i)$ is the rank in the lexicographic order of the $i-th$ circular shift of the word w. For instance, for $w = aabacacb$, we have:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 6 & 3 & 7 & 4 & 8 & 5 \end{pmatrix}.$$

Let $F(w)$ denote the first column of the matrix $M(w)$. If c_i denotes the first letter of the word w_i, for $i = 1,2,\ldots,n$, then $F(w) = c_1 c_2 \cdots c_n$ is the nondecreasing rearrangement of w. By definition, we have, for each index i, with $1 \leq i \leq n$,

$$a_i = c_{\sigma(i)} \tag{1.7.26}$$

The permutation σ transforms the first column of $M(w)$ into its first row, i.e. into the word w. We have also the following formula expressing $T(w)$ using σ:

$$b_i = a_{\sigma^{-1}(i)-1}. \tag{1.7.27}$$

Indeed, $b_{\sigma(i)}$ is the last letter of $w_{\sigma(i)} = a_j \cdots a_n a_1 \cdots a_{j-1}$, hence $b_{\sigma(j)} = a_{j-1}$, which is equivalent to the above formula.

Given a primitive word $w \in A^*$, let π_w, or simply π when no confusion arises, be the permutation defined by

$$\pi(i) = \sigma(\sigma^{-1}(i) + 1), \tag{1.7.28}$$

where the addition is to be taken modulo n.

Remark 1.7.1 *Observe that π is just the permutation defined by writing σ as a word and interpreting it as a n-cycle. Thus we have also $\sigma(i) = \pi^{i-1}(1)$ and*

$$a_i = c_{\pi^{i-1}(1)}.$$

In the previous example we have, written as a cycle,

$$\pi = (1\,2\,6\,3\,7\,4\,8\,5)$$

and as an array

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 6 & 7 & 8 & 1 & 3 & 4 & 5 \end{pmatrix}.$$

The following proposition is fundamental for defining the inverse transform.

Proposition 19 *If $c_1c_2\cdots c_n$ and $b_1b_2...b_n$ are the first and the last columns, respectively, of the matrix $M(w)$, then $c_i = b_{\pi(i)}$, for $i = 1,2,\ldots,n$.*

Proof. Substituting in formula 1.7.27 the value a_i given by formula 1.7.26, we obtain $b_i = c_{\sigma(\sigma^{-1}(i)-1)}$, which is equivalent to the statement of the proposition. ∎

The previous proposition states that the permutation π_w transforms the last column of the matrix $M(w)$ into the first one. Actually, it can be noted that π_w transforms any column of the matrix $M(w)$ into the following one.

1.7.1 The inverse transform

We now show how the word w can be recovered from $T(w)$. For this we prove a property of the matrix $M(w)$ stating that, for any letter $a \in A$, its occurrences in $F(w)$ appear in the same order as in $T(w)$, i.e. the $k-th$ instance of a in $T(w)$ corresponds (through π) to its $k-th$ instance in $F(w)$. In order to formalize this property we introduce the following notation.

The *rank* of the index i in the word $z = z_1z_2\cdots z_n$, denoted by $rank(i,z)$, is the number of occurrences of the letter z_i in $z_1z_2\cdots z_i$. For instance, if $z = babccaaa$, then $rank(4,z) = 1$ and $rank(6,z) = 2$.

Proposition 20 *Given the words $T(w) = b_1b_2\cdots b_n$ and $F(w) = c_1c_2\cdots c_n$, for each index $i = 1,2,\ldots,n$, we have*

$$rank(i,F(w)) = rank(\pi(i),T(w)).$$

Proof. We first note that, for two words u,v of the same length, and for any letter $a \in A$, one has

$$au < av \Longleftrightarrow ua < va.$$

Thus, for all indices i, j, $i < j$ and $c_i = c_j$ implies $\pi(i) < \pi(j)$. Hence, the number of

occurrences of c_i in $c_1c_2\cdots c_i$ is equal to the number of occurrences of $b_{\pi(i)} = c_i$ in $b_1b_2\cdots b_{\pi(i)}$. ∎

To obtain w from $T(w) = b_1b_2...b_n$, we first compute $F(w) = c_1c_2...c_n$ by rearranging the letters b_i in nondecreasing order. Proposition 20 shows that $\pi(i)$ is the index j such that $c_i = b_j$ and $rank(j, T(w)) = rank(i, F(w))$. This defines the permutation π, from which σ can be obtained expressing π as a n-cycle, and then, by using Formula (1.7.26), the word w can be reconstructed.

Remark 1.7.2 *Proposition 20 further shows that the permutation π is related to the* standard *permutation of the word $T(w)$. Recall that the standard permutation of a word $v = b_1b_2\cdots b_n$ on a totally ordered alphabet A is the permutation τ such that, for $i, j \in 1,2,\ldots,n$, the condition $\tau(i) < \tau(j)$ is equivalent to $b_i < b_j$ or $b_i = b_j$ and $i < j$. The permutation τ may be obtained by numbering from left to right the letters of v, starting from the smallest letter, then the second smallest, and so on. For example, for $v = babccaaa$, we have that, written as a word, $\tau = 51678234$, and as an array:*

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 1 & 6 & 7 & 8 & 2 & 3 & 4 \end{pmatrix}.$$

It is easy to see that the permutation π corresponds to the inverse of the standard permutation of $T(w)$.

Remark 1.7.3 *The Burrows-Wheeler transform $T(w)$ of a primitive word w depends only on the conjugacy class of w. Therefore, T defines an* injective *mapping from the primitive necklaces over an alphabet A to the words of A^*. However such a mapping is not* surjective*. Remark 1.7.2 indeed shows that, if we consider a word $u \in A^*$ such that the standard permutation τ of u (and then also the permutation $\pi = \tau^{-1}$) is not a cycle, then does not exists any word w such the $T(w) = u$. Let us, for instance, consider the word $u = bccaaab$. Its standard permutation*

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 6 & 7 & 1 & 2 & 3 & 5 \end{pmatrix}$$

is the product of two cycles

$$\tau = (1\ 4)(2\ 6\ 3\ 7\ 5).$$

It follows that there does not exist any word w such that $T(w) = u$.

1.7.2 Descents of a permutation

A *descent* of a permutation π is an index i such that $\pi(i) > \pi(i+1)$. We denote by $Des(\pi)$ the set of descents of the permutation π. Consider the permutation π_w corresponding to a word w. It is clear from Proposition 20 that if i is a descent of

π_w, then $c_i \neq c_{i+1}$. Thus the number of descents of π_w is at most equal to $k-1$, where k is the number of distinct symbols appearing in the word w. For instance, for $w = aabacacb$, $\pi_w(4) > \pi_w(5)$, moreover 4 is the only descent of π_w and so $Des(\pi_w) = \{4\}$.

