LENGTH DISTRIBUTIONS AND REGULAR SEQUENCES

FRÉDÉRIQUE BASSINO , MARIE-PIERRE BÉAL , AND DOMINIQUE PERRIN *

Abstract. This paper presents a survey on length distributions of regular languages. The accent is on problems in coding theory and the relation with symbolic dynamics.

Key words. Regular sequences, finite automata, prefix codes, bifix codes, symbolic dynamics, zeta functions.

1. Introduction. The notion of a length distribution for a formal language is a simple one: it is the generating series $u(z) = \sum_{n \geq 0} u_n z^n$ of the number of words of each length. This series carries important information concerning a formal language since it measures in a sense the size of the language. It is moreover appropriate in the case of coding. In fact, a length-preserving encoding defines a one-to-one correspondence between words. The two sets of words in such a correspondence will have the same length distribution.

It is a classical result that the length distribution of a formal language carries also some information concerning the structure of the language, in the sense that algebraic operations on series correspond to operations on formal languages. Thus, as we shall see below in more detail, length distributions which are rational series correspond to regular languages.

This correspondence between operations on series and on sets is the basis of the method of generating series in enumerative combinatorics. Numerous examples of applications can be found in the book of Graham, Knuth and Pataschnik [23].

We present here a survey on length distributions of formal languages with emphasis on the problems related to coding and finite automata. We insist on the following general problem: given a family $\mathcal{F}$ of sets of words, characterize the length distributions of the elements of $\mathcal{F}$. For example, the length distributions of prefix codes on k-symbols are the sequences satisfying Kraft's inequality

$$\sum_{n \geq 0} u_n k^{-n} \leq 1,$$

i.e. $u(1/k) \leq 1$.

Our emphasis is on the property of regularity which is the definability by a finite automaton. This places our work at the intersection between

*Institut d'Électronique et d'Informatique Gaspard-Monge, Université de Marne la Vallée, 5, Boulevard Descartes, Champs-sur-Marne, 77454 Marne la Vallée Cedex 2, France. http://www-igm.univ-mlv.fr/

coding theory and automata theory. For example, one of the main results presented here is a finite-state version of Kraft-McMillan's theorem characterizing the length distributions of regular prefix codes.

We also make connexions with the field of symbolic dynamics. This is natural since the basic notion of symbolic dynamics, namely the conjugacy of subshifts is based on a one-to-one correspondence between paths in finite graphs, giving rise to an invariance of the length distributions.

Our paper is organized as follows. The first sections (Sections 2,3) present the basic notions on automata and formal series used in the paper. In Section 4, we present the finite-state version of Kraft-McMillan theorem mentioned above. The particular case of bifix codes is studied in Section 5. The last section (Section 6) presents several interconnected notions concerning subshifts of finite type and circular codes.

2. Length distributions. We consider the set A^* of all words on a given alphabet A. A subset of A^* is often called a *formal language.* For sets $X, Y \subset A^*$, we denote

$$\begin{aligned} &X + Y = X \cup Y, \\ &XY = \{xy \mid x \in X, y \in Y\}, \\ &X^* = \{x_1 x_2 \cdots x_n \mid x_i \in X, n \geq 0\} \end{aligned}$$

We say that the pair (X, Y) is unambiguous if for each $z \in XY$ there is at most one pair $(x, y) \in X \times Y$ such that $z = xy$.

We say that a set of nonempty words X is a *code* if for each $x \in X^*$ there is at most one sequence $(x_1, x_2, \ldots, x_n)$ with $x_i \in X$ such that $x = x_1 x_2 \cdots x_n$ (one also says that X is uniquely decipherable). A particular case of a code is a *prefix code.* It is a set of words X such that no element of X is a prefix of another one. It is easy to see that such a set is either reduced to the empty word or does not contain the empty word and is then a code.

The *length distribution* of a set of words X is the sequence $u_X = (u_n)_{n \geq 0}$ with

$$u_n = \mathrm{Card}(X \cap A^n).$$

We denote by u_X the formal series

$$u_X(z) = \sum_{n \geq 0} u_n z^n.$$

which is the ordinary generating series of the sequence u_X.

For example, the length distribution of $X = A^*$ is $u(z) = \frac{1}{1-kz}$ where $k = \mathrm{Card}(A)$.

The *entropy* of a formal language X is

$$h(X) = \log(1/\rho),$$

where ρ is the radius of convergence of the series $u_X(z)$. It is well defined provided X is infinite and thus ρ is finite. If the alphabet A has k elements, we have $h(X) \leq \log k$.

The following result relates the basic operations on sets with operations on series.

PROPOSITION 2.1. *The following properties hold for any subsets X, Y of A^*.*

(i) *If $X \cap Y = \emptyset$, then $u_{X+Y} = u_X + u_Y$.*
(ii) *If the pair (X, Y) is unambiguous, then $u_{XY} = u_X u_Y$.*
(iii) *If X is a code, then $u_{X^*} = 1/(1 - u_X)$.*

Proof. The first two formulae are clear. If X is a code, every word in X^* has a unique decomposition as a product of words in X. This implies that

$$u_{X^n} = (u_X)^n$$

and thus,

$$u_{X^*} = 1 + u_X + \cdots + u_{X^n} + \cdots = 1/(1 - u_X)\,.$$

□

EXAMPLE 1. *The set $X = \{b, ab\}$ is a prefix code. The series u_{X^*} is*

$$u_{X^*}(z) = \frac{1}{1 - z - z^2}.$$

Let $(F_n)_{n\geq 0}$ be the sequence of Fibonacci numbers defined by $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$. It follows from the recurrence relation that

$$\frac{z}{1 - z - z^2} = \sum_{n\geq 0} F_n z^n.$$

Consequently, $u_{X^*}(z) = \sum_{n\geq 0} F_{n+1} z^n$. It can also be proved by a combinatorial argument that the number of words of length n in X^* is F_{n+1}.

There are several variants of the generating series considered above. One may first define

$$p_X(z) = \sum_{n\geq 0} \frac{u_n}{k^n} z^n,$$

where $k = \mathrm{Card}(A)$. The coefficients of z^n in $p_X(z)$ is the probability for a word of length n to be in the set X. The relation between u_X and p_X is simple since $p_X(z) = u_X(z/k)$. Another variant of the generating series is the *exponential generating series* of the sequence $(u_n)_{n\geq 0}$ defined as

$$e(z) = \sum_{n\geq 0} \frac{u_n}{n!} z^n.$$

We will also use the zeta function of a sequence $(u_n)_{n\geq 1}$ defined as

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

3. Regular distributions. In this section, we describe the connection between the notions of a regular language and a rational series. We prove the classical result (Theorem 3.4) characterizing the regular sequences as the length distributions of regular languages. We mention finally the possible extension to more general classes of formal languages, such as the context-free languages. These results are well-known in the theory of automata and we include them here for the sake of the reader's convenience.

