Minimal Forbidden Words and Symbolic Dynamics

Marie-Pierre Béal[1], Filippo Mignosi[2] and Antonio Restivo[2]

[1] LITP - Institut Blaise Pascal, Université Denis Diderot
2 Place Jussieu 75251 Paris cedex 05. `beal@litp.ibp.fr`
[2] University of Palermo, Dipartimento di Matematica ed Applicazioni
Via Archirafi 34. 90123 Palermo. `mignosi|restivo@altair.math.unipa.it`

Abstract. We introduce a new complexity measure of a factorial formal language L: the growth rate of the set of minimal forbidden words. We prove some combinatorial properties of minimal forbidden words. As main result we prove that the growth rate of the set of minimal forbidden words for L is a topological invariant of the dynamical system defined by L.

Classification: Automata and Formal Languages

1 Introduction

Let $L \subset A^*$ be a *factorial* language, i.e. a language containing all factors of its words. A word $w \in A^*$ is a *minimal forbidden word* for L if $w \notin L$ and all proper factors of w belong to L. We denote by $MF(L)$ the language of minimal forbidden words for L.

It turns out (as also stressed by the results of this paper) that the combinatorial properties of $MF(L)$ provide an usefull tool to investigate the structure of the language L or of the system that it describes. Consider, for instance, the case of locally testable factorial languages (cf [14]): they are characterized by the fact that the corresponding languages of minimal forbidden words are finite. In the context of Symbolic Dynamics they correspond to systems of finite type. Another example is given by a language L which is the set of factors of an infinite word (or of a set of infinite words): in this case, as we show in Sec. 2, the elements of $MF(L)$ are closely related to the *bispecial* factors (cf. [7], [8] and [6]) of the infinite word. Minimal forbidden words have been also considered in the study of complexity in the framework of a hierarchical modeling of physical systems (cf. [2] and [3]).

In this paper we are mainly concerned with the function $F_L(n)$, that counts, for any n, the number of words of $MF(L)$ of length n. Our main result is expressed in the framework of symbolic dynamics and it states, roughly speaking, that the growth of the function $F_L(n)$ is a topological invariant of the dynamical system defined by L.

There are deep connections between the theory of automata and formal languages and symbolic dynamics (cf. [16] and references therein). Several results from symbolic dynamics have a natural interpretation in terms of formal languages and conversely.

Let us recall some basic definitions. If A^Z is the set of all biinfinite sequences of elements of a finite alphabet A, endowed with the usual topology, a *symbolic* *dynamical system* is defined as a closed subset $\mathcal{S}$ of A^Z which is invariant under the shift function. It is easy to show that a system $\mathcal{S} \subseteq A^Z$ is uniquely specified by the associated language $L(\mathcal{S}) \subseteq A^*$ of finite factors of elements in $\mathcal{S}$. The language $L(\mathcal{S})$ is a factorial language. We denote by $F_{\mathcal{S}}(n) = F_{L(\mathcal{S})}(n)$ the function that counts the number of minimal forbidden words of $\mathcal{S}$ of length n.

Two symbolic dynamical systems are *conjugate* or *isomorphic* if there exists a bi-continuous bijection between them which commutes with the shift. A long standing open problem of the theory (cf. [18], [11] and [16]) is to decide whether two given sofic systems are isomorphic. Recall that a system $\mathcal{S}$ is *sofic* if $L(\mathcal{S})$ is a rational language.

A property of a system $\mathcal{S}$ that is preserved under isomorphism is said to be a *topological invariant* of $\mathcal{S}$. Several topological invariants have been found, like topological entropy and zeta functions. None of them is characteristic; for instance there exist systems $\mathcal{S}$ and $\mathcal{T}$ that are not isomorphic but that have same topological entropy and same zeta function (cf. next Example 1).

Let g, f be two functions from N to N. We say that f and g are *linearly equivalent*, and we write $f \simeq g$, if there exist two constants K_1 and K_2 such that

1) For any $n \geq K_1$, $f(n) \leq K_1 \sum_{i=-K_1}^{K_1} g(n+i)$.
2) For any $n \geq K_2$, $g(n) \leq K_2 \sum_{i=-K_2}^{K_2} f(n+i)$.

It is easy to verify that $\simeq$ is an equivalence relation.