Let $A = \{a_1, a_2, \ldots, a_k\}$ be a totally ordered alphabet with $a_1 < a_2 < \ldots < a_k$. If w is a word of A^*, denote by $P(w)$ the *Parikh vector* of w: $P(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k})$. It is clear that, if two words are conjugate, then they have the same Parikh vector, and so one can define the Parikh vector of a necklace. We say that a vector $V = (n_1, n_2, ..., n_k)$ is *positive* if $n_i > 0$ for $i = 1, 2, ..., k$. We denote by $\rho(V)$ the set of integers $\rho(V) = \{n_1, n_1 + n_2, ..., n_1 + ... + n_{k-1}\}$. When V is positive, $\rho(V)$ has $k-1$ elements. Let π_w be the permutation corresponding to word w and let $P(w)$ be the Parikh vector of w. It is clear from Proposition 20 that we have the inclusion $Des(\pi_w) \subset \rho(P(w))$.

Example 19 *The Parikh vector of the word $w = aabacacb$ is $V = (4,2,2)$ and $\rho(V) = \{4,6\}$. The permutation π_w corresponding to w is*

$$\pi_w = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 6 & 7 & 8 & 1 & 3 & 4 & 5 \end{pmatrix}.$$

Thus $Des(\pi_w) = \{4\} \subset \rho(V)$.

The following statement, due to Crochemore, Désarménien and Perrin [12], results from the previous considerations and Remark 1.7.2.

Theorem 1.7.4 *For any positive vector $V = (n_1, n_2, ..., n_k)$, with $n = n_1 + n_2 + ... + n_k$, the map $w \to \pi_w$ is one-to-one from the set of primitive necklaces of length n with Parikh vector V onto the set of cyclic permutations π on $\{1, 2, ..., n\}$ such that $\rho(V)$ contains $Des(\pi)$.*

This result actually is a particular case of a result stated in [31] and closely related to the Gessel-Reutenauer bijection introduced in the next section. Since each conjugacy class of primitive words can be represented by a Lyndon word, Theorem 1.7.4 establishes a bijection between Lyndon words and cyclic permutations having special descent sets. The extension of this result (Theorem 11.6.1 of [31]) establishes a bijection between words and permutations, relating the Lyndon factorization of words and the cycle structure of the permutations.

1.8 The Gessel-Reutenauer bijection

We have shown that the Burrows-Wheeler transform $T(w)$ of a word w depends only on the conjugacy class of w. Therefore, T defines an injective mapping from the

primitive necklaces over an alphabet A to the words of A^*. However (cf. Remark 1.7.3) such a mapping is not surjective.

We now extend the Burrows-Wheeler transform by defining a *bijective* mapping Φ from the *multisets* of primitive necklaces over a totally ordered alphabet A to the words of A^*. This bijection has been introduced by Gessel and Reutenauer in [20]. In order to define the mapping Φ we introduce a new order on A^*. For a word $z \in A^*$, z^ω denotes the infinite word $zzz\cdots$ obtained by infinitely iterating z. Given $u, v \in A^*$, $u \preccurlyeq_\omega v$ if and only if $u^\omega \leq v^\omega$ in the lexicographic order. Remark that the order $\preccurlyeq_\omega$ is different from the usual lexicographic order: for instance, $aba \preccurlyeq_\omega ab$.

Following [33] (see also [23]), we give here a presentation of the mapping Φ that emphasizes its relation with the Burrows-Wheeler transform.

Let $S = \{s_1, s_2, \ldots, s_m\}$ be a multiset of necklaces, represented by their Lyndon words, i.e. s_k is the Lyndon word corresponding to the k-th necklace of S. In some cases, it is convenient to denote by $(z_1 z_2 \cdots z_t)$ the necklace containing the word $z_1 z_2 \cdots z_t$. Denote also by $n = |s_1| + |s_2| + \ldots + |s_m|$ the *size* of S and by L the least common multiple of the lengths of the words in S.

In order to sort the elements in the necklaces of S according to the order $\preccurlyeq_\omega$, consider the collection of all words of the form $u^{L\backslash|u|}$, where u is an element of a necklace. All these words then have a common length L. We order this set of words lexicographically to yield a matrix $M(S)$ with n rows and L columns.

Example 20 *Let* $S = \{aab, ab, abb\}$. *Then* $n = 8$ *and* $L = 6$ *and*

$$M(S) = \begin{bmatrix} a & a & b & a & a & b \\ a & b & a & a & b & a \\ a & b & a & b & a & b \\ a & b & b & a & b & b \\ b & a & a & b & a & a \\ b & a & b & a & b & a \\ b & a & b & b & a & b \\ b & b & a & b & b & a \end{bmatrix}$$

The transform $\Phi(S) = b_1 b_2 \cdots b_n$ corresponds to the last column of the matrix $M(S)$, read from top to bottom. In the previous example, $\Phi(S) = babbaaba$.

It is clear that if the multiset S has only one necklace represented by its Lyndon word w, then $\Phi(S) = T(w)$, i.e. $\Phi(S)$ corresponds to the Burrows-Wheeler transform of w. Note also that, if there are non trivial multiplicities in the multiset S, then there are repeated rows in the matrix $M(S)$.

Several properties of the Burrows-Wheeler matrix $M(w)$ of a word w can be easily extended to the matrix $M(S)$. In particular, if we denote by $F(S) = c_1 c_2 \cdots c_n$ the first column of the matrix $M(S)$, by using the same argument as in the proof of Proposition 20, it can be shown that, for any letter $a \in A$, its occurrences in $F(S)$ appear in the same order as in $\Phi(S)$. This defines a permutation π that transforms the

last column of the matrix $M(S)$ into the first one. Actually (cf. Remark 1.7.2), π is the inverse of the standard permutation of the word $\Phi(S)$.

Example 21 *The inverse of the standard permutation of the word babbaaba is*

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 5 & 6 & 8 & 1 & 3 & 4 & 7 \end{pmatrix}.$$

In order to reverse the Burrows-Wheeler transform, given the word $T(w) = b_1 b_2 \cdots b_n$, we considered in Section 1.7.1 the inverse π of its standard permutation, then we expressed it as a n-cycle $(j_1, j_2, \ldots, j_n)$ and we associated to this n-cycle the necklace $(c_{j_1} c_{j_2} \cdots c_{j_n})$.

Recall that the Burrows-Wheeler transform is injective, but not surjective. This is a consequence of the fact that, for some words u, the inverse of its standard permutation cannot be expressed by a single n-cycle, but its decomposition contains several cycles (cf. Remark 1.7.3). This remark is at the base of the surjectivity of the Gessel-Reutenauer transform.

