ZETA FUNCTIONS OF RECOGNIZABLE LANGUAGES

Jean Berstel and Christophe Reutenauer

LITP
4 place Jussieu
75005 Paris

Abstract Motivated by symbolic dynamics and algebraic geometry over finite fields, we define cyclic languages and the zeta function of a language. The main result is that the zeta function of a cyclic language which is recognizable by a finite automaton is rational.

1. Introduction

Motivated by algebraic geometry over finite fields and symbolic dynamics, we call zeta function of a formal language L the function

$$\zeta(L) = \exp\left(\sum a_n \frac{t^n}{n}\right)$$

where a_n is the number of words of length n in L. Moreover, we say that a language is cyclic if it is conjugation - closed ($u\,v \in L \Leftrightarrow v\,u \in L$) and if for any two words having a power in common, if one of them is in L, then so is the other.

Our main result states that if L is a cyclic language which is recognizable by a finite automaton (i.e regular), then its zeta function is rational (th. 1), and effectively computable, as the proof shows.

One consequence is that the zeta function of a sofic system in symbolic dynamics is rational, a fact which was claimed in [20] and [6]. Moreover, it is effectively computable, if the sofic system is given by a semigroup or a graph.

There is of course a striking analogy with one of the Weil ex-conjectures, stating that the zeta function of an algebraic variety over a finite field is rational (Dwork's theorem [7], see also [12]). In fact, several constructions allow to associate to each such variety a cyclic language (see Sect. 3); however, these constructions do not produce recognizable languages, so th.1 does not give a new proof of Dwork's theorem.

We prove in fact a structure theorem on cyclic recognizable languages: such a language is

(informally speaking) a linear combination of $\mathbb{Z}$ over traces of finite deterministic automata (th. 2).

This result is actually the difficult point; it has th.1 as a simple consequence, using a variant (prop.2) of a theorem of Bowen, Lanford [4] which rests essentially on Jacobi's identity det (m) = exp o tr o log(m), for any matrix m (when defined). The proof of th.2 is rather involved. It heavily relies on the theory of the minimal ideal of syntactic monoids, and will be published elsewhere.

A byproduct of theorem 2, which should be explored elsewhere, is that cyclic recognizable languages constitute a new class of languages which, as biprefix codes [17], have a semisimple syntactic algebra.

2. Definitions and main result

We assume the reader familiar with the elementary notions of finite automata theory (see [13], [8]). Recall that when L is a language contained in some free monoid A^*, then L is regular if and only if L is recognizable (by a finite automaton).

Let a_n be the number of words of length n in the language L. Then the usual generating function of L is

$$\sum_{n\geq 0} a_n t^n$$

It is a well-known result, which goes back to Chomsky and Schützenberger [5], that when L is recognizable by a finite automaton, then its generating function is a rational function.

Call zeta function of L the function

$$\zeta(L) = \exp\left(\sum a_n \frac{t^n}{n}\right)$$

This definition and the following ones will be motivated in the next section. This function is in general neither rational nor has integer coefficients, even if L is recognizable: for instance, for $L = \{a\}$ and $L = \{(ab)^n \mid n \in \mathbb{N}\}$, the zeta functions are respectively $\exp(t)$ and $1/\sqrt{1-t^2}$.

We say that a language is cyclic if for any words u, v, w and integer $n \geq 1$, the two following conditions hold:

(1) $uv \in L \Leftrightarrow vu \in L$

(2) $w \in L \Leftrightarrow w^n \in L$

Recall that two words x and y are said to be conjugate if for some words u and v, one has $x = uv$, $y = vu$. Hence, Eq. (1) means that L is conjugation-closed. Moreover, each word x in A^* is the power y^n of a unique primitive word y, which means that y is not a nontrivial power of another word (see e.g. [2] or [15]): y is called the primitive root of x. Then, Eq. (2) means that L is closed for the equivalence relation described by: x and y are equivalent iff x and y have the same primitive root iff x and y have a nontrivial power in common.