Our main result states that if two symbolic dynamical systems $\mathcal{S}$ and $\mathcal{T}$ are isomorphic then the two functions $F_{\mathcal{S}}$ and $F_{\mathcal{T}}$, that count, respectively, the number of minimal forbidden words of $\mathcal{S}$ and of $\mathcal{T}$ are linearly equivalent; in other word, our main result states that the equivalence class of $F_{\mathcal{S}}$ is a topologial invariant of $\mathcal{S}$. This new invariant in not characteristic too. As a corollary we have that the topological entropy $H_{MF(L)}$ of the set of minimal forbidden words of L is a topological invariant of the dynamical system associated to L.

This result provides a new tool to show that some systems are not isomorphic. We give indeed examples of non-isomorphic sofic systems having the same zeta function and different growths of their minimal forbidden words. As a more special application of our main result, we show that two (non sofic) systems associated to Sturmian words having "different slope" are not isomorphic.

From the point of view of formal languages the fact that the growth of the function $F_L(n)$ is invariant under some natural transformations suggests that this function is a "good" tool to investigate combinatorial properties of factorial languages and, in particular, properties of infinite words. The growth of $F_L(n)$ provides a new complexity measure of the language L.

As a final remark we note that the statement of our main result presents some analogies with a theorem of Maurer and Nivat (cf. [13]) concerning rational bijections of rational languages. However the results are very different as explained with more details in the last section.

2 Minimal Forbidden Words

For any notation not explicitely defined in this paper we refer to [12] and to [9].

Let A be a finite alphabet and A^* the set of finite words of letters from A, the empty word ϵ included.

Let $L \subset A^*$ be a factorial language, i.e. $\forall u, v \in A^*$, $uv \in L \Rightarrow u, v \in L$. The complement $A^* \setminus L$ of L is a (two sided) ideal of A^*. Denote by $MF(L)$ its *base*: $A^* \setminus L = A^* MF(L) A^*$.

$MF(L)$ is the set of Minimal Forbidden words for L. A word $v \in A^*$ is forbidden for L if $v \notin L$. It is minimal if it has no proper factors that are forbidden.

Remark 2.1: The set $MF(L)$ uniquely characterizes L. Indeed $L = A^* \setminus A^* MF(L) A^*$.

Remark 2.2: A word $u = a_1 a_2 \cdots a_n \in MF(L)$ if and only if (1) u is forbidden, i.e. $u \notin L$, *and* (2) both $a_1 a_2 \cdots a_{n-1} \in L$ and $a_2 \cdots a_n \in L$.

From this last remark we derive the following expression of $MF(L)$: $MF(L) = AL \cap LA \cap (A^* \setminus L)$.

As a consequence of both previous remarks we have the following

Proposition 1. *The language $L \in Rat(A^*)$ if and only if $MF(L) \in Rat(A^*)$*

Example 1.

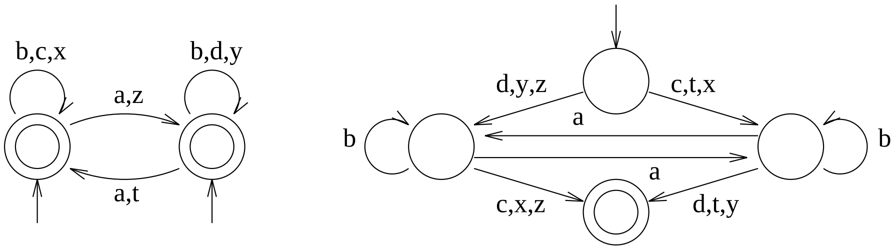

Figure 1. Languages L and MF(L)

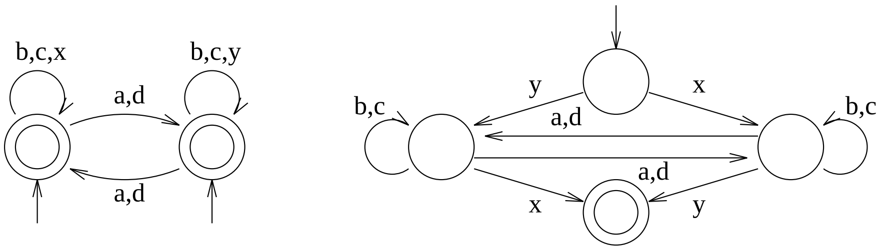

Figure 2. Languages S and MF(S)

A word $v \in L$ is *special on the left* with respect to B, where $B \subset A$ and *Card(B)*≥ 2, if for any $b \in B$, bv belongs to L. Anagously we define words *special on the right.* Given $B, C \subset A$ such that *Card(B)*≥ 2 and *Card(C)*≥ 2, we say that a word $v \in L$ is *bispecial* with respect to (B, C) if it is special on the left with respect to B and special on the right with respect to C.