Now we show how to reverse the transform Φ, that is how the multiset of necklaces S can be recovered from the word $\Phi(S) = b_1 b_2 \cdots b_n$.

As for the Burrows-Wheeler transform, first compute the first column $F(S) = c_1 c_2 \cdots c_n$ of $M(S)$ by rearranging the letters b_i in nondecreasing order.

Then, consider the inverse π of the standard permutation associated to the word $\Phi(S) = b_1 b_2 \cdots b_n$. With each cycle $(j_1, j_2, \ldots, j_i)$ of π, associate the necklace $(c_{j_1} c_{j_2} \cdots c_{j_i})$. The multiset S is given by

$$S = \{(c_{j_1} c_{j_2} \cdots c_{j_i}) | (j_1, j_2, \ldots, j_i) \text{ is a cycle of } \pi\}.$$

Remark that different cycles of π could give rise to the same necklace, and this explains the use of multisets.

Example 22 *Let $\Phi(S) = babbaaba$ (see Example 20). Rearranging the letters in nondecreasing order, one obtains $F(S) = aaaabbbb$. Then the permutation π is*

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 5 & 6 & 8 & 1 & 3 & 4 & 7 \end{pmatrix}.$$

By decomposing π in cycles

$$\pi = (1\ 2\ 5)(3\ 6)(4\ 8\ 7),$$

one obtains the multiset of necklaces $S = \{(aab), (ab), (abb)\}$.

The following theorem, due to Gessel and Reutenauer [20], results from the preceding considerations.

Theorem 1.8.1 *The map Φ defines a bijection between words over a totally ordered alphabet A and multisets of primitive necklaces over A.*

A similar, but different, bijection has been proved in [19], where, instead of the lexicographic order, is used the *alternate lexicographic* order.

1.8.1 Gessel-Reutenauer bijection and de Bruijn cycles

In this section we present an interesting connection, pointed out in [23], between the Gessel-Reutenauer bijection and the de Bruijn cycles.

A multiset $S = \{s_1, s_2, \ldots, s_m\}$ of necklaces is a *de Bruijn set of span n* over an alphabet A if $|s_1| + |s_2| + \ldots + |s_m| = \mathrm{Card}(A)^n$ and every word $w \in A^n$ is a prefix of some power of some word in a necklace of S.

Remark 1.8.2 *The number of distinct prefixes of length n of powers of the words in the necklaces of S is at most* $\mathrm{Card}(A)^n$. *So, given that S is a de Bruijn set of span n, every word in* A^n *can be read exactly once within the necklaces of S. It also follows, in particular, that no two necklaces in S are equal, so that S is indeed a set, as opposed to a multiset, of necklaces.*

Remark 1.8.3 *If S is a de Bruijn set of span n, then S contains a necklace of length at least n. To show this, consider a Lyndon word u of length n (for instance,* $u = ab^{n-1}$, *where* $a < b$*). By definition, u is prefix of some power of a word in a necklace of S. Since u, as a Lyndon word, is unbordered, it cannot arise as a prefix of a proper power in a necklace of S. It follows that S contains a necklace of length at least n.*

If A is an alphabet of cardinality k, denote by Γ the set of all $k!$ products of distinct elements of A:

$$\Gamma = \{a_1 a_2 \cdots a_k | a_i \in A \text{ for } i = 1, \ldots, k \text{ and } a_i \neq a_j \text{ for } i \neq j\}.$$

For instance, for $A = \{a, b, c\}$,

$$\Gamma = \{abc, acb, bac, bca, cab, cba\}.$$

The following result is due to Higgins [23].

Theorem 1.8.4 *A set S is a de Bruijn set of span n if and only if* $\Phi(S) \in \Gamma^{k^{n-1}}$.

Proof. Let us first suppose that S is a de Bruijn set of span n. Consider the matrix $M(S)$. By Remark 1.8.3, the length L of the rows of $M(S)$ is at least n. Consider the sub-matrix consisting of the first n columns of $M(S)$. Since S is a de Bruijn set, the rows of this sub-matrix form the set A^n. Each word $u \in A^{n-1}$ is prefix of k successive rows of $M(S)$. We show that these successive rows of $M(S)$ end with distinct letters of A. Suppose, by contradiction, that two of these rows v_1 and v_2, end with the same letter a, i.e. $v_1 = uxa$ and $v_2 = uya$ for some $x, y \in A^*$, with $x \neq y$. Since the conjugates aux and auy, of v_1 and v_2, respectively, correspond to distinct rows in $M(S)$, it follows that $au \in A^n$ would be a prefix of a power of distinct words in the necklaces of S, contrary to S being a de Bruijn set of span n. Hence the final column $\Phi(S)$ of $M(S)$ is a product of k^{n-1} elements (possibly with repetitions) taken from the set Γ.

In order to prove the converse implication, let S be a multiset of necklaces such that $\Phi(S) = w \in \Gamma^{k^{n-1}}$. We first prove, by induction on the integer r, with $1 \leq r \leq n$,

that any word $u \in A^*$ of length r is the prefix of k^{n-r} consecutive rows of the matrix $M(S)$. In particular, we show that there exists an integer j such that u appears as a prefix in the rows of $M(S)$ ranging from the index jk^{n-r} to the index $(j+1)k^{n-r}-1$. Remark that the sequence of the last letters of these rows, read from top to bottom, returns a factor of w which is again a concatenation of elements of Γ.

The statement is true for $r = 1$. Indeed, since $w \in \Gamma^{k^{n-1}}$, $|w| = k^n$ and, for any letter $a \in A$, $|w|_a = k^{n-1}$. It follows that the first column $F(S)$ of $M(S)$, read from top to bottom, consists of k^{n-1} occurrences of the first (in the order) letter of A, followed by k^{n-1} occurrences of the second letter, and so on. Actually, we have also that, if z is the word corresponding to an arbitrary column of $M(S)$, for each $a \in A$, $|z|_a = |z|/k$.