A first fact, rather classical (compare to [11] prop. 11.1.3), implies that the zeta function of a cyclic language has integer coefficients.

Proposition 1 Let L be a cyclic language. Then its zeta function has the infinite product expansion

$$\zeta(L) = \prod_{n\geq 1} \frac{1}{(1-t^n)^{\alpha_n}} \tag{3}$$

where α_n is the number of conjugation classes of primitive words contained in L. In particular, $\zeta(L)$ has integer coefficients.

Equivalently, α_n may be defined as the number of Lyndon words of length n in L (see e.g. [15]). For instance, when L is simply the whole free monoid A^*, then the α_n's are the Witt numbers, which count the Lyndon words, the homogeneous dimensions of free Lie algebras, the ranks of the quotients of the lower central series of a free group, the primitive necklaces and the irreducible polynomials over a finite field (see e.g. [15], [19]).

Proof We have to show that

$$\exp\Big(\sum_{n\geq 1} a_n \frac{t^n}{n}\Big) = \prod_{k\geq 1} \frac{1}{(1-t^k)^{\alpha_k}}$$

Take the logarithmic derivative of both members and multiply by x, obtaining

$$\sum_{n\geq 1} a_n t^n = \sum_{k\geq 1} \alpha_k \frac{k\, t^k}{1-t^k}$$

$$= \sum_{k\geq 1} \sum_{p\geq 1} \alpha_k k\, t^{kp}$$

This is equivalent to

$$a_n = \sum_{k|n} k\alpha_k$$

But this expresses the fact that each word x of length n is the power of a unique primitive word of length k dividing n and has k conjugates. □

Our main result is

Theorem 1 If L is a cyclic language which is recognizable by a finite automaton, then its zeta function is rational.

Example 1 Let L be the set of words on the alphabet {a, b, c} of the form

a . . . abc a . . . a bc a . . . . . a bc a . . . .a

with at least one occurence of bc, or of the form

c a . . . abc a . . .a bc a . . . . . .a bc a . . . ab

Then L is cyclic and recognizable. The number a_n of words of length n in L is $F_n - 1 + F_{n-2}$, where F_n is the n-th Fibonacci number. Hence, $a_n = \theta_n + \overline{\theta}_n - 1$ where

$$\theta = \frac{1+\sqrt{5}}{2}, \quad \overline{\theta} = \frac{1-\sqrt{5}}{2}.$$

Thus

$$\zeta(L) = \exp\left(\sum_{n\geq 1} (\theta^n + \overline{\theta}^n - 1)\frac{t^n}{n}\right)$$

$$= \exp\left(\sum \theta^n \frac{t^n}{n}\right) \exp\left(\sum \overline{\theta}^n \frac{t^n}{n}\right) \exp\left(\sum \frac{-t^n}{n}\right)$$

$$= \frac{1-t}{(1-\theta t)(1-\overline{\theta} t)} = \frac{1-t}{1-t-t^2}$$

Hence $\zeta(L)$ is rational.

It is known in general (see e.g. [11] prop. 11.1.1) that a series $\exp\left(\sum a_n \frac{t^n}{n}\right)$ is rational if and only if one has $a_n = \alpha_1 \lambda_1^n + \ldots + \alpha_k \lambda_k^n$ for some integers α_i and some complex numbers λ_i. This arithmetic approach is however impossible in our case.

As the proof of theorem 1 will show, the zeta function may be effectively computed. It should be noted that it is decidable if a given recognizable language L is cyclic: indeed, condition

(1) and (2) may be tested in the syntactic monoid of L, and this monoid is finite (see [13], [8]). Moreover, it is a well-known fact that the conjugation-closure of a recognizable language L (that is, the smallest conjugation-closed language containing L) is still recognizable; however, it is not true that the cyclic closure of a recognizable language is always recognizable: take for instance the language $a^* b^* = \{a^i b^j \mid i, j \geq 0\}$ and use the pumping lemma for recognizable languages (see e.g. [8] prop. 2.5.1) .