A word on the terminology used here. We use constantly the term *regular* where a richer terminology is often used. In particular, what we call here a regular sequence is, in Eilenberg's terminology, an $\mathbb{N}$-rational sequence (see [20], [33] or [16]). A regular set is also called a *rational* or *recognizable* set.

3.1. Regular sequences. A sequence $u = (u_n)_{n\geq 0}$ of integers is *regular* if there exists a finite graph G and two sets of vertices I, T of G such that for all $n \geq 0$,

$$u_n = \mathrm{Card}(P(n, I, T)),$$

where $P(n, I, T)$ is the set of paths of length n from a vertex of I to a vertex of T. The graph G is one in which multiples edges are allowed (sometimes called a multigraph). We say that the graph G *recognizes* the sequence u.

An equivalent definition of regular sequences is obtained by considering nonnegative matrices.

PROPOSITION 3.1. *A sequence $u = (u_n)_{n\geq 0}$ of integers is regular iff there exists a nonnegative matrix $M \in \mathbb{N}^{k\times k}$ and two vectors $l, c \in \mathbb{N}^k$ such that*

$$u_n = lM^n c,$$

where l is considered as a row vector and c as a column vector.

Proof. Let u be a regular sequence defined by a graph G on the set $\{1, \ldots, k\}$ of vertices. We choose M to be the adjacency matrix of G, i.e. for each pair v, w of vertices, $M_{v,w}$ is the number of edges from v to w. Let l be the row vector defined by $l_v = 1$ if $v \in I$ and 0 otherwise. Let c be the column vector defined by $c_v = 1$ if $v \in T$ and 0 otherwise. The number of paths of length n from a vertex of I to a vertex of T is for each $n \geq 1$ equal to $lM^n c$.

Conversely, let G be the graph with adjacency matrix M. Since the family of regular sequences is closed under addition, we may suppose that the vectors l, c have $0, 1$ coefficients. We can then consider l, c as the characteristic vectors of sets I, T of vertices. It is then obvious that the graph thus constructed recognizes u. □

EXAMPLE 2. *Let G be the graph of Figure 1. The number of paths of length n from vertex $i = 1$ to vertex $t = 2$ is the Fibonacci number F_n.*

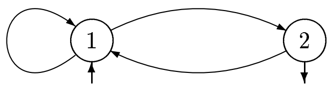

FIG. 1. *The Fibonacci graph.*

Accordingly, let M be the matrix

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

The same sequence is defined by the equation

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

We say that a sequence u of integers is *rational* if $u(z) = p(z)/q(z)$ for some polynomials $p(z), q(z)$ with integer coefficients. The following result is classical.

THEOREM 3.1. *Any regular sequence u of nonnegative integers is rational.*

Proof. Let (l, M, c) be such that $u_n = lM^n c$. We have

$$u(z) = \sum_{n\geq 0} lM^n cz^n = l(\sum_{n\geq 0}(Mz)^n)c = l(I - Mz)^{-1}c.$$

The result follows since the coefficients of $(I - Mz)^{-1}$ are rational fractions. □

EXAMPLE 3. *The generating function of the Fibonacci sequence is*

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

The converse of Theorem 3.1 is not true. We have actually the following result, due to Jean Berstel (see [20] or [16]).

THEOREM 3.2. *For any regular sequence u, there is an integer p such that the set of poles of minimal modulus is the set of complex numbers $\rho\varepsilon$ where ρ is the radius of convergence of u and $\varepsilon^p = 1$ for some $p \geq 1$.*

In particular, the radius of convergence is a pole.

The following example (from [20] Example 6.1, Chapter VIII) shows the existence of rational series with non-negative integer coefficients which are not regular.

EXAMPLE 4. *Let $0 < \theta < \pi/2$ be such that $\cos\theta = a/c$ with $0 < a < c$ and $c \neq 2a$. The sequence*

$$u_n = c^{2n}\cos^2 n\theta$$

is rational but not regular (poles: $1, e^{2i\theta}, e^{-2i\theta}$*).*

A sequence u is a *merge* of sequences

$$u^{(0)}, \dots, u^{(p-1)}$$

if for $n \geq 0, 0 \leq i < p$,

$$u_{pn+i} = u_n^{(i)}.$$

We say that a pole of a rational series is *dominating* if it is strictly less than the modulus of all other ones. The following result is due to Soittola (see [33]).

THEOREM 3.3. *A sequence of non-negative integers is regular iff it is an merge of rational sequences with a dominating pole.*

EXAMPLE 5. *The sequence*

$$1, 1, 2, 1, 4, 2, 8, 3, 16, 5, \dots$$

is the merge of the sequence of powers of 2 *and the Fibonacci sequence.*

A third equivalent definition of regular sequences is possible. One can indeed show that a series $u(z)$ is regular iff it can be obtained by a finite number of operations of sum, product and star with

$$u^*(z) = \frac{1}{1 - u(z)},$$

starting from polynomials with nonnegative integer coefficients. An expression of this form is usually called a *regular expression.*

EXAMPLE 6. *The sequence* $(0, 1, 3, 8, 21, \dots)$ *formed of the Fibonacci numbers of even index is regular. Indeed we have*

$$F_{2n} = lM^{2n}c$$

with the triple (l, M, c) *of Example 2. We have*

$$M^2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

and thus F_{2n} *is the number of paths of length* n *from* 1 *to* 2 *in the graph of Figure 2. The series* $s(z) = \sum_{n \geq 0} F_{2n} z^n$ *can accordingly be written*

$$s(z) = z(2z + z^2 z^*)^* = \frac{z(1-z)}{1 - 3z + z^2}.$$

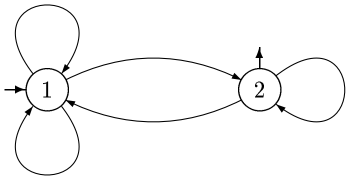

FIG. 2. *One every other Fibonacci number*

3.2. Finite automata. We present here a brief introduction to the concepts used in automata theory. For a general reference, see [31] or [20].

An *automaton* over the alphabet A is composed of a set Q of *states*, a set $E \subset Q \times A \times Q$ of *edges* or *transitions* and two sets $I, T \subset Q$ of *initial* and *terminal* states.