In the case of a two letters alphabet A, special and bispecial words have been extensively studied (cf. [5], [7], [8], [6]). Remark that, since *Card(A)*$= 2$, there is no need to specify the sets B and C (both must be equal to A). In this case let us denote by $BS(L)$ the set of bispecial elements of L.

Example 2. Let K be the set of factors of the Fibonacci infinite word $\mathbf{f}$ (cf. [5]). Then

$$BS(K) = \{v|\ v \text{ is a palindrome prefix of } \mathbf{f}\} =$$

$$= \{\epsilon, a, aba, abaaba, abaababaaba, abaababaabaababaaba, \cdots\}.$$

$$MF(K) = \{w|w = bvb,\ v \text{ is the } n\text{-th palindrome prefix of } \mathbf{f},\ n \text{ is even}\} \cup$$

$$\cup\{w|w = ava,\ v \text{ is the } n\text{-th palindrome prefix of } \mathbf{f},\ n \text{ is odd}\} =$$

$$= \{bb, aaa, babab, aabaabaa, babaababaabab, aabaababaabaababaabaa, \cdots\}.$$

As can be seen by previous example, the sets $MF(K)$ and $BS(K)$ are "similar". This fact is motivated by the following two easy propositions, that relate bispecial and minimal forbidden words.

Proposition 2. *If $u \in A^*$ is bispecial with respect to (B, C) and $buc \notin L$ for some $b \in B$ and $c \in C$, then $buc \in MF(L)$.*

Proof. If u is bispecial with respect to (B, C) then $bu \in L$ and $uc \in L$. Since $buc \notin L$, by Remark 2, $buc \in MF(L)$.

The converse of previous proposition holds true under the supplementary hypothesis that L is *extensible*. Recall that a Language L is extensible if, for any $v \in L$ there exist $x, y \in A$ such that $xv \in L$ and $vy \in L$.

Proposition 3. *Let L be a factorial extensible language. If $w = buc \in MF(L)$, b, $c \in A$, then there exist $B, C \subset A$, Card(B)≥ 2, Card(C)≥ 2, $b \in B$ and $c \in C$ such that u is bispecial with respect to (B, C).*

Proof. We prove that u is left special with respect to B. A symmetric argument proves that u is right special with respect to C.

Since $buc \in MF(L)$, $uc \in L$. Since L is extensible there exists a letter x such that $xuc \in L$. This letter x is different from b because $buc \notin L$. If we set $B = \{x, b\}$ then it is easy to verify that u is left special with respect to B.

Given a language $L \subset A^*$, the *complexity function* $f_L(n)$ of L is defined as

$$f_L(n) = Card\{v \in L \text{ such that } |v| = n\}.$$

In this paper we study the complexity function of the set of minimal forbidden words $MF(L)$, $f_{MF(L)}(n)$ (that we denote by $F_L(n)$ in order to simplify the notation). It is often useful to consider also the formal series $g_L(z)$ defined as follows:

$$g_L(z) = \Sigma_{n>0} F_L(n) z^n.$$

A measure of the complexity of a language L is given by its *topological entropy* H_L, defined as follows:

$$H_L = \limsup_{n \to \infty} \log_2(\sqrt[n]{f_L(n)}).$$

It is known that for a factorial extensible language previous lim sup can be substituted by lim.

Example 3. If L and S are the languages of Example 1, then $H_L = H_S$ while $1 = H_{MF(L)} \neq H_{MF(S)} = 2$. Indeed $g_L(z) = \frac{18z^2(1-z^2)}{1-2z}$ and $g_S(z) = \frac{2z^2(1-4z^2)}{1-4z}$.

If K is the set of factors of the infinite Fibonacci word (cf. Example 2) then for any $n \geq 0$, $f_K(n) = n + 1$ and for any $n \geq 2$, $F_K(n) = 1$ if n is a fibonacci number, $F_K(n) = 0$ otherwise. Consequently $H_K = 0$ and $H_{MF(K)} = 0$.

3 Symbolic Dynamics

Symbolic dynamics is a field born at the beginning of the years 20 with the work in topology of Marston Morse (cf. [15]). Later the theory was developed as a branch of ergodic theory. There are deep connections between the theory of automata and formal languages and symbolic dynamics (cf. [16] and references therein). Several results from symbolic dynamics have a natural interpretation in terms of formal languages and conversely.