In order to prove theorem 1, we will prove a more general result, which gives some insight in the structure of cyclic recognizable languages.

Denote by $\mathbb{Z}\langle\langle A \rangle\rangle$ the set of noncommutative formal power series over $\mathbb{Z}$ on the alphabet A. Each language L defines a series, its characteristic series defined by

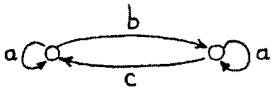
$$\underline{L} = \sum_{w \in L} w$$

Now, let A be a finite automaton over A, and define a formal power series, called the trace of A and denoted by tr(A), by

$$tr(A) = \sum_{w \in A^*} \alpha_w w$$

where the coefficient α_w of the word w is equal to the number of couples (q, c) where q is a state in A and c a path $q \to q$ in A labelled w.

Example 2 The trace of the automaton



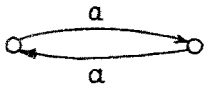
is the series having as coefficient of w

2 if w is a power of a

1 if w is a shuffle of a's and a word of the form $(bc)^i$ or $(cb)^i$, $i \geq 1$.

0 otherwise.

The trace of the automaton

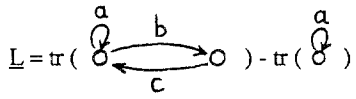


is $\sum_{n \geq 0} 2a^{2n}$.

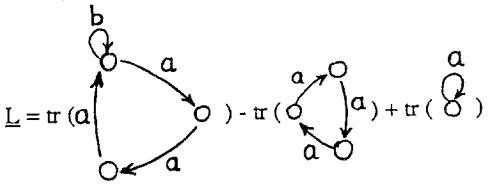
In what follows, a deterministic automaton will always be a non necessarily complete deterministic automaton.

Theorem 2 The characteristic series of each cyclic regular language is a linear combination over $\mathbb{Z}$ of traces of finite deterministic automata.

Example 3 For L = the language of example 1, one has



Now, let L be the set of words on the alphabet {a, b} such that between any two b's, the number of a's is a multiple of 3, even cyclically. Then



Theorem 2 will be proved in section 4. In the sequel of this section, we show how one may deduce theorem 1 from theorem 2. In fact, we shall prove a little bit more.

Let π: $\mathbb{Z}\langle\langle A\rangle\rangle \to \mathbb{Z}[[A]]$ be the natural homomorphism, where $\mathbb{Z}[[A]]$ is the usual commutative algebra of formal power series in the variables $a \in A$. Let $S \in \mathbb{Z}\langle\langle A\rangle\rangle$ be a noncommutative series. Then one has

$$S = \sum_{n \geq 0} S_n$$

where each S_n is the homogeneous part of S of degree n. Call generalized zeta function of S the commutative series

$$Z(S) = \exp\left(\sum_{n \geq 1} \frac{\pi(S_n)}{n} \right) \quad \in \quad \mathbb{Z}[[A]]$$

Note that if L is a language, then

$$\zeta(L) = \theta(Z(L))$$

where θ: $\mathbb{Z}[[A]] \longrightarrow \mathbb{Z}[[t]]$ is the homomorphism $\theta(a) = t$, for any letter a in A.

Hence it suffices to show that Z(L) is rational, under the hypothesis of theorem 1.

Call matrix of an automaton A the matrix E in $\mathbb{Z}[A]^{Q \times Q}$ (where Q is the set of states of A) defined by

$$E_{p,q} = \sum_{p \xrightarrow{a} q} a$$

where $p \xrightarrow{a} q$ means that there is an edge labelled a from p to q. Call determinant (cf. [20], [2] VIII.2) of A the polynomial in $\mathbb{Z}[[A]]$

$$\det (A) = \det (I - E)$$

where I is the Q x Q identity matrix.

Proposition 2 The generalized zeta function of the trace of a finite automaton is equal to the inverse of the determinant of this automaton.