A *path* in the automaton $\mathcal{A}$ is a sequence

$$(p_1, a_1, p_2), (p_2, a_2, p_3), \ldots, (p_n, a_n, p_{n+1})$$

of consecutive edges. Its label is the word $x = a_1 a_2 \cdots a_n$. A path is *successful* if it starts in an initial state and ends in a terminal state. The set *recognized* by the automaton is the set of labels of its successful paths.

An automaton is *deterministic* if, for each state p and each letter a, there is at most one edge which starts at p and is labeled by a. The term *right resolving* is also used.

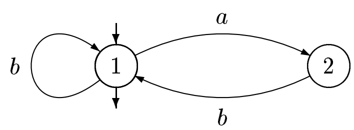

FIG. 3. *Golden mean automaton.*

EXAMPLE 7. *Let $\mathcal{A}$ be the automaton given in Figure 3 with* 1 *as unique initial and terminal state. It recognizes the set X^* where X is the prefix code $X = \{b, ab\}$.*

A set of words X over A is *regular* if it can be recognized by a finite automaton.

It is a classical result that a set of words is regular iff it can be obtained by a finite number of operations union, product and star, starting form the finite sets.

The following result is also classical.

PROPOSITION 3.2. *Every regular set can be recognized by a finite deterministic automaton having a unique initial state.*

Proof. Let $\mathcal{A} = (Q, E, I, T)$ be a finite automaton over A recognizing a set X. Let $\mathcal{B} = (\mathcal{R}, F, \{I\}, \mathcal{T})$ be the automaton defined as follows. Its states are the subsets

$$Q(u) = \{q \in Q \mid i \xrightarrow{u} q \text{ for some } i \in I\}$$

for all u in A^*. Since Q is finite, there is a finite number of subsets $Q(u)$. The edges of $\mathcal{B}$ are all triples

$$(Q(u), a, Q(ua)).$$

The set of terminal states is

$$\mathcal{T} = \{U \in \mathcal{R} \mid U \cap T \neq \emptyset\}.$$

It is easy to verify that $\mathcal{B}$ is deterministic and recognizes X. □

THEOREM 3.4. *The length distributions of regular sets are the regular sequences.*

Proof. Let X be a regular set. By Proposition 3.2, it can be recognized by a deterministic automaton $\mathcal{A}$. Since $\mathcal{A}$ is deterministic, there is at most one path with given label, origin and end. Thus the number of paths of length n from the initial state to a terminal state is equal to the number u_n of words of X of length n.

Conversely, let u be a regular sequence enumerating the paths in a graph G from I to T. We consider the graph G as an automaton with all edges with distinct labels. Let X be the set of labels of paths from I to T. The sequence u is the length distribution of the set X. □

EXAMPLE 8. *If $X = a^*b$, then*

$$u_X(z) = \frac{z}{1-z}.$$

3.3. Beyond regular sequences. There are several natural classes of series beyond the rational ones. The algebraic series are those satisfying an algebraic equation. More generally, the hypergeometric series are those such that the quotient of two successive terms is given by a rational fraction (see [23]).

The class of algebraic series is linked with the class of context-free sets (see [21]). A typical example of a context-free set is the set of words on the binary alphabet $\{a, b\}$ having as many a's as b's. We compute below its length distribution which is an algebraic series.

EXAMPLE 9. *The set of words on $A = \{a, b\}$ having an equal number of occurrences of a and b is a submonoid of A^* generated by a prefix code D. Since any word of D^* of length $2n$ is obtained by choosing n positions among $2n$, we have*

$$u_{D^*}(z) = \sum_{n \geq 0} \binom{2n}{n} z^{2n}.$$

By a simple application of the binomial formula, we obtain

$$u_{D^*}(z) = (1 - 4z^2)^{-\frac{1}{2}}.$$

This follows indeed, using the simple identity

$$\binom{-\frac{1}{2}}{n} = \frac{1}{(-4)^n}\binom{2n}{n}.$$

We have $u_D(z) = 1 - 1/u_{D^*}(z)$ *and thus*

$$u_D(z) = 1 - \sqrt{1 - 4z^2}.$$

Thus $u_D(z)$ *is an algebraic series, solution of the equation*

$$f^2 - 2f + 4z^2 = 0.$$

4. A finite-state version of the Kraft-McMillan theorem. Let X be a prefix code on an alphabet with k symbols. It is classical that its length distribution $u = (u_n)_{n\geq 1}$ satisfies Kraft's inequality

$$\sum_{n\geq 1} u_n k^{-n} \leq 1,$$

or equivalently $u(1/k) \leq 1$. The number $u(1/k)$ can actually be interpreted as the probability that a long enough word has a prefix in X.

There is also a connexion with the notion of entropy. Actually, if X is a prefix code, the entropy of X^* is equal to $\log(1/\rho)$ where ρ is the solution of the equation $u_X(\rho) = 1$. Thus Kraft's inequality expresses the fact that $h(X^*) \leq \log k$.

Conversely, Kraft-McMillan's theorem states that for any such sequence $u = (u_n)_{n\geq 1}$, there exists a prefix code X on a k-symbol alphabet such that $u = u_X$.

Let us briefly describe the proof. We suppose by induction to have already built a prefix code X formed of words of length at most $n-1$ with length distribution $(u_1, u_2, \ldots, u_{n-1})$ on the alphabet $A_k = \{0, 1, \ldots, k-1\}$. We have

$$\sum_{i=1}^{n} u_i k^{-i} \leq 1,$$

and thus

$$\sum_{i=1}^{n} u_i k^{n-i} \leq k^n.$$

This allows us to choose u_n words on the alphabet A_k of length n without a prefix in X. For the sake of a complete description of the construction, we have to specify the choice made at each step among the words of length

n which do not have already a prefix in X. A possible policy is to choose the earlier ones in the alphabetic order.

The equality case in Kraft's inequality corresponds to a particular class of prefix codes often called *complete*. A prefix code X on the alphabet A is complete if any word on A has either a prefix in X or is a prefix of a word of X.

The notion of a prefix code is related to the notion of a tree. A prefix code on k symbols corresponds to a k-ary tree. The length distribution of the prefix code is the enumerative sequence of the leaves of the tree. We call it the *length distribution* of the tree. Usually, the interest is focused on finite trees, as in Huffman algorithm for example.

We are interested here in the case of infinite trees and, more especially of regular trees arising from prefix codes which are regular, in the sense defined above. The notion of a regular tree can also be defined directly as an infinite tree with only a finite number of non-isomorphic subtrees.