We present in this section a short introduction to the concepts of symbolic dynamics. Basic definitions and notations are from [16]. For any other notations not explicitely defined here we refer to [4].

Let A^Z be the set of bi-infinite sequence of letters of an alphabet A. An element of A^Z, $\mathbf{x} = (x_n)_{n \in Z}, x_i \in A$ is called an *infinite word.*

Let us endow A with the discrete topology and A^Z with the product topology. Then A^Z is a compact space.

The *shift* σ is a function defined on A^Z. It associates to $\mathbf{x}$ the element $\mathbf{y} = \sigma(\mathbf{x})$ defined by the rule: $y_n = x_{n+1}$ for any integer n.

A *symbolic dynamical system* $\mathcal{S}$ is a closed subset of A^Z that is invariant by the shift function; i.e. $\sigma(\mathcal{S}) = \mathcal{S}$.

Given a system $\mathcal{S}$, the set $L(\mathcal{S})$ of all factors of all infinite words in $\mathcal{S}$ is a factorial and extensible language. Conversely it is possible to associate to a factorial and extensible language L a system $\mathcal{S}(L)$ composed by all the infinite words $\mathbf{x}$ such that all the factors of $\mathbf{x}$ belong to L. It is easy to see that the system associated to $L(\mathcal{S})$ is again $\mathcal{S}$ and that $L(\mathcal{S}(L))$ is again L. Hence any system $\mathcal{S}$ is uniquely specified by the associated language $L(\mathcal{S})$. We will write $F_{\mathcal{S}}$ instead of $F_{L(\mathcal{S})}$. The entropy $H_{\mathcal{S}}$ of a system $\mathcal{S}$ is the topological entropy of the corresponding language $L(\mathcal{S})$.

A *morphism* between two systems $\mathcal{S}$ and $\mathcal{T}$ is a map ϕ: $\mathcal{S} \longrightarrow \mathcal{T}$ which is continuous and commutes with the shift, i.e. such that $\sigma \circ \phi = \phi \circ \sigma$.

An *isomorphism* is a bijective morphism. A property of a system $\mathcal{S}$ that is preserved under isomorphism is said to be a *topological invariant* of $\mathcal{S}$.

Let k be an integer greater or equal to 1. A k-*block* map (or k-*local* function) ψ from a system $\mathcal{S}$ into a system $\mathcal{T}$ is defined by a partial function f from A^k into B, by two integers m (for memory) and a (for anticipation) with $m + a = k$, and satisfies, for any integer i, $\psi((a_n)_{n \in Z})_i = f(a_{i-m+1} a_{i-m+2} \dots a_{i-1} a_i a_{i+1} \dots a_{i+a})$. A k-block map is continuous and commutes with the shift. The Curtis-Hedlund theorem (cf. [10]) assures that any continuous map from A^Z into B^Z that commutes with the shift is a k-block map for some integer k.

A *finite automaton* is a couple (Q, F) of a finite set of states Q and of a set of edges F labeled by letters of A. All states are both initial and terminal.

A *sofic* system (or *sofic shift*) can be defined as the set of bi-infinite labels of bi-infinite paths on a finite automaton. We say that the automaton recognizes the sofic system. A long standing open problem of the theory (cf. [18], [11] and [16]) is to decide whether two given sofic systems are isomorphic. We should also mention that there is a strong connection between sofic systems and the theory of constrained coding, a sub-field of coding theory. In terms of codes an isomorphism is a sliding-block code which is one-to-one and onto (cf. [1] and [4]).

An automaton is *local* if it does not admit two distinct equally labeled cycles. For any local automaton, there exist two integers m and a (m for memory and a for anticipation) such that two equally labeled paths of length $m + a$ go through the same state after their beginnings of length m. Such an automaton is said to be (m, a)-local.

Systems recognized by a local automata are called systems *of finite type*. A system of finite type can be also defined as a system consisting of all infinite words avoiding a finite set of words. Clearly, any system of finite type is also a sofic system. One can prove that the notion "of finite type" and the notion of "sofic" are invariant under isomorphism in the sense that any system that is isomorphic to a system of finite type (resp. sofic) is again of finite type (resp. sofic). Remark that the isomorphism problem is open even for the systems of finite type.