Proof. Let A be a finite automaton. Then it is an easy consequence of a well-known fact in automata theory (see e.g. [8] prop. VI. 6.1) that

$$\pi(\mathrm{tr}\,(A)) = \mathrm{tr}\,(\sum_{n \geq 0} E^n) = \mathrm{tr}\,((I - E)^{-1})$$

Actually, this equality justifies the terminology "trace of an automaton". More precisely, let

$$\mathrm{tr}(A) = \sum_{n \geq 0} S_n$$

be the decomposition into homogeneous parts. Then

$$\pi(S_n) = \mathrm{tr}\,(E^n).$$

Hence the generalized zeta function of tr (A) is

$$Z = \exp\left(\sum_{n\geq 1} \frac{\pi(S_n)}{n}\right)$$

$$= \exp\left(\sum_{n\geq 1} \frac{1}{n} \operatorname{tr}(E^n)\right)$$

$$= \exp\left(\operatorname{tr} \sum_{n\geq 1} \frac{E^n}{n}\right)$$

$$= \exp\left(\operatorname{tr}\left(\log (I - E)^{-1}\right)\right)$$

Now, the Jacobi's identity tells us that det(M) = exp (tr (log (M))) for any matrix M where it is defined (instead of the Jacobi's identity, one may use e.g. [10] appendix, (lemma 4.1). Hence

$$Z = \det (I - E)^{-1}$$

which was to be shown. □

In order to deduce theorem 1, note that if $S = \sum \alpha_i S_i$ for some series S, S_i and integers α_i, then $Z(S) = \prod_i Z(S_i)^{\alpha_i}$. Thus theorem 1 may be deduced from theorem 2 and proposition 2. Note that the condition $\alpha_i \in \mathbb{Z}$ is crucial to rationality.

Example 4 For L = the first language of example 3, we have

$$Z(L) = \begin{vmatrix} 1-a, & -b \\ -c, & 1 \end{vmatrix}^{-1} \quad . \; |\, 1- a \,|$$

$$= \frac{1- a}{1- a - bc}$$

Its ordinary zeta function is

$$\zeta(L) = \frac{1 - t}{1 - t - t^2}$$

For the second one, we have

$$Z(L) = \begin{vmatrix} 1-b, -a, & 0 \\ 0, & 1, -a \\ -a, & 0, & 1 \end{vmatrix} \cdot \begin{vmatrix} -1 \end{vmatrix} \cdot \begin{vmatrix} 1, & -a, & 0 \\ 0, & 1, & -a \\ -a, & 0, & 1 \end{vmatrix} \cdot | 1 - a |^{-1}$$

$$= \frac{1 - a^3}{(1 - b - a^3)(1 - a)} = \frac{1 + a + a^2}{1 - b - a^3}$$

Its ordinary zeta function is

$$\zeta(L) = \frac{1 + t + t^2}{1 - t - t^3}$$

Another consequence of theorem 2 is that the syntactic algebra of each cyclic recognizable language is semi-simple (and finite dimensional): indeed, it suffices to apply prop. II.2.1(i) of [16]. Thus, cyclicity of a language is a combinatorial property which, as biprefixity, implies the semisimplicity of the syntactic algebra [17].

3. Motivations and applications

a. Let $\mathbb{F}_q$ be the finite field with q elements, $\mathbb{F}_{q^\infty}$ its algebraic closure and $f \in \mathbb{F}_q[x_1, .., x_k]$. Let V be the set of solutions in $\mathbb{F}_{q^\infty}$ of the algebraic equation

(4) $$f(x_1, ..., x_k) = 0$$

and let a_n be the number of those solutions which lie in the field $\mathbb{F}_{q^n}$. Then the zeta function of f is the series

$$\zeta(f) = \exp\left(\sum_{n \geq 1} a_n \frac{t^n}{n} \right)$$

It was one of the Weil conjectures, proved by Dwork [7], that this function is rational (more generally, the same holds for any algebraic variety V defined over $\mathbb{F}_q$).