By Theorem 3.4, if X is regular, then the sequence u_X is also regular. The following result shows that conversely the conjunction of the two conditions (of being regular and to satisfy Kraft's inequality) is sufficient to ensure the existence of a regular prefix code on a k-symbol alphabet.

THEOREM 4.1. *A sequence u of integers is the length distribution of a regular prefix code on k symbols iff*

(i) *it is regular.*
(ii) *it satisfies Kraft's inequality $u(1/k) \leq 1$.*

The essence of this result is a constructive method allowing one to build the regular prefix code X given the sequence u.

Two simple methods come to mind at first glance. The first one is to apply directly the proof of the Kraft's theorem. The following example shows that the result need not be a regular set, although the sequence u is itself regular.

EXAMPLE 10. *Let $u(z) = z^2/(1-2z^2)$. Since $u(1/2) = 1/2$, we may apply the Kraft construction to build a binary tree with length distribution u. The result is the set*

$$X = \bigcup_{n \geq 0} 01^n0\{0,1\}^n$$

which is not regular.

The second method takes into account the hypothesis that the sequence is regular. It will fail in its naive version but the solution is a refinement of this idea. Let G be a graph such that u_n is the number of paths of length n from I to T. We can normalize the graph G to obtain a graph such that $I = \{i\}$, $T = \{t\}$ and that no edge goes out of t. We label each edge in such a way that edges with a common start have different labels. The set recognized by the automaton thus constructed is a prefix code with length distribution equal to u.

The trouble is that the number of symbols used may well be larger than k as shown by the following example.

EXAMPLE 11. *Let u be the regular sequence given by the graph of Figure 4 on the left with $i = 1$ and $t = 4$. We have also $u(z) = 3z^2/(1-z^2)$. Furthermore $u(1/2) = 1$ and thus u satisfies Kraft's equality. However there are four edges going out of vertex 2 and the method described above fails to build a binary prefix code. A solution on $A = \{a, b\}$ is the regular prefix code*

$$X = (aa)^*(ab + ba + bb).$$

The corresponding automaton is given on Figure 4 on the right.

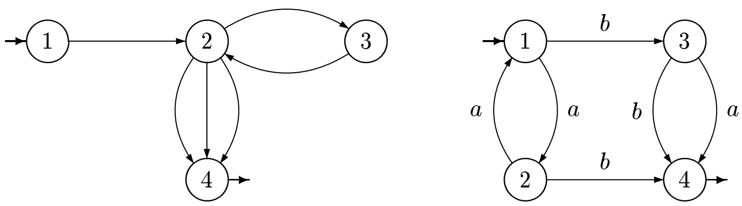

FIG. 4. *Graphs recognizing $u(z) = 3z^2/(1-z^2)$.*

The proof of Theorem 4.1 consists in building a new graph with all vertices of outdegree at most k. It relies on a transformation called the *multiset construction* described in [8]. The proof uses the following combinatorial lemma also used in symbolic dynamics by Adler and Marcus [28],[2], and quoted in [4] as a nice variant of the pigeon-hole principle.

LEMMA 4.1. *Let $k_1, k_2, \ldots, k_n$ be positive integers. Then there is a subset $S \subset \{1, 2, \ldots, n\}$ such that $\sum_{s \in S} k_s$ is divisible by n.*

The graph obtained is shown in an example below.

EXAMPLE 12. *Let*

$$u(z) = \frac{z^2}{1-z^2} + \frac{z^2}{1-5z^3}. \tag{4.1}$$

We have $u(1/2) = 1$. A regular binary tree with length distribution u is given in Figure 5 (note that, by convention, a vertex labeled v has its sons represented only once on the figure. Thus, for example the vertex labeled 1 on the right has the same sons as the root. The leaves of the tree are indicated by a black box).

To check that the length distribution is equal to u, one may compute from the graph the following regular expression of u and check by an elementary computation (possibly with the help of a symbolic computation system) that it is equal to u.

$$u(z) = (z^6)^*(2z^2 + z^4 + 2z^5 + z^6 + (z^2 + 3z^5)(5z^3)^*3z^3).$$

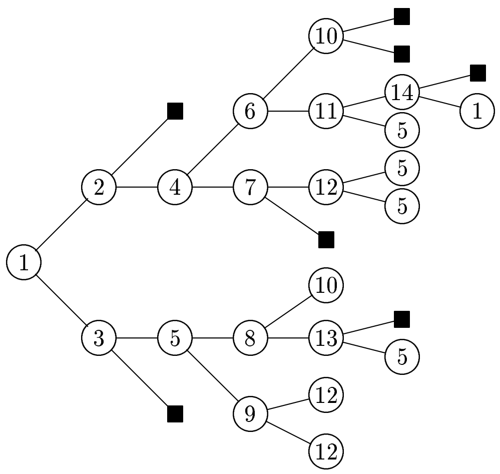

FIG. 5. *Regular binary tree with length distribution u.*

(note for a reader unfamiliar with regular expressions: the first factor $(z^6)^*$ corresponds to the vertex labeled 1 *at level* 6 *of the tree. The term $2z^2 + z^4 + 2z^5 + z^6$ corresponds to the leaves reached by a path which does not use a vertex labeled* 5*. The factor $(z^2 + 3z^5)(5z^3)^*$ corresponds to the paths from the root to a vertex labeled* 5*. Finally, the factor $3z^3$ corresponds to the direct paths from* 5 *to a leaf.)*

This example (suggested to us by Christophe Reutenauer) shows an interesting feature of this problem. In fact, from the point of view of regular expressions, the difficult operation in this problem is the sum. It would be a simple matter to build a rational tree for each term of the sum in the expression (12) (see Example 11). The difficulty would then be to merge these two trees to obtain one corresponding to the sum.

A curious consequence of Theorem 4.1 is the following property of regular sequences.

COROLLARY 4.1. *Let $k \geq 2$ be an integer and let u be regular sequence such that $u(1/k) \leq 1$ and $u(0) = 0$. Then there exist k regular sequences $u_1, \ldots, u_k$ such that $u_i(1/k) \leq 1$ and*

$$u(z) = \sum_{i=1}^{k} z u_i(z).$$

Proof. It is a simple consequence of Theorem 4.1. Indeed, if X is a regular prefix code on the k element alphabet A, then $X = \sum_{a \in A} aX_a$ where each X_a is a regular prefix code on the alphabet A. ◻

We don't know of a direct proof of this result.

5. Bifix codes. We investigate here the length distributions of a particular class of prefix codes, called bifix. Several other classes of prefix codes could give rise to a similar study (for a description to these classes, see [19]).