4 Main Result

In this section we state and prove our main result. We begin with a definition:

Let g, f be two functions from N to N. We say that f and g are *linearly equivalent*, and we write $f \simeq g$, if there exist two constant K_1 and K_2 such that

1) For any $n \geq K_1$, $f(n) \leq K_1 \sum_{i=-K_1}^{K_1} g(n+i)$.
2) For any $n \geq K_2$, $g(n) \leq K_2 \sum_{i=-K_2}^{K_2} f(n+i)$.

It is easy to verify that $\simeq$ is an equivalence relation. In the following we will consider two systems $\mathcal{S}$ and $\mathcal{T}$, $\mathcal{S}$ over the alphabet A, and $\mathcal{T}$ over the alphabet B. Let $F_{\mathcal{S}}$ and $F_{\mathcal{T}}$, be the functions that count, respectively, the number of minimal forbidden words of $\mathcal{S}$ and of $\mathcal{T}$.

We can now state our main result.

Theorem 4. *If two symbolic dynamical systems $\mathcal{S}$ and $\mathcal{T}$ are isomorphic then the two functions $F_{\mathcal{S}}$ and $F_{\mathcal{T}}$ are linearly equivalent.*

In other word, our main result states that the equivalence class of $F_{\mathcal{S}}$ is a topologial invariant of $\mathcal{S}$.

Since $\mathcal{S}$ and $\mathcal{T}$ are isomorphic then there exists a k-block map ψ partially defined from A^Z to B^Z such that ψ is an isomorphism between $\mathcal{S}$ and $\mathcal{T}$. Without loss of generality (eventually by using a composition with a power of the shift), we can assume that the map has memory k and no anticipation.

A *transducer* is an automaton (Q, F) with a set of edges F labeled by $A \times B$, where A and B are two finite alphabets; we, as usual, write a label $a|b$, $a \in A, b \in B$ instead of writing the ordered couple (a, b). The input of the transducer is the automaton labeled by A and obtained from the transducer after removing the second components of the labels of the edges. The output of the transducer is obtained from the transducer after removing the first components of the labels of the edges. It is an automaton labeled by B.

All block maps ψ defined by a partial function f from A^k into B, with integers $m = k$ and $a = 0$ can be represented, for any integer n with $n \geq k$, by a transducer labeled by $A \times B$, $T_n = (Q, F)$ where $Q = \{(a_1 a_2 \ldots a_n), a_i \in A\}$ and F is the set of the edges :

$$(a_1 a_2 \ldots a_n) \xrightarrow{a|f(a_{n-k+2} \ldots a_n a)} (a_2 a_3 \ldots a_n a),$$

for all $a_i, a \in A$. We can remark that if $(u|v)$ is the label of a path of a transducer, then $(\sigma(u)|\sigma(v))$ also.

We say that T_n is a representation of ψ: any path of T_n labeled $(u|v)$ with $u \in \mathcal{S}$ verifies $v = \psi(u)$. Of course it is not true that any input label of a path of T_n belongs to $\mathcal{S}$.

Remark also that the input automaton of the transducer T_n is a local automaton; it has memory n and no anticipation. This input automaton is usually known as a De Bruin graph.

We now define, for any $n \geq k$, the transducer T'_n obtained from T_n by removing all states defined by forbidden words of length n of $\mathcal{S}$.

We will first state and prove the following four lemmas:

Lemma 5. *For any state of T'_n, there exists a bi-infinite path of T'_n going through it.*

Proof. Let $u = a_1 \ldots a_n$ be a state of T'_n. The word u is a factor of a bi-infinite word w of S. There exists a path of T_n labeled by w in its input. Then there exists also a path of T'_n labeled by w in its input since w belongs to S. This path goes through the state u at a time because u is a factor of length n of w. This proves also that any finite path of T'_n can be extended to the infinity on the left and on the right.

Lemma 6. *There exists an integer n_0 such that for any $n \geq n_0$, T'_n has a local output.*

Proof. We will consider in the following only integers n greater or equal to k. As the input of T_n is local, there exists at most one bi-infinite path of the input of T_n with a given label. The input of T'_n is also a local automaton. To show that the output of T'_n is local for n big enough, it is sufficient to show that it is impossible to have the following situation:
For any integer N, there exists an integer $n > N$ and two bi-infinite paths of T'_n labeled by $((a_i^n)_{i \in Z} | (x_i^n)_{i \in Z})$ and $((b_i^n)_{i \in Z} | (y_i^n)_{i \in Z})$, with $(a_i^n)_{i \in Z} \neq (b_i^n)_{i \in Z}$ and $(x_i^n)_{i \in Z} = (y_i^n)_{i \in Z}$. We get a sequence of these bi-infinite paths, indexed by $P = \{p(n), n \in \mathbf{N}\}$, where p is an increasing function from $\mathbf{N}$ to $\mathbf{N}$.