Now, let A be the alphabet $A = (\mathbb{F}_q)^k$. Then there exist a mapping $\varphi: A^* \longrightarrow \mathbb{F}_{q^\infty}$ with the following properties:

(i) For any n, $\varphi \mid A^n$ is a bijection from A^n onto $\mathbb{F}_{q^n}$.

(ii) For any words u and v, $\varphi(u\,v)$ and $\varphi(vu)$ are conjugate points over $\mathbb{F}_q$.

(iii) For any word w and integer $n \geq 1$, $\varphi(w) = \varphi(w^n)$.

Such a mapping may be constructed using a family of primitive elements, one for each $\mathbb{F}_{q^n}$, following Golomb [9] ; or using a family of normal bases [14] of $\mathbb{F}_{q^n}$ (see e.g. [19]).

Note that an algebraic variety V defined over $\mathbb{F}_q$ (such as the set of solutions of Eq. (4)) is $\mathbb{F}_q$-conjugation - closed. Hence $L = \varphi^{-1}(V)$ will be a cyclic language, which encodes V, and the zeta function of L is equal to the zeta function of the algebraic variety V.

Unfortunately, no known mapping φ as above allows to obtain a recognizable language L: this would give a new proof of the rationality of the zeta function of V, which was one of the motivation of this paper. The construction of a mapping φ such that algebraic varieties correspond to recognizable languages is an open problem, certainly difficult, related to the construction of a natural bijection between irreducible polynomials over $\mathbb{F}_q$ of degree n and primitive necklaces over $\mathbb{F}_q$ of length n (see e.g. [19]).

b. Let A be a finite alphabet and $\sigma: A^{\mathbb{Z}} \longrightarrow A^{\mathbb{Z}}$ be the shift mapping, that is

$$\sigma: (a_n)_{n\in\mathbb{Z}} \longrightarrow (a_{n+1})_{n\in\mathbb{Z}}$$

If $S \subset A^{\mathbb{Z}}$ is closed under σ, then its zeta function is

$$\exp\Big(\sum_{n\geq 1} u_n \frac{t^n}{n}\Big)$$

where u_n is the number of points x in S such that $\sigma^n(x) = x$ (in other words, x is periodic and has n as a period). Call pattern of a periodic word $x = (a_n)_{n\in\mathbb{Z}}$ a word $w \in A^*$ such that p = length (w) is a period of x and that for some $n \in \mathbb{Z}$, one has $w = a_{n+1} \cdots a_{n+p}$. In other words

$$x = \ldots\ldots\ w\ w\ w \ldots\ldots w \ldots\ldots$$

with the origin 0 somewhere. Then associate to each σ-closed subset S of $A^{\mathbb{Z}}$ the language L of its patterns. Evidently, L is a cyclic language, whose zeta function is equal to that of S.

Now, let S be a sofic system [21]. Then its set of patterns is a recognizable language, which is a consequence of

Proposition 3 Let L be a recognizable language such that for any word w and integer $n \geq 1$, one has: $w \in L \Rightarrow w^n \in L$. Then the cyclic closure of L is recognizable. The set of patterns of the periodic words of a sofic system is recognizable.

All the constructions are effective, of course.

Proof. (i) Let $L = \mu^{-1}(P)$, where $\mu: A^* \longrightarrow M$ is the natural monoid homomorphism from A^* onto the (finite) syntactic monoid of M (see [13], [8]).

Let $P' = \{ m \in M \mid \exists\, p, q \in M,\ \exists\, n \geq 1, m = pq, (qp)^n \in P \}$

Then $L' = \varphi^{-1}(P')$ is a recognizable language, which is the cyclic closure $\underline{L}$ of L: indeed, if w is in $\underline{L}$, then for some u, v and n, one has $w = uv$ and $(vu)^n \in L$; this implies that $\varphi(w) = \varphi(u)\,\varphi(v)$, $(\varphi(v)\,\varphi(u))^n \in P$, hence $\varphi(w) \in P'$ and $w \in L'$; conversely, if $w \in L'$, then a power of w is conjugate to some word in L, thus w is in $\underline{L}$.