Let us now suppose that the statement is true for some $r < n$, and consider a word $v \in A^*$ of length $r+1$. If a is the first letter of v, we have $v = au$, with $|u| = r$. By the inductive hypothesis, there exists an integer j such that u is the prefix of length r of k^{n-r} consecutive rows of $M(S)$ ranging from the index jk^{n-r} to the index $(j+1)k^{n-r}-1$. The sequence of the last letters of these rows, read from top to bottom, forms a factor z_u of w (the word corresponding to the last column of $M(S)$), and moreover z_u is product of elements of Γ. Thus, for any $a \in A$, $|z_u|_a = k^{n-r-1}$. It follows that, within the k^{n-r} consecutive rows of $M(S)$ having u as prefix, k^{n-r-1} of them end with the letter a. By taking into account their conjugates, we have that k^{n-r-1} consecutive rows of $M(S)$ have as prefix the same word $au = v$. If b is the last letter of v, i.e. $v = u'b$, since $|u'| = r$, by the inductive hypothesis there exists an integer i such that u' appears as prefix of the rows of $M(S)$ ranging from the index ik^{n-r} to the index $(i+1)k^{n-r}-1$. The k different letters of A split the interval $[ik^{n-r}, (i+1)k^{n-r}-1]$ into k sub-intervals of equal length in such a way that each sub-interval contains the rows of $M(S)$ having as prefix of length $r+1$ the word $u'c$, for some $c \in A$. We conclude that there is an integer t such that the k^{n-r-1} consecutive rows of $M(S)$, having as prefix the word $v = u'b$, have indexes that range from tk^{n-r-1} to $(t+1)k^{n-r-1}-1$. So, we have proved that, if $\Phi(S) = w \in \Gamma^{k^{n-1}}$, then, for any r, with $1 \leq r \leq n$, every word $u \in A^*$ of length r is the prefix of k^{n-r} consecutive rows of $M(S)$. In particular, for $r = n$, every word $u \in A^*$ of length n is the prefix of exactly one row of $M(S)$. This implies that S is a de Bruijn set of span n.

■

By Theorem 1.8.4, one can generate a de Bruijn set S of span n, on an alphabet A of cardinality k, by taking a word $v \in \Gamma^{k^{n-1}}$ and by computing $\Phi^{-1}(v)$.

Example 23 *Consider the alphabet $A = \{a, b\}$ with $a < b$. Then $\Gamma = \{\alpha, \beta\}$, where $\alpha = ab$ and $\beta = ba$. Let $n = 4$, and consider the word $v = \beta\alpha\beta\beta\alpha\alpha\alpha\beta = baabbabaababbabba \in \Gamma^8$. Rearranging the letters of v in nondecreasing order, one obtains the first column $F(S)$ of the matrix $M(S)$: $F(S) = aaaaaaaabbbbbbbb$. The inverse π of the standard permutation of the word v is*

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 2 & 3 & 6 & 8 & 9 & 11 & 13 & 16 & 1 & 4 & 5 & 7 & 10 & 12 & 14 & 15 \end{pmatrix}.$$

By decomposing π in cycles

$$\pi = (1\ 2\ 3\ 6\ 11\ 5\ 9)(4\ 8\ 16\ 15\ 14\ 12\ 7\ 13\ 10),$$

one obtains the set of necklaces

$$S = \{(baaaaba), (baabbbbab)\}.$$

One can verify that any word of A^4 is prefix of some word in a necklace of S, i.e. S is a de Bruijn set of span 4.

Given a totally ordered alphabet $A = \{a_1, a_2, \ldots, a_k\}$, of cardinality k, with $a_1 < a_2 < \ldots < a_k$, denote by α the element $a_1 a_2 \cdots a_k \in \Gamma$. Now we look at the special case of Theorem 1.8.4 where v is a power of α. In such a case, by specializing the arguments in the proof of Theorem 1.8.4, (cf.[23]), one can prove the following result.

Theorem 1.8.5 *Let $v = \alpha^{k^{n-1}}$, let $S = \Phi^{-1}(v)$ and let $M = M(S)$ be the matrix corresponding to S. Then the rows of M are simply the elements of A^n. Moreover S is the set of necklaces of the Lyndon words of length dividing n.*

Example 24 *Consider the alphabet $A = \{a, b\}$ with $a < b$, and the word $\alpha^{2^4} = (ab)^{16}$. The inverse π of the standard permutation of the word $(ab)^{16}$ is*

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ 1 & 3 & 5 & 7 & 9 & 11 & 13 & 15 & 2 & 4 & 6 & 8 & 10 & 12 & 14 & 16 \end{pmatrix}.$$

By decomposing π in cycles

$$\pi = (1)(2\ 3\ 5\ 9)(4\ 7\ 13\ 10)(6\ 11)(8\ 15\ 14\ 12)(16),$$

one obtains the set of necklaces

$$S = \{(a), (aaab), (aabb), (ab), (abbb), (b)\},$$

which is the set of necklaces of the Lyndon words of length dividing 4*. If we consider the concatenation of such Lyndon words, we obtain the word*

$$a.aaab.aabb.ab.abbb.b$$

which is indeed the first de Bruijn word of span 4 *in the lexicographic order. That this is always the case is the well known theorem of Frederickson and Maiorana (see Theorem 1.5.6).*

Actually, as a consequence of Theorem 1.8.5 and of the theorem of Frederickson and Maiorana, we obtain the following result.

Proposition 21 *The concatenation in ascending order of the Lyndon words of the necklaces of $S = \Phi^{-1}(\alpha^{k^{n-1}})$ is the first de Bruijn word of span n in the lexicographic order.*

1.9 Suffix arrays

Suffix array is a widely used data structure in string algorithms (see [32] or [22]). The *suffix array* of a word w of length n is essentially a permutation of $\{1,2,...,n\}$ corresponding to the starting positions of all the suffixes of w sorted lexicographically.

Let $A = \{a_1, a_2, \ldots, a_k\}$ be a totally ordered alphabet of size k, where $a_1 < a_2 < \ldots < a_k$. Given a word $w = z_1 z_2 \cdots z_n$ of length n on the alphabet A, the suffix array of w is the permutation ϑ_w (or simply ϑ when no confusion arises) of the set $\{1,2,\ldots,n\}$ such that $\vartheta(i) = j$ if the suffix $z_j z_{j+1} ... z_n$ has rank i in the lexicographic ordering of all the suffixes of w. For instance, if $w = baaababa$, then

$$\vartheta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 2 & 3 & 6 & 4 & 7 & 1 & 5 \end{pmatrix}.$$

1.9.1 Suffix arrays and Burrows-Wheeler transform

We show here, following [28], the close connection between the suffix array of a word and its Burrows-Wheeler transform. For this purpose, it is convenient to introduce the *Burrows-Wheeler array* (or simply the *BW-array*) of a primitive word of A^*.

Given a primitive word $w = z_1 z_2 \cdots z_n$ of length n on the ordered alphabet A, the *BW*-array of w is the permutation φ_w (or simply φ when no confusion arises) of $\{1,2,\ldots,n\}$ such that $\varphi(i) = j$ if the conjugate $z_j \cdots z_n z_1 \cdots z_{j-1}$ has rank i in the lexicographic sorting of all the conjugates of w. By definition, the *BW*-array of a word w is just the inverse of the permutation σ defined by the relation 1.7.25. i.e. $\varphi = \sigma^{-1}$.