The definition of a suffix code is symmetric to the definition of a prefix code. It is a set of words X such that no element of X is a suffix of another one. The notion of a complete suffix code is also symmetric. A *bifix code* is a set X of words which is both a prefix and a suffix code.

Any set of words of fixed length is obviously a bifix code but there are more complicated examples.

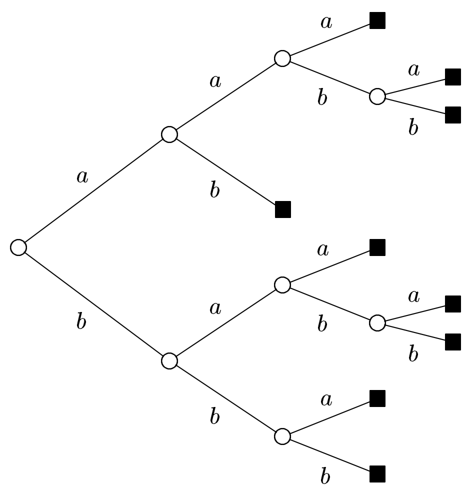

FIG. 6. *The bifix code X.*

EXAMPLE 13. *The set*

$$X = \{aaa, aaba, aabb, ab, baa, baba, babb, bba, bbb\}$$

is a complete prefix code pictured in Figure 6. It is also a complete suffix code as one may check by reading its words backwards.

Surprisingly, it is an open problem to characterize the length distributions of bifix codes. The following simple example shows that they are more constrained than those of prefix codes.

EXAMPLE 14. *The sequence $u(z) = z + 2z^2$ is not realizable as the length distribution of a bifix code on a binary alphabet although $u(1/2) = 1$. Indeed, one of the symbols has to be in X, say a. Then bb is the only word of length 2 that can be added.*

The following nice partial result is due to Ahlswede, Balkenhol and Khachatrian [3]. We state the result for a binary alphabet. It can be readily generalized to k symbols but it presents less interest.

THEOREM 5.1. *For any integer sequence u such that*

$$u(1/2) \leq 1/2,$$

there is a bifix code X such that $u = u_X$.

Proof. The proof is by induction. We suppose that we have already built a bifix code X formed of words of length at most $n-1$ with length distribution $(u_1, u_2, \dots, u_{n-1})$. We have

$$\sum_{i=1}^{n} u_i 2^{-i} \leq 1/2,$$

and thus

$$2\sum_{i=1}^{n} u_i 2^{n-i} \leq 2^n.$$

Finally, we obtain

$$u_n \leq 2^n - 2\sum_{i=1}^{n-1} u_i 2^{n-i}.$$

The expression of the right handside is at most equal to the number of elements of the set $A^n - XA^* - A^*X$. Thus, we can choose u_n words of length n which do not have a prefix or a suffix in X. This proves the result by induction. □

The authors of [3] formulate the interesting conjecture that Theorem 5.1 is still true if the hypothesis $u(1/2) \leq 1/2$ is replaced by $u(1/2) \leq 3/4$.

There are known additional conditions imposed on length distributions of bifix codes. For example, one has the following result, originally due to Schützenberger (see [14]).

THEOREM 5.2. *If X is a finite complete bifix code on k symbols, then $u_X(1/k) = 1$ and $\frac{1}{k}u'_X(1/k)$ is an integer.*

The number $\frac{1}{k}u'_X(1/k)$ can be interpreted as the average length of the words of X. Indeed

$$zu'_X(z) = \sum_{x \in X} |x| z^{|x|}.$$

EXAMPLE 15. *For the bifix code of Example 13, we have*

$$u_X(z) = z^2 + 4z^3 + 4z^4$$

and thus

$$u'_X(z) = 2z + 12z^2 + 16z^3.$$

Hence $\frac{1}{2}u'_X(1/2) = 3$. The conditions of Theorem 5.2 show directly that the sequence of Example 14 is not realizable. Indeed, it satisfies the first condition but not the second one. The conditions of Theorem 5.2 are not sufficient. Indeed, if $u(z) = z + 4z^3$ we have $u(1/2) = 1$ and $u'(1/2) = 4$ although it is clearly impossible that $u = u_X$ for a bifix code X.

6. Zeta functions, subshifts of finite type and circular codes. In this section, we present a number of results on interrelated objects which are connected with cyclic permutation of words. We begin with notions classical in symbolic dynamics (see [25] or [24] for a general reference; see [13] or [22] for the link with finite automata).

6.1. Subshifts of finite type. A *subshift* is a set of biinfinite words on a finite alphabet A which avoids a given set F of forbidden words. It is a topological space as a closed subset of the space $A^{\mathbb{Z}}$ of functions from $\mathbb{Z}$ into the set A. The *full shift* on A is the set of all biinfinite words on A. It corresponds to the case $F = \emptyset$.

A *sofic* subshift is the set of biinfinite labels of paths in a finite automaton. A sofic subshift is called *irreducible* if the automaton can be chosen strongly connected. A *subshift of finite type* is the set of biinfinite words avoiding a finite set of finite words. Any subshift of finite type is sofic but the converse is not true. The *edge shift* of a finite graph G is the set S_G of biinfinite paths in G (viewed as biinfinite sequences of edges). It is a subshift of finite type.

The *shift* σ is the function on a subshift S which maps a point x to the point $y = \sigma(x)$ whose ith coordinate is $y_i = x_{i+1}$.

A *morphism* from a subshift S into a subshift T is a function $f : S \to T$ which is continuous and invariant under the shift. A bijective morphism is called a *conjugacy*. Any subshift of finite type is conjugate to some edge shift.

The *entropy* $h(S)$ of a subshift S is the entropy of the formal language formed by the finite blocks occurring in words of S. It can be shown that the entropy is a topological invariant, in the sense that two conjugate subshifts have the same entropy.

While the entropy is a measure of number of forbidden words, it is possible to study the number of minimal forbidden words. It gives rise to another invariant of subshifts [11], [12].

An integer p is a *period* of a point $x = (a_n)_{n\in\mathbb{Z}}$ if $a_{n+p} = a_n$ for all $n \in \mathbb{Z}$. Equivalently, p is a period of x if $\sigma^p(x) = x$. The *zeta function* of a subshift S, is defined as the series

$$\zeta(S) = \exp \sum_{n\geq 1} \frac{p_n}{n} z^n$$

where p_n is the number of words with period n in S. It is also a topological invariant, since a point of period n is mapped by a conjugacy on a point of the same period.

The following result due to Bowen and Lanford [18] is classical (see [25]).