Let us assume that the above sentence is true. For each integer $n \in P$, there exists an index q such that $a_q^n \neq b_q^n$. Without loss of generality, we can assume that $q = 0$ for each n by shift invariance of the property of being label of a path of T'_n. We have now an infinite number of integers n and bi-infinite paths of T'_n labeled by $((a_i^n)_{i \in Z} | (x_i^n)_{i \in Z})$ and $((b_i^n)_{i \in Z} | (y_i^n)_{i \in Z})$, with $a_0^n \neq b_0^n$ and $(x_i^n)_{i \in Z} = (y_i^n)_{i \in Z}$. As A^Z is a compact set, we extract from each sequence indexed by P, $((a_i^n)_{i \in Z})_{n \in P}$ and $((b_i^n)_{i \in Z})_{n \in P}$, a convergent sequence. After renumbering (by changing the set P), we can assume now that these sequences converge respectively to the two sequences $(a_i)_{i \in Z}$ and $(b_i)_{i \in Z}$.

We also denote by w_n the word of length n : $a^n_{-\lfloor \frac{n}{2} \rfloor + \delta_n} \dots a^n_1 a^n_0 a^n_1 \dots a^n_{\lfloor \frac{n}{2} \rfloor}$, where δ_n is equal to 1 if n is even and is equal to 0 otherwise.

As the word w_n is an input label of length n of a path of T'_n, it is not a forbidden word of S. As $((a^n_i)_{i \in Z})_{n \in P}$ converge to $(a_i)_{i \in Z}$, any finite word factor of $(a_i)_{i \in Z}$ is a factor of a word of S because it is a factor of a word w_n for some integer n. The bi-infinite sequence $(a_i)_{i \in Z}$ is then a limit of sequences of S. As a consequence the sequences $(a_i)_{i \in Z}$ and $(b_i)_{i \in Z}$ belong to S, by compactnes of the symbolic system S.

As ψ is a continuous map, the outputs $(x^n_i)_{i \in Z}$ and $(y^n_i)_{i \in Z}$ also converge to $(x_i)_{i \in Z}$ and $(y_i)_{i \in Z}$, and we have $(x_i)_{i \in Z} = \psi((a_i)_{i \in Z})$, $(y_i)_{i \in Z} = \psi((b_i)_{i \in Z})$. As $(x^n_i)_{i \in Z} = (y^n_i)_{i \in Z}$ for all $n \in P$, we have $(x_i)_{i \in Z} = (y_i)_{i \in Z}$. But $(a_i)_{i \in Z} \neq (b_i)_{i \in Z}$ because we have $a_0 \neq b_0$ since $a^n_0 \neq b^n_0$ for all $n \in P$. This contradicts the fact that ψ is one-to-one from S to T.

Lemma 7. *We assume that T'_{n_0} has a (m, a)-local output, where m and a are chosen greater than or equal to k and n_0. Then for each integer $n \geq n_0$, T'_n has a $(m+(n-n_0), a)$-local output.*

Proof. We fix an integer $n \geq n_0$. We assume that T'_n is (k_1, k_2)-local in its output. We assume that k_1 and k_2 are greater or equal to k and that k_1 is greater or equal to n. We just have to prove that T'_{n+1} is $(k_1 + 1, k_2)$-local in its output.
We consider two paths of T'_{n+1} of length k_1+1+k_2, with the word u as output label. We assume that the first path (resp. the second one) goes through state p (resp. p') after the first $k_1 + 1$ symbols. We want to show that $p = p'$. These two finite paths can be extended to two bi-infinite paths of T'_{n+1} (by Lemma 5). The first bi-infinite path is labeled by (w, z), the second one by (w', z'). We have $u = z_{i+1} \dots z_{i+k_1+k_2+1}$ and $u = z'_{i+1} \dots z'_{i+k_1+k_2+1}$ for some index i. The labels (w, z) and (w', z') are also labels of paths of T'_n because all labels of path of T'_{n+1} are labels of paths of T'_n. As T'_n is (k_1, k_2)-local in its output, the two bi-infinite paths of T'_n go through the same state q after the symbol of index $i + k_1$, and they also go through the same state r after the symbol of index $i + k_1 + 1$. Let $q = a_1 \dots a_n$ and $r = a_2 \dots a_{n+1}$. As T'_n is $(n, 0)$-local in its input, the input symbols of length $n+1$, $w_{i+k_1-n+1} \dots w_{i+k_1+1}$ and of $w'_{i+k_1-n+1} \dots w'_{i+k_1+1}$ are a fixed word $a_1 \dots a_{n+1}$. Back to T'_{n+1}, we also have $w_{i+k_1-n+1} \dots w_{i+k_1+1} = p$ and $w'_{i+k_1-n+1} \dots w'_{i+k_1+1} = p'$. We get $p = p'$.