In order to show the connection between the suffix array and the Burrows-Wheeler transform, we first introduce a *sentinel* symbol at the end of the word. Consider a symbol $\sharp \notin A$, and the ordered alphabet $A' = \{\sharp, a_1, ..., a_k\}$ where $\sharp < a_1 < \ldots < a_k$. We will examine the suffix array of the word $w' = w\sharp$. In the sequel, we denote by S_n be the set of permutations of $\{1,2,\ldots,n\}$. Moreover, for $\vartheta \in S_n$, $\tilde{\vartheta} \in S_{n+1}$ denotes the permutation

$$\tilde{\vartheta} = \begin{pmatrix} 1 & 2 & \ldots & n+1 \\ n+1 & \vartheta(1) & \ldots & \vartheta(n) \end{pmatrix}.$$

Remark 1.9.1 *There is a one-to-one correspondence between the suffix arrays of the words $w \in A^n$ and the suffix arrays of the words $w' \in A^n\sharp$. In particular, if the permutation $\vartheta_w \in S_n$ is the suffix array of $w \in A^n$, then the permutation $\vartheta_{w'} \in S_{n+1}$ is the suffix array of $w' = w\sharp$ if and only if $\vartheta_{w'} = \tilde{\vartheta}$.*

Remark 1.9.2 *It is easy to see that, for words in $A^*\sharp$, conjugate sorting is equivalent*

to suffix sorting. It follows that the suffix array of the word $w' = w\sharp$ coincides with its BW-array, i.e. $\vartheta_{w'} = \varphi_{w'}$.

The following statement follows from the previous remarks.

Proposition 22 *A permutation $\vartheta \in S_n$ is the suffix array of a word $w \in A^n$ if and only if the permutation $\tilde{\vartheta} \in S_{n+1}$ is the BW-array of the word $w' = w\sharp$.*

Consider now the mapping $\Psi : S_n \to S_{n+1}$ defined as follows. If $\vartheta \in S_n$, $\Psi(\vartheta)$ is the permutation $\mu \in S_{n+1}$ defined by $\mu(i) = \tilde{\vartheta}^{-1}(\tilde{\vartheta}(i)+1)$, where the addition is taken modulo $n+1$. Actually, $\Psi(\vartheta)$ is just the permutation obtained by writing $\tilde{\vartheta}^{-1}$ as a word and interpreting it as a $(n+1)$-cycle.

Example 25 *Consider the permutation*

$$\vartheta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 2 & 3 & 6 & 4 & 7 & 1 & 5 \end{pmatrix}.$$

Then we have that

$$\tilde{\vartheta} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 8 & 2 & 3 & 6 & 4 & 7 & 1 & 5 \end{pmatrix}$$

and

$$\tilde{\vartheta}^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 3 & 4 & 6 & 9 & 5 & 7 & 2 & 1 \end{pmatrix}.$$

It follows that

$$\Psi(\vartheta) = (8\,3\,4\,6\,9\,5\,7\,2\,1) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 1 & 4 & 6 & 7 & 9 & 2 & 3 & 5 \end{pmatrix}.$$

The following theorem, that appears in [28], gives a characterization of suffix arrays.

Theorem 1.9.3 *Let $n_1, n_2, \ldots, n_k$ be positive integers such that $n_1 + n_2 + \ldots + n_k = n$. A permutation $\vartheta \in S_n$ is the suffix array of a word $w \in A^n$, with Parikh vector $P(w) = (n_1, n_2, \ldots, n_k)$ if and only if $Des(\Psi(\vartheta)) \subseteq \{1, 1+n_1, \ldots, 1+n_1+\ldots+n_{k-1}\}$. Moreover, in this case, ϑ is the suffix array of exactly one such word.*

Proof. Let us suppose that $\vartheta \in S_n$ is the suffix array of a word $w \in A^n$. Then, by Proposition 22, $\tilde{\vartheta} \in S_{n+1}$ is the *BW*-array of the word $w' = w\sharp$. By observing that the *BW*-array of w' corresponds to the inverse of the permutation $\sigma_{w'}$ defined by the relation 1.7.25, we have that $\tilde{\vartheta} = (\sigma_{w'})^{-1}$. We show that $\Psi(\vartheta) = \pi_{w'}$, where $\pi_{w'}$ is the permutation defined by formula 1.7.28. Indeed, if $\mu = \Psi(\vartheta)$, we can write

$$\mu(i) = \tilde{\vartheta}^{-1}(\tilde{\vartheta}(i)+1) = \sigma_{w'}(\sigma_{w'}^{-1}(i)+1) = \pi_{w'}.$$

We now observe that, if $P(w) = (n_1, \ldots, n_k)$ is the Parikh vector of w, then the Parikh vector of $w' = w\sharp$ is $P(w') = (1, n_1, \ldots, n_k)$. Therefore, by Theorem 1.7.4,

$$Des(\Psi(\vartheta)) = Des(\pi_{w'}) \subseteq \rho(P(w')) = \{1, 1+n_1, \ldots, 1+n_1+\ldots+n_{k-1}\}.$$

Conversely, given a Parikh vector $V = (n_1, \ldots, n_k)$ and a permutation $\vartheta \in S_n$ such that $Des(\Psi(\vartheta)) \subseteq \{1, 1+n_1, \ldots, 1+n_1+\ldots+n_{k-1}\}$, we show that there exists a unique word $w \in A^n$ having Parikh vector $P(w) = V$ and suffix array $\vartheta_w = \vartheta$. Actually, we provide a construction of this word. Since $a_1 < a_2 < \ldots < a_k$, in the starting positions of the first n_1 suffixes of w, in the lexicographic order, there is the letter a_1, in the starting positions of the suffixes of w having rank from n_1+1 to n_1+n_2, in the lexicographic order, there is the letter a_2, and so on. Therefore, for $w = z_1 z_2 \cdots z_n$, if $1 \leq i \leq n_1$ then $z_{\vartheta(i)} = a_1$ and, for $1 < r \leq k$, if $n_1 + \ldots + n_{r-1} < i \leq n_1 + \ldots + n_r$, then $z_{\vartheta(i)} = a_r$. This concludes the proof. ∎

Example 26 *Given the permutation $\vartheta \in S_8$ in Example 25 and the vector $V = (5,3)$, we construct a word w on a binary alphabet $A = \{a,b\}$, with $a < b$, having V as Parikh vector and ϑ as suffix array. From Example 25, we have that*

$$\Psi(\vartheta) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 1 & 4 & 6 & 7 & 9 & 2 & 3 & 5 \end{pmatrix}.$$

$Des(\Psi(\vartheta)) = \{1,6\}$ verifies the condition of the Theorem 1.9.3. The word $w = z_1 \ldots z_8$ having Parikh vector $(5,3)$ and suffix array ϑ is obtained as follows:

$$z_{\vartheta(1)} = z_{\vartheta(2)} = z_{\vartheta(3)} = z_{\vartheta(4)} = z_{\vartheta(5)} = a$$

and

$$z_{\vartheta(6)} = z_{\vartheta(7)} = z_{\vartheta(8)} = b.$$

Therefore $w = baaababa$.