PROPOSITION 6.1. *Let G be a finite graph and let M be the adjacency matrix of G. Then*

$$\zeta(S_G) = \det(I - Mz)^{-1}.$$

Proof. We first have for each $n \geq 1$

$$\mathrm{Tr}(M^n) = p_n$$

since the coefficient (i,j) of M^n is the number of paths from i to j. Thus

$$\begin{aligned}\zeta(S_G) &= \exp \sum_{n\geq 1} \frac{p_n}{n} z^n \\ &= \exp \sum_{n\geq 1} \frac{\mathrm{Tr}(M^n)}{n} z^n \\ &= \exp \mathrm{Tr}(\log(I - Mz)^{-1}) \\ &= \det(I - Mz)^{-1}\end{aligned}$$

since, by the formula of Jacobi, $\exp \mathrm{Tr} = \det \exp$. □

EXAMPLE 16. *Let S be the edge shift of the graph G of Figure 7. We have*

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

Consequently

$$\zeta(S) = \frac{1}{1 - z - z^3}.$$

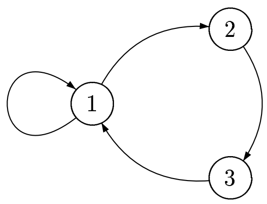

FIG. 7. *A subshift of finite type*

Let S be a subshift of finite type and let p_n be the number of points with period n. Let q_n be the number of points with least period n. Since q_n is a multiple of n, we also denote $q_n = nl_n$. We have then the formula expressing the zeta function as an infinite product using the integers l_n as exponents.

$$\zeta(S) = \Pi_{n\geq 1}(1 - z^n)^{-l_n},$$

as one may verify using $p_n = \sum_{d|n} dl_d$ and the definition of $\zeta(S)$.

A classical result, related with what follows, is the following statement, known as Krieger's embedding theorem.

THEOREM 6.1. *Let S, T be two subshifts of finite type. There exists an injective morphism $f : S \to T$ with $f(S) \neq T$ iff*

1. *$h(S) < h(T)$*
2. *for each $n \geq 1$, $q_n(S) \leq q_n(T)$ where $q_n(S)$ (resp. $q_n(T)$) is the number of points of S (resp. T) of least period n.*

The following result is the basis of many applications of symbolic dynamics to coding. It is due to Adler, Coppersmith and Hassner [2].

THEOREM 6.2. *If S is an irreducible subshift of finite type such that $h(S) \geq \log k$, it is conjugate to a subshift of finite type S_G where the graph G has outdegree at least k.*

The proof is based on a state-splitting algorithm using approximate eigenvectors and Lemma 4.1. This result is part of a number of constructions leading to sliding block codes used in magnetic recording (see [29], [9] or [25]). It gives at the same time the following result.

THEOREM 6.3. *It S is a subshift of finite type such that $h(S) \leq \log k$, then there is a graph G of outdegree at most k such that S is conjugate to S_G.*

There is a connexion between this theorem and Theorem 4.1. Let indeed u be a regular sequence of integers such that $u(1/k) \leq 1$. Let G be a normalized graph recognizing u (in the sense of Section 4). Let $\bar{G}$ be the graph obtained by merging the initial and terminal vertex. Then $h(S_{\bar{G}}) \leq \log k$. We can apply Theorem 6.3 to obtain a graph H with outdegree at most k such that S_G and S_H are conjugate. This gives the conclusion of Theorem 4.1 provided the initial-terminal vertex did not split in the construction. The following examples show both cases (for details, see [6] and [7]).

EXAMPLE 17. *Let G be the graph of Figure 4. The splitting of vertex 2 gives a graph of outdegree 2. A normalization gives the automaton on the right.*

EXAMPLE 18. *The sequence of Example 12 is recognized by a graph G such that $\bar{G}$ has three cycles of length 2. The solution as a binary tree has only two cycles of length 2 and thus could not be obtained by state-splitting.*

6.2. Circular codes. A *circular word*, or necklace, is the equivalence class of a word under cyclic permutation. For a word w, we denote by $\bar{w}$ the circular word represented by w.

Let X be a set of words and $w = x_1 x_2 \cdots x_n$ with $x_i \in X$. The set of cyclic permutations of the sequence $(x_1, x_2, \ldots, x_n)$ is called a factorization of the circular word $\bar{w}$.

A *circular code* is a set X of words such that the factorization of circular words is unique.

EXAMPLE 19. *The set $X = \{a, aba\}$ is a circular code. Indeed, the position of the symbols b determines uniquely the occurrences of aba.*

EXAMPLE 20. *The set* $X = \{ab, ba\}$ *is not a circular code. Indeed, the circular word* $\bar{w}$ *for* $w = abab$ *has two factorizations namely* (ab, ab) *and* (ba, ba).

The following characterization is useful (see [14]).

PROPOSITION 6.2. *A set* X *is a circular code if and only if it is a code and for all* $u, v \in A^*$,

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

EXAMPLE 21. *We obtain another way to prove that the set* $X = \{ab, ba\}$ *is not a circular code. Indeed, otherwise we would have* $a, b \in X^*$ *which is contradictory.*

Let X be a finite code. The *flower automaton* of X, denoted $\mathcal{A}_X$, is the following automaton. The set of its states is

$$Q = \{(u, v) \in A^+ \times A^+ \mid uv \in X\} \cup (1, 1)$$

The transitions are of the form $(u, av) \xrightarrow{a} (ua, v)$ or $(1, 1) \xrightarrow{a} (a, v)$ or $(u, a) \xrightarrow{a} (1, 1)$. The unique initial and final state is $(1, 1)$.

EXAMPLE 22. *The flower automaton of the circular code* $\{a, aba\}$ *is pictured in Figure 8.*

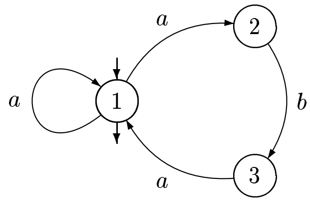

FIG. 8. *The flower automaton of* $\{a, aba\}$.

The following result is easy to prove.

PROPOSITION 6.3. *The flower automaton* $\mathcal{A}_X$ *recognizes* X^*. *The code* X *is circular iff for each word* w, *there is at most one cycle with label* w.

We now study the length distributions of circular codes. Let X be a circular code and let $u_(z) = (u_n)_{n\geq 1}$ be its length distribution. For each $n \geq 1$, let p_n be the number of words w of length n such that $\bar{w}$ has a factorization in words of X.