Lemma 8. *Any word of length $n - k + 1$ which is an output label of a path of T'_n is a factor of a word of $\mathcal{T}$.*

Proof. We consider a word u of length $n-k+1$, output label of a path of T'_n. Then by Lemma 5 there exists a bi-infinite path of T'_n labeled (w, z) such that, for an index i, we have $u = z_{i+k} \dots z_{i+n}$. The word $w_{i+1} \dots w_{i+n}$ is a word of length n input label of a path of T'_n. It is then a factor of a word of $\mathcal{S}$. Then there exists a path in T'_n, labeled (w', z') with $w' \in \mathcal{S}$ and such that $w_{i+1} \dots w_{i+n} = w'_{i+1} \dots w'_{i+n}$. As w' belongs to $\mathcal{S}$, z' belongs to $\mathcal{T}$. By definition of T'_n, we have $z_{i+k+j} = f(w_{i+j+1} \dots w_{i+j+k})$, for $j = 0 \dots (n-k)$. We get $u = z_{i+k} \dots z_{i+n} = z'_{i+k} \dots z'_{i+n}$. Then u is a factor of $\mathcal{T}$ since z' belongs to $\mathcal{T}$.

We have now all notations and elements to prove the fundamental lemma:

Lemma 9. *If $\mathcal{S}$ and $\mathcal{T}$ are isomorphic, for each integer n greater than n_0, we have*

$$F_{L(\mathcal{S})}(n) \leq K \sum_{c=-k+1}^{m+a-n_0} F_{L(\mathcal{T})}(n+c)$$

where $K = (2\mathtt{card}(B))^l$ and $l = m + a - n_0 + k - 1$.

Proof. Let x be a minimal forbidden word of length n of $\mathcal{S}$. It is an input label of a path of T'_{n-1} labeled (w, z), which is not the label of a path of T'_n. Let us assume that $x = w_1 w_2 \dots w_n$. We consider the part of the output label of this path $y = z_{-(m-n_0-1)} \dots z_0 z_1 \dots z_n z_{n+1} \dots z_{n+a}$. The word y is of length $m + (n - n_0) + a$. Let us assume that y is a factor of $\mathcal{T}$. It is then the output label of a path of T'_{n-1} labeled (w', z') with $w' \in \mathcal{S}$, $z' \in \mathcal{T}$ and $y = z'_{-(m-n_0-1)} \dots z'_0 z'_1 \dots z'_n z'_{n+1} \dots z'_{n+a}$. As T'_{n-1} is $(m + (n - 1 - n_0), a)$-local in its output, we have $w'_1 w'_2 \dots w'_n = w_1 w_2 \dots w_n = x$. This contradicts the fact that x is not a factor of $\mathcal{S}$. Then y is not a factor of $\mathcal{T}$. To each minimal forbidden word x of length n we associate a word y defined as above. The previous argument shows also that the words y associated to two distinct words x are themselves distinct.
By Lemma 8 we know that each word of length $n - k$, which is an output label of a path of T'_{n-1}, belongs to $\mathcal{T}$. So each factor of length $n - k$ of y is a factor of a word of $\mathcal{T}$ and y is not factor of $\mathcal{T}$. Each word y can be obtained with a minimal forbidden word m_y of $\mathcal{T}$ of length between $n - k + 1$ and $m - n_0 + n + a$, completed left or right by symbols of B. Now the number of such words y, greater or equal to the number of minimal forbidden words of $\mathcal{S}$ of length n, is less or equal to

$$K \sum_{c=-k+1}^{m+a-n_0} F_{L(\mathcal{T})}(n+c)$$

This gives the announced formula.