The following corollary of Theorem 1.9.3 will be useful in the next section.

Proposition 23 *A permutation $\vartheta \in S_n$ is the suffix array of some word w of length n on an alphabet of cardinality k if and only if*

$$\mathrm{Card}(Des(\Psi(\vartheta)) \setminus \{1\}) \leq k-1.$$

1.9.2 Counting suffix arrays

The results of previous sections are used here to solve three enumeration problems concerning suffix arrays. The results are essentially due to Schurmann and Stoye [38] (see also [15], [3] and [28]).

The first problem approached here is to count the number $s(n,k)$ of distinct permutations that are suffix arrays of some word of length n over an alphabet of size k.

The following table gives the values of $s(n,k)$ for $2 \leq k \leq n \leq 9$.

k / n	2	3	4	5	6	7	8	9
2	2							
3	5	6						
4	12	23	24					
5	27	93	119	120				
6	58	360	662	719	720			
7	121	1312	3728	4919	5039	5040		
8	248	4541	20160	35779	40072	40319	40320	
9	503	15111	103345	259535	347769	362377	362879	362880

In next theorem we show that the function $s(n,k)$ is related to the *Eulerian numbers* $\left\langle {n \atop d} \right\rangle$, i.e. the number of permutations of $\{1,2,...,n\}$ with exactly d descents. Recall (cf.[21]) that the Eulerian numbers can be defined by the following recurrence relation

$$\left\langle {n \atop d} \right\rangle = (d+1)\left\langle {n-1 \atop d} \right\rangle + (n-d)\left\langle {n-1 \atop d-1} \right\rangle$$

with $\left\langle {1 \atop 0} \right\rangle = 1$ and $\left\langle {1 \atop d} \right\rangle = 0$ when $d \geq 1$.

Theorem 1.9.4 *The number $s(n,k)$ of distinct permutations that are suffix arrays of some word of length n over an alphabet of size k is*

$$s(n,k) = \sum_{d=0}^{k-1} \left\langle {n \atop d} \right\rangle$$

In order to prove the theorem we need a preliminary lemma. In the following it is convenient to represent a permutation $\varphi \in S_n$ by the word $\varphi(1)\varphi(2)...\varphi(n)$ on the alphabet $\{1,2,\ldots,n\}$. Now we define a mapping that, for any $\varphi \in S_n$ and for any $s \in \{2,3,\ldots,n+1\}$, gives a permutation $\psi \in S_{n+1}$. Such a mapping is described as a transformation on words performed in three steps.

For a permutation $\varphi(1)\varphi(2)\cdots\varphi(n)$ and an integer $s \in \{2,3,\ldots,n+1\}$, in the *first* step we obtain the word

$$E_s(\varphi) = \varphi_s(1)\varphi_s(2)\cdots\varphi_s(n),$$

where $\varphi_s(i) = \varphi(i)$ for $\varphi(i) < s$, and $\varphi_s(i) = \varphi(i)+1$ for $\varphi(i) \geq s$. Remark that $\varphi_s(1)\varphi_s(2)\cdots\varphi_s(n)$ is a word on the alphabet $\{1,2,\ldots,n,n+1\}$, but it does not represent a permutation, because the integer s does not appear in the word. For instance, consider the permutation $\varphi \in S_6$ represented by the word 364215 and $s = 3$. Then $E_3(\varphi) = 475216$.

In the second step I_s, which is the most important, we move $\varphi_s(1)$ from the first position in the word to the position $s-1$. It is called the *insertion step* and it is formally defined as follows:

$$I_s(\varphi_s(1)\varphi_s(2)\cdots\varphi_s(n)) = \varphi_s(2)\cdots\varphi_s(s-1)\varphi_s(1)\varphi_s(s)\cdots\varphi_s(n).$$

For instance, $I_3(475216) = 745216$.

In the third step C_s we simply insert the symbol s in the first position of the word. For instance, $C_3(745216) = 3745216$.

The compositions of the above operations define the transformation $T(\varphi,s) = C_s(I_s(E_s(\varphi)))$. Remark that the word $T(\varphi,s)$ represents a permutation of $\{1,2,\ldots,n,n+1\}$. For instance, for $\varphi = 364215$ and $s = 3$, we have $T(\varphi,s) = 3745216$. Moreover, it is straightforward to check that, if φ is *cyclic*, then $T(\varphi,s)$ is *cyclic* too. Therefore, if we denote by S_n^c the set of *cyclic* permutations of $\{1,2,\ldots,n\}$, the transformation T defines a mapping

$$T : S_n^c \times \{2,3,\ldots,n+1\} \to S_{n+1}^c.$$

Lemma 4 *The mapping T is a bijection from $S_n^c \times \{2,3,\ldots,n+1\}$ onto S_{n+1}^c.*

Proof. We first prove that T is injective by showing that, given a permutation $\psi \in S_{n+1}^c$, one can uniquely reconstruct the pair (φ,s), with $\varphi \in S_n^c$ and $s \in \{2,\ldots,n+1\}$, such that $T(\varphi,s) = \psi$. Let $\psi = \psi(1)\psi(2)\cdots\psi(n+1)$. Since ψ is a cyclic permutation, $\psi(1) \neq 1$. By the definition of T, $s = \psi(1)$. We delete $\psi(1) = s$ from the word $\psi(1)\psi(2)\cdots\psi(n+1)$, and we obtain the word $\psi(2)\cdots\psi(n+1)$. Then we take the element $\psi(s)$ and move this element in the first position of the word. We obtain the word $\psi(s)\psi(2)\cdots\psi(s-1)\psi(s+1)\cdots\psi(n+1)$. Now we substitute each $\psi(j) > s$ with $\psi(j)-1$ and we obtain the permutation $\varphi \in S_n^c$ such that $T(\varphi,s) = \psi$. In order to show that the mapping T is surjective, it suffices to verify that $\mathrm{Card}(S_n^c \times \{2,3,\ldots,n+1\}) = \mathrm{Card}(S_{n+1}^c)$. Indeed $\mathrm{Card}(S_n^c \times \{2,3,\ldots,n+1\}) = (n-1)!n = n! = \mathrm{Card}(S_{n+1}^c)$. ∎

Proof of Theorem 1.9.4. According to Proposition 23, there is a bijection between the suffix arrays of words $w \in A^n$ and the cyclic permutations $\psi \in S_{n+1}^c$ such that $\mathrm{Card}(Des(\psi)\setminus\{1\}) \leq k-1$. We have then to count the number of such permutations.