PROPOSITION 6.4. *The sequences* $(p_n)_{n\geq 1}$ *and* $(u_n)_{n\geq 1}$ *are related by*

$$\exp \sum_{n\geq 1} \frac{p_n}{n} z^n = \frac{1}{1 - u(z)}. \tag{6.1}$$

Proof. Each (p_n) depends only on the first n terms of the sequence (u_n). It is therefore possible to suppose that the sequence (u_n) is finite, i.e. that the code X is finite. Let $\mathcal{A}$ be the flower automaton of X. Let S be the subshift of finite type associated with the graph of $\mathcal{A}$. Then p_n is the number of elements of period n in S. Indeed, each word w such that $\bar{w}$ has a factorization is counted exactly once as the label of a cycle in $\mathcal{A}$. We have also

$$\det(I - Mz) = 1 - u(z).$$

Thus, the result follows from Proposition 6.1. □

The explicit relation between the numbers u_n and p_n is the following. For each $i \geq 1$, let $u^{(i)} = (u_n^{(i)})_{n\geq 1}$ be the length distribution of X^i. Equivalently, $u_n^{(i)}$ is the coefficient of degree n of $u(z)^i$. Then for each $n \geq 1$

$$p_n = \sum_{i=1}^{n} \frac{n}{i} u_n^{(i)}.$$

We also have for each $n \geq 1$

$$p_n = nu_n + \sum_{i=1}^{n-1} p_i u_{n-i}. \tag{6.2}$$

This formula can be easily deduced from Formula (6.1) by taking the logarithmic derivative of each side of the formula. It shows directly that for any sequence $(u_n)_{n\geq 1}$ of nonnegative integers, the sequence p_n defined by Formula (6.1) is formed of nonnegative integers.

Formula (6.2) is known as Newton's formula in the field of symmetric functions. Actually, the numbers u_n can be considered, up to the sign, as elementary symmetric functions and the p_n as the sums of powers (see [26]). The link between Witt vectors and symmetric functions was established in [34].

Let $p_n = \sum_{d|n} dl_d$. Then l_n is the number of non-periodic circular words of length n with a factorization. In terms of generating series, we have

$$\exp \sum_{n\geq 1} \frac{p_n}{n} z^n = \prod_{n\geq 1} (1 - z^n)^{-l_n}. \tag{6.3}$$

Putting together Formulae (6.1) and (6.3), we obtain

$$\frac{1}{1 - u(z)} = \prod_{n\geq 1} (1 - z^n)^{-l_n}. \tag{6.4}$$

For any sequence $(u_n)_{n\geq 1}$ of nonnegative integers, the sequence $l = (l_n)_{n\geq 1}$ thus defined is formed of nonnegative integers. This can be proved either by a direct computation or by a combinatorial argument since any sequence u of nonnegative integers is the length distribution of a circular code on a large enough alphabet. We denote $l = \phi(u)$ and we say that l is the *ϕ-transform* of the sequence u.

We denote by $\varphi_n(k)$ the number of non-periodic circular words of length n on k symbols. The numbers $\varphi_n(k)$ are called the *Witt numbers*. It is clear that the sequence $(\varphi_n(k))_{n\geq 1}$ is the ϕ-transform of the sequence $(k^n)_{n\geq 1}$.

The corresponding particular case of Identity (6.4)

$$1 - kz = \prod_{n\geq 1}(1 - z^n)^{\varphi_n(k)}$$

is known as the *cyclotomic identity*.

The following arrays display a tabulation of the Witt numbers for small values of n and k.

n	$\varphi_n(2)$	$\varphi_n(3)$	$\varphi_n(4)$
1	2	3	4
2	1	3	6
3	2	8	20
4	3	18	60
5	6	48	204
6	9	116	670
7	18	312	2340
8	30	810	8160
9	56	2184	29120
10	99	5880	104754

The value $\varphi_3(4) = 20$ is famous because of the genetic code: there are precisely 20 amino-acids coded by words of length 3 over a 4-symbol alphabet A,C,G,U.

For any sequence $a = (a_n)_{n\geq 1}$, let

$$p_n = \sum_{d|n} d a_d^{n/d}.$$

The pair (a, p) is called a *Witt vector* (see [30]). The numbers p_n are the *ghost components*. In terms of generating series, one has

$$\exp \sum_{n\geq 1} \frac{p_n}{n} z^n = \prod_{n\geq 1}(1 - a_n z^n)^{-1}.$$

The following result is due to Schützenberger (see [14]).

THEOREM 6.4. *Let $u = (u_n)_{n\geq 1}$ be a sequence of nonnegative integers and let $l = (l_n)_{n\geq 1}$ be the ϕ-transform of u. The sequence $(u_n)_{n\geq 1}$ is the length distribution of a circular code on k symbols iff for all $(n \geq 1)$*

$$l_n \leq \varphi_n(k).$$

Several complements to Theorem 6.4 appear in [5]. In particular, the relation with Kraft's inequality is studied. The equality case in Kraft's inequality is characterized in terms of the sequence of inequalities above.

There is a connexion between Theorem 6.4 and Krieger's embedding theorem (Theorem 6.1), in the sense that Theorem 6.4 gives a simple proof of Theorem 6.1 in a particular case. Actually, let us consider the particular case of subshift of finite type, called a *renewal system*.

A renewal system S is the edge shift of a graph G made up of cycles sharing exactly one vertex. Such a graph is determined by the sequence $u = (u_i)_{1\leq i\leq n}$ where u_i is the number of loops with length i. Let T_k be the full shift on k symbols. Suppose that the pair formed by S and T_k satisfies the hypotheses of Krieger's theorem. The number $q_n(S)$ of points of least period n is nl_n where $l = (l_n)_{n\geq 1}$ is the ϕ-transform of the sequence u and $q_n(T_k) = n\varphi_n(k)$. Thus, the sequence u satisfies the hypotheses of Theorem 6.4. Consequently, there is circular code X such that $u_X = u$. The flower automaton of X defines an embedding of S_G into the full shift on k symbols. This gives an alternative proof of Krieger's theorem in this case.

It would be interesting to have a proof of Krieger's theorem along the same lines in the general case.

To close this section, we mention the following open problem: If the sequence u is regular and satisfies the inequalities

$$l_n \leq \varphi_n(k) \qquad (n \geq 1),$$

where $l = \phi(u)$, does there exist a rational circular code on k symbols such that $u = u_X$?