(ii) Let S be a sofic system. Then S is equal to the set of bi-infinite paths of a finite automaton A. Let L be the language consisting of all words w such that for some state q in A, there is a path $q \xrightarrow{w} q$. Then L is recognizable and satisfies to the condition of the proposition. Hence, its cyclic closure $\underline{L}$ is recognizable. But $\underline{L}$ is equal to the set of patterns of S. □

This proposition, together with theorem 1, shows that the zeta function of a sofic system is rational, a result which was claimed by Weiss [21] and Coven, Paul [6]. Moreover, it is computable as shows the proofs of theorem 1 and 2. The authors were told that M.-P. Béal had also obtained this result, in the particular case of an aperiodic semigroup defining the sofic system [1].

Actually, we show a little bit more: the generalized zeta function of a sofic system is rational; the latter function is a series in several variables which gives some information on the commutative composition of the patterns of the periodic words. Note also that when S is an irreducible sofic system, then the cyclic language L associated to it as above allows to recover S: indeed, the set of periodic points is then dense in S. Observe that proposition 2 is a variant of Bowen, Lanford's result [4], and that in the case of an irreducible subshift of finite type (i.e the set of bi-infinite paths

of a transitive graph), the inverse of the generalized zeta function is an irreducible polynomial, as shown in [18] (th. 3).

References

[1] M.-P. Béal, Personal communication (1987).

[2] J. Berstel, D. Perrin, The theory of codes, Acad. Press (1986).

[3] J. Berstel, C. Reutenauer, Rational series and their languages, Springer-Verlag (to appear).

[4] R. Bowen, O. Lanford, Zeta functions of restrictions of the shift transformation, Proc. Sympos. Pure Maths, 14 (1970) 43-50, A.M.S.

[5] N. Chomsky, M.P. Schützenberger, The algebraic theory of context-free languages, in: P. Braffort and D. Hirschberg (ed.), Computer Programming and formal systems, (1963) 118-161, North Holland.

[6] E.M. Coven, M.E. Paul, Sofic systems, Israel J. Maths 20 (1975) 165-177.

[7] B. Dwork, On the rationality of the zeta function of an algebraic variety, Amer. J. Math. 82 (1960) 631-648.

[8] S. Eilenberg, Automata, languages and machines, vol. A, Acad. Press (1974).

[9] S. W. Golomb, Irreducible polynomials, synchronization codes, primitive necklaces and the cyclotomic algebra, Univ. of North Carolina Monograph Series in Proba. and Stat. 4 (1967) 358-370.

[10] R. Hartshorne, Algebraic geometry, Springer-Verlag 1977.

[11] K. Ireland, M. Rosen, A classical introduction to modern number theory, Springer-Verlag (1982).

[12] N. Koblitz, P-adic numbers, p-adic analysis and zeta functions, Springer-Verlag 1984.

[13] G. Lallement, Semigroups and combinatorial applications, John Wiley 1979.

[14] R. Lidl, H. Niederreiter, Finite fields, Encycl. of Maths (1983), Addison Wesley.

[15] M. Lothaire, Combinatorics on words, Encycl. of Maths., Addison Wesley (1983).

[16] C. Reutenauer, Séries formelles et algèbres syntaxiques, J. Algebra 66 (1980) 448-483.

[17] C. Reutenauer, Semisimplicity of the algebra associated to a biprefix code, Semigroup Forum 23 (1981) 327-342.

[18] C. Reutenauer, Ensembles libres de chemins dans un graphe, Bull. Soc. Math. France 114 (1986) 135-152.

[19] C. Reutenauer, Mots circulaires et polynômes irréductibles, to appear.

[20] M.P. Schützenberger, Sur certains sous-monoïdes libres, Bull. Soc. Math. France 93 (1965) 209-223.

[21] B. Weiss, Subshifts of finite type and sofic systems, Monatsh. Math. 77 (1973) 462-474.