In order to prove our main result we have to prove that there exists two constant K_1 and K_2 such that

1) For any $n \geq K_1$, $F_{L(\mathcal{S})}(n) \leq K_1 \sum_{i=-K_1}^{K_1} F_{L(\mathcal{T})}(n+i)$.
2) For any $n \geq K_2$, $F_{L(\mathcal{T})}(n) \leq K_2 \sum_{i=-K_2}^{K_2} F_{L(\mathcal{S})}(n+i)$.

If we set $K_1 = max\{K, m + a - n_0, k - 1\}$ then inequality (1) is an immediate consequence of previous lemma. Inequality (2) comes also by previous lemma, by exchanging the roles played by $\mathcal{S}$ and $\mathcal{T}$.

5 Applications

In this section we present some consequences and applications of our main result.

Let us first remark that, as a particular case of Theorem 4, we obtain the following well known result of Symbolic Dynamics concerning systems of finite type. Recall that a system $\mathcal{S}$ is of finite type if $L(\mathcal{S})$ is a locally testable language.

Corollary 10. *Let $\mathcal{S}$ be a system of finite type and let $\mathcal{T}$ be a system isomorphic to $\mathcal{S}$. Then $\mathcal{T}$ is also of finite type.*

It is well known that the entropy of a system is a topological invariant. Theorem 4 allows us to state that the entropy of minimal forbidden words is an invariant too.

Corollary 11. *Let $\mathcal{S}$ be a dynamical system. $H_{MF(L(\mathcal{S}))}$ is a topological invariant.*

Theorem 4 and Corollary 11 provide useful tools to prove that two systems are not isomorphic. In particular, if $\mathcal{S}$ is a sofic system, i.e. $L(\mathcal{S})$ is a rational language, by Proposition 1, $MF(L(\mathcal{S}))$ is also a rational language and its entropy can be easily computed. For instance the two sofic systems in Example 1 are not isomorphic, because $H_{MF(L(\mathcal{S}))} \neq H_{MF(L(\mathcal{T}))}$, but they have same entropy (and also same zeta function).

However there exist sofic systems that have different zeta functions (and hence are non isomorphic) but that have linearly equivalent growths of minimal forbidden words. Moreover there exist sofic systems that have equal zeta functions and that have linearly equivalent growths of minimal forbidden words but that are non isomorphic.

Let us now consider the case of dynamical systems associated to Sturmian words. Recall that a Sturmian word can be also defined by considering the intersections with a squared-lattice of a semi-line having a slope which is an irrational number $\alpha > 0$ (cf. [7]). A vertical intersection is denoted by the letter a, a horizontal intersection by b and the intersection with a corner by ab or ba. It is possible to prove that the language of factors of a Sturmian word defined in this way (and hence the associated system $\mathcal{S}_\alpha$) depends only on the slope of the line.

Let $\mathcal{S}_\alpha$, $\mathcal{S}_\beta$ be the dynamical systems associated to two Sturmian words such that $\alpha \neq \beta$ and $\alpha \neq \frac{1}{\beta}$. It is possible to verify that the entropy of both systems and the entropy of their minimal forbidden words are zero. However in what follows, as a consequence of Theorem 4, we prove in a purely combinatorial way that that $\mathcal{S}_\alpha$, $\mathcal{S}_\beta$ are not isomorphic; this also shows that Theorem 4 is stronger than Corollary 11.

Before stating next theorem we need some more preliminaries.

We construct the infinite sequence of pairs of words (A_n, B_n), $n \geq 0$, as follows. We set $(A_0, B_0) = (a, b)$. For any $n \geq 0$ the pair (A_{n+1}, B_{n+1}) is obtained from (A_n, B_n) by using one of the following two rules:

First rule: $(A_{n+1}, B_{n+1}) = (A_n, A_n B_n)$.

Second rule: $(A_{n+1}, B_{n+1}) = (B_n A_n, B_n)$.

Let $\alpha > 0$ be an irrational number and let $[q_0, q_1, q_2, \cdots]$ be the development in continued fraction of α. It has been proved (cf. [17]) that if one applies q_0 times first rule, q_1 times second rule, q_2 times first rule, and so on, then the sequences $\{A_n\}_{n\geq 0}$, $\{B_n\}_{n\geq 0}$ converge to the same infinite Sturmian word $\mathbf{x}_\alpha$ that is associated to the semi-line that have slope α and that starts from the origin. Moreover, if $|A_n| \geq 2$ (resp. $|B_n| \geq 2$) then A_n (resp. B_n