6.3. Zeta functions. Theorem 6.1 admits the following generalization due to Reutenauer [32].

THEOREM 6.5. *The zeta function of a sofic subshift is regular.*

We have seen already (Theorem 6.1) that the zeta function of a subshift of finite type is a rational fraction, and indeed the inverse of a polynomial. The stronger statement that it is regular follows from the following formula allowing to compute $\det(I - Mz)$ when M is the adjacency matrix of a $n \times n$ graph G. One has

$$\det(I - Mz) = (1 - v_1(z)) \cdots (1 - v_n(z)),$$

where $v_i(z)$ is the length distribution of the set of first returns to state i using only states $\{i, i+1, \ldots, n\}$ (see [10]).

The proof that the zeta function of a sofic subshift is rational is a result of Manning and Bowen [27], [17]. For an exposition, see [25] or [10]. A generalization appears in [15].

7. Acknowledgments. The authors wish to thank for the help received during the preparation of this paper. We are indebted to Julia Abrahams for the reference of the work of Ahlswede et al. and several other recent references concerning bifix codes (see [1]). The link between length distributions of circular codes and symmetric functions was disclosed to us by Jacques Désarménien and Jean-Yves Thibon. We also thank Véronique Bruyère for improving our work.

REFERENCES

[1] J. Abrahams, *Code and parse trees for lossless source encoding*, in Compression and Complexity of Sequences 1997, B. C. et al., ed., IEEE Computer Society, 1998, pp. 145–171.

[2] R. L. Adler, D. Coppersmith, and M. Hassner, *Algorithms for sliding block codes*, IEEE Trans. Inform. Theory, IT-29 (1983), pp. 5–22.

[3] R. Ahlswede, B. Balkenhol, and L. Khachatrian, *Some properties of fix-free codes*, Tech. Rep. 039, University Bielefeld, 1997.

[4] M. Aigner and G. M. Ziegler, *Proofs from The Book*, Springer-Verlag, 1998.

[5] F. Bassino, *Generating functions of circular codes*, Adv. in Appl. Math, 22 (1999), pp. 1–24.

[6] F. Bassino, M.-P. Béal, and D. Perrin, *Enumerative sequences of leaves in rational trees*, in ICALP'97, no. 1256 in Lecture Notes in Computer Science, Springer-Verlag, 1997, pp. 76–86.

[7] ———, *Enumerative sequences of leaves and nodes in rational trees*, Theoret. Comput. Sci., (1999), pp. 41–60.

[8] ———, *A finite state version of version of Kraft-McMillan theorem*, SIAM J. Comput., (2000). To appear.

[9] M.-P. Béal, *Codage Symbolique*, Masson, 1993.

[10] ———, *Puissance extérieure d'un automate déterministe, application au calcul de la fonction fonction zêta d'un système sofique*, RAIRO Inform. Théor. Appl., 29 (1995), pp. 85–103.

[11] M.-P. Béal, F. Mignosi, and A. Restivo, *Minimal forbidden words and symbolic dynamics*, in STACS'96, C. Puech and R. Reischuk, eds., vol. 1046 of Lecture Notes in Computer Science, Springer-Verlag, 1996, pp. 555–566.

[12] M.-P. Béal, F. Mignosi, A. Restivo, and M. Sciortino, *Forbidden words in symbolic dynamics*, Tech. Rep. 99-15, I.G.M., Université de Marne-la-Vallée, 1999. To appear in Adv. in Appl. Math.

[13] M.-P. Béal and D. Perrin, *Symbolic dynamics and finite automata*, in Handbook of Formal Languages, G. Rosenberg and A. Salomaa, eds., vol. 2, Springer-Verlag, 1997, ch. 10.

[14] J. Berstel and D. Perrin, *Theory of Codes*, Academic Press, 1985.

[15] J. Berstel and C. Reutenauer, *Zeta functions of formal languages*, Trans. Amer. Math. Soc., 321 (1990), pp. 533–546.

[16] ———, *Rational Series and their Languages*, Springer-Verlag, 1998.

[17] R. Bowen, *On Axiom A diffeomorphisms*, in AMS-CBMS Reg. Conf., vol. 35, Providence, 1978.

[18] R. Bowen and O. E. Lanford, *Zeta functions of restrictions of the shift transformation*, in Proc. Symp. Pure Math. AMS, vol. 14, 1970, pp. 43–50.

[19] V. Bruyère and M. Latteux, *Variable-length maximal codes*, in Proc. 23rd Inter-

national Colloquium on Automata, Languages and Programming (ICALP'96), F. Meyer and B. Monien, eds., vol. 1099, Springer-Verlag, 1996, pp. 24–47.

[20] S. Eilenberg, *Automata,Languages and Machines*, vol. A, Academic Press, 1974.

[21] P. Flajolet, *Analytic models and ambiguity of context-free languages*, Theoret. Comput. Sci., 49 (1987), pp. 283–309.

[22] G. D. Forney, B. H. Marcus, N. T. Sindhushayana, and M. Trott, *A multilingual dictionary: System theory, coding theory, symbolic dynamics and automata theory*, in Proceedings of Symposia in Applied Mathematics, no. 50, 1995, pp. 109–138.

[23] R. L. Graham, D. Knuth, and O. Pataschnik, *Concrete Mathematics*, Addison Wesley, 1988.

[24] B. P. Kitchens, *Symbolic Dynamics*, Springer-Verlag, 1997.

[25] D. A. Lind and B. H. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge, 1995.

[26] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, Oxford University Press, 1995.

[27] A. Manning, *Axiom A difeomorphisms hava rational zeta functions*, Bull. London Math. Soc., 3 (1971), pp. 215–220.

[28] B. H. Marcus, *Factors and extensions of full shifts*, Monats. Math, 88 (1979), pp. 239–247.

[29] B. H. Marcus, R. M. Roth, and P. H. Siegel, *Constrained systems and coding for recording channels*, in Handbook of Coding Theory, V. S. Pless and W. C. Huffman, eds., vol. II, North Holland, 1998, ch. 20, pp. 1635–1764.

[30] N. Metropolis and G.-C. Rota, *Witt vectors and the algebra of necklaces*, Advances in Math., 50 (1983), pp. 95–125.

[31] D. Perrin, *Finite automata*, in Handbook of Theoretical Computer Science, J. van Leeuwen, ed., vol. B, Elsevier, 1990, ch. 1.

[32] C. Reutenauer, *$\mathbb{N}$-rationality of zeta functions*, Adv. in Appl. Math., 29 (1997), pp. 1–17.

[33] A. Salomaa and M. Soittola, *Automata Theoretic Properties of Formal Power Series*, Springer-Verlag, 1978.

[34] T. Scharf and J.-Y. Thibon, *On Witt vectors and symmetric functions*, Algebra Colloq., 3 (1996), pp. 231–238.