Let $P(n,d)$ denote the number of permutations $\psi \in S_{n+1}^c$ such that $\mathrm{Card}(Des(\psi)\setminus\{1\}) = d$. To prove the theorem, we show that $P(n,d)$ is equal to the Eulerian number $\left\langle {n \atop d} \right\rangle$.

The proof is by induction on n. Trivially, $P(1,0) = 1 = \left\langle {1 \atop 0} \right\rangle$, and $P(1,d) = 0 = \left\langle {1 \atop d} \right\rangle$ when $d \geq 1$.

We now show that $P(n,d) = (d+1)P(n-1,d) + (n-d)P(n-1,d-1)$.

By Lemma4, a permutation $\psi \in S_{n+1}^c$ can be obtained, through the transform T, from a permutation $\varphi \in S_n^c$ with the "insertion" of an element $s \in \{2,\ldots,n+1\}$.

We now examine how the transform T affects the number of descents of φ. Remark that the steps 1 and 3 in the definitions of the transform T do not affect the number of descents. This number can be affected only in step 2 (the *insertion step* I_s). If φ has d descents in the interval $\{2,\ldots,n+1\}$, also $E_s(\varphi)$, the word obtained after the first step, has d descents, independently from the choise of s. We can thus factorize $E_s(\varphi)$ in $d+1$ monotonic (increasing) runs. The second step in the transform T (the insertion step I_s) may or may not create a new descent, depending on the position in which is inserted the first symbol $\varphi_s(1)$ of the word $E_s(\varphi)$. In each monotonic run of $E_s(\varphi)$ there is exactly one position where $\varphi_s(1)$ can be placed without creating a new descent. Otherwise one creates exactly one new descent.

How many permutations $\psi = T(\varphi,s)$ can we obtain with $\mathrm{Card}(Des(\psi)\setminus\{1\}) = d$? For each $\varphi \in S_n^c$ with $\mathrm{Card}(Des(\varphi)\setminus\{1\}) = d$, we have $d+1$ possibilities to choose s (because in $E_s(\varphi)$ there are $d+1$ monotonic runs). For each $\varphi \in S_n^c$ with $\mathrm{Card}(Des(\varphi)\setminus\{1\}) = d-1$, we have $n-d$ possibilities to choose s. Since T is a bijection, there is no other way to get a permutation $\psi \in S_{n+1}^c$ with $\mathrm{Card}(Des(\psi)\setminus\{1\}) = d$. It follows that

$$P(n,d) = (d+1)P(n-1,d) + (n-d)P(n-1,d-1).$$

■

We now consider the problem of counting the number of words that share the same suffix array.

Theorem 1.9.5 *Given a permutation $\vartheta \in S_n$, the number of words of length n over an alphabet of size k having ϑ as their suffix array is*

$$\binom{n+k-1-d}{k-1-d},$$

where $d = \mathrm{Card}(Des(\Psi(\vartheta))\setminus\{1\})$.

Proof. By Theorem 1.9.3, a word $w \in A^n$, with $|A| = k$, has ϑ as suffix array if and only if w has a Parikh vector $P(w) = (n_1, n_2, \ldots, n_k)$ such that

$$Des(\Psi(\vartheta)) \subseteq \{1, 1+n_1, \ldots, 1+n_1+\ldots+n_{k-1}\}.$$

Therefore, given the permutation ϑ, and then given the set

$$D_\vartheta = Des(\Psi(\vartheta))\setminus\{1\} = \{m_1, m_2, \ldots, m_d\},$$

we need to count the number of tuples $(n_1,\ldots,n_k)$, with $n_1+\ldots+n_k = n$ such that

$$D_\vartheta \subseteq \{1+n_1, 1+n_1+n_2, \ldots, 1+n_1+\ldots+n_{k-1}\}.$$

We represent the tuple $(n_1,\ldots,n_k)$ by a word z on the alphabet $\{x,y\}$:

$$z = x^{n_1} y x^{n_2} y \cdots x^{n_{k-1}} y x^{n_k},$$

with $n_i \geq 0$ and $n_1 + \ldots + n_k = n$. We have that $|z| = n + k - 1$. The condition $D_\vartheta = \{m_1, \ldots, m_d\} \subseteq \{1 + n_1, \ldots, 1 + n_1 + \ldots + n_{k-1}\}$ defines the positions of d occurrences of the letter y in z. The remaining $k - 1 - d$ occurrences of y can be placed in arbitrary positions. This can be done in

$$\binom{n+k-1-d}{k-1-d}$$

ways. ∎

Note that if $k - 1 < \mathrm{Card}(Des(\Psi(\vartheta)) \setminus \{1\})$, there is no word on an alphabet of size k which has ϑ as its suffix array. This is confirmed by Theorem 1.9.5, since $\binom{m}{n} = 0$ for $m < n$.

In the next theorem, we require that each letter of the alphabet occurs at least once in the words that we count.

Theorem 1.9.6 *Given a permutation $\vartheta \in S_n$, the number of words of length n over an alphabet of size k that have at least one occurrence of each of the k letters and have ϑ as their suffix array is*

$$\binom{n-1-d}{k-1-d},$$

where $d = \mathrm{Card}(Des(\Psi(\vartheta)) \setminus \{1\})$.

Proof. The proof of Theorem 1.9.5 is modified in order to ensure that each letter occurs at least once. In the representation of the tuple $(n_1, \ldots, n_k)$ by the word $z = x^{n_1} y x^{n_2} y \cdots x^{n_{k-1}} y x^{n_k}$, we require that the n_i are strictly positive, i.e. $n_i > 0$ for $i = 1, \ldots, k-1$. Then we have to distribute the occurrences of the letter y among the $n-1$ possible positions. As in the proof of Theorem 1.9.5, the positions of d occurrences of y is determined by the permutation ϑ, and the remaining $k - 1 - d$ are distributed among the $n - 1 - d$ remaining positions. ∎

From Theorem 1.9.4 and Theorem 1.9.5 we can derive a long known summation identity of Eulerian numbers. The identity

$$k^n = \sum_j \left\langle {n \atop j} \right\rangle \binom{k+j}{n},$$

as given in [21, Eq.6.37], was proven by J. Worpitzki, already in 1883. In order to prove it, we observe that the number of words of length n over an alphabet of size k can be obtained by summing the number of words for each suffix array. Thus, we have:

$$k^n = \sum_{d=0}^{k-1} \left\langle {n \atop d} \right\rangle \binom{n+k-d-1}{k-d-1}.$$

By using the symmetry rule for Eulerian and binomial numbers, from the previous equality we derive

$$k^n = \sum_{d=0}^{k-1} \left\langle {n \atop n-1-d} \right\rangle \binom{n+k-d-1}{n}.$$

By setting $j = n-d-1$, we obtain

$$k^n = \sum_{j=n-k}^{n-1} \left\langle {n \atop j} \right\rangle \binom{k+j}{n} = \sum_{j} \left\langle {n \atop j} \right\rangle \binom{k+j}{n},$$

where the last equality is motivated by the remark that $\left\langle {n \atop j} \right\rangle = 0$ for all $j \geq n$ and $\binom{k+j}{n} = 0$ for all $j < n-k$.

References

[1] Tanja van Aardenne-Ehrenfest and Nicolaas Govert de Bruijn. Circuits and trees in oriented linear graphs. *Simon Stevin*, 28:203–217, 1951.

[2] Yu Hin Au. Shortest sequences containing primitive words and powers. 2013. arXiv:0904.3997.

[3] Hideo Bannai, Shunsuke Inenaga, Ayumi Shinohara, and Masayuki Takeda. Inferring strings from graphs and arrays. volume 2747 of *Lecture Notes in Computer Science*, pages 208–217. Springer Berlin Heidelberg, 2003.

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

[5] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2009.

[6] Francine Blanchet-Sadri. Algorithmic combinatorics on partial words. *Internat. J. Found. Comput. Sci.*, 23(6):1189–1206, 2012.

[7] Francine Blanchet-Sadri, N. C. Brownstein, Andy Kalcic, Justin Palumbo, and T. Weyand. Unavoidable sets of partial words. *Theory Comput. Syst.*, 45(2):381–406, 2009.

[8] Carl Wilhelm Borchardt. Ueber eine der Interpolation entsprechende Darstellung der Eliminations-Resultante. *J. reine angew. Math.*, 57:111–121, 1860.

[9] Michael Burrows and David J. Wheeler. A block sorting data compression algorithm. Technical report, DIGITAL System Research Center, 1994.

[10] Jean-Marc Champarnaud and Georges Hansel. Ensembles inévitables et classes de conjugaison. *Bull. Belg. Math. Soc. Simon Stevin*, 10(suppl.):679–691, 2003.

[11] Jean-Marc Champarnaud, Georges Hansel, and Dominique Perrin. Unavoidable sets of constant length. *Internat. J. Algebra Comput.*, 14(2):241–251, 2004.

[12] Maxime Crochemore, Jacques Désarménien, and Dominique Perrin. A note on the Burrows-Wheeler transformation. *Theoret. Comput. Sci.*, 332(1-3):567–572, 2005.

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

[14] Jean-Pierre Duval. Génération d'une section des classes de conjugaison et arbre des mots de Lyndon de longueur bornée. *Theoret. Comput. Sci.*, 60(3):255–283, 1988.

[15] Jean-Pierre Duval and Arnaud Lefebvre. Words over an ordered alphabet and suffix permutations. *RAIRO Theor. Inform. Appl.*, 36(3):249–259, 2002.

[16] Steven R. Finch. *Mathematical constants*, volume 94 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2003.

[17] Harold Fredricksen and James Maiorana. Necklaces of beads in k colors and k-ary de Bruijn sequences. *Discrete Math.*, 23(3):207–210, 1978.

[18] Michael R. Garey and David S. Johnson. *Computers and intractability*. W. H. Freeman and Co., San Francisco, Calif., 1979. A guide to the theory of NP-completeness, A Series of Books in the Mathematical Sciences.

[19] Ira M. Gessel, Antonio Restivo, and Christophe Reutenauer. A bijection between words and multisets of necklaces. *European Journal of Combinatorics*, 33(7):1537 – 1546, 2012.

[20] Ira M. Gessel and Christophe Reutenauer. Counting permutations with given cycle structure and descent set. *J. Combin. Theory Ser. A*, 64(2):189–215, 1993.

[21] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics*. Addison-Wesley Publishing Company, Reading, MA, second edition, 1994. A foundation for computer science.

[22] Roberto Grossi. A quick tour on suffix arrays and compressed suffix arrays. *Theoret. Comput. Sci.*, 412(27):2964 – 2973, 2011.

[23] Peter M. Higgins. Burrow-Wheeler transformations and de Bruijn words. *Theoret. Comput. Sci.*, 457(0):128 – 136, 2012.

[24] Donald E. Knuth. Oriented subtrees of an arc digraph. *J. Comb. Theory*, 3:309–314, 1967.

[25] Donald E. Knuth. *The Art of Computer Programming, volume 1, Fundamental Algorithms*. Addison Wesley, 1968. Second edition, 1973.

[26] Donald E. Knuth. *The Art of Computer Programming , Volume 4A, Combinatorial Algorithms: Part 1*. Addison Wesley, 2012.

[27] Tomasz Kociumaka, Jakub Radoszewski, and Wojciech Rytter. Computing k-th lyndon word and decoding lexicographically minimal de Bruijn sequence. In *Combinatorial Pattern Matching*, volume 8486 of *Lecture Notes in Computer Science*, pages 202–211, 2014.

[28] Gregory Kucherov, Lilla Tóthmérész, and Stéphane Vialette. On the combinatorics of suffix arrays. *Inform. Process. Lett.*, 113(22-24):915–920, 2013.

[29] Douglas Lind and Brian H. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge, 1995.

[30] M. Lothaire. *Combinatorics on Words*. Cambridge University Press, second edition, 1997. (First edition 1983).

[31] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.

[32] Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

[33] Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, and Marinella Sciortino. An extension of the Burrows-Wheeler Transform. *Theoret. Comput. Sci.*, 387(3):298–312, 2007.

[34] Eduardo Moreno. On the theorem of Fredricksen and Maiorana about de Bruijn sequences. *Adv. in Appl. Math.*, 33(2):413–415, 2004.

[35] Eduardo Moreno and Dominique Perrin. Corrigendum to: 'on the theorem of Fredricksen and Maiorana about de Bruijn sequences'. *Adv. in Appl. Math.*, 2014. to appear.

[36] Johannes Mykkeltveit. A proof of Golomb's conjecture for the de Bruijn graph. *J. Combinatorial Theory Ser. B*, 13:40–45, 1972.

[37] Christophe Reutenauer. *Free Lie algebras*. The Clarendon Press Oxford University Press, New York, 1993. Oxford Science Publications.

[38] Klaus-Bernd Schurmann and Jens Stoye. Counting suffix arrays and strings. *Theor. Comput. Sci.*, pages 220–234, 2008.

[39] Arseny M. Shur. Growth of power-free languages over large alphabets. *Theory Comput. Syst.*, 54(2):224–243, 2014.

[40] Cedric A. Smith and William T. Tutte. On unicursal paths in a network of degree 4. *Amer. Math. Monthly*, 48, 1941.

[41] Richard P. Stanley. *Enumerative combinatorics. Vol. 1*. Cambridge University Press, Cambridge, 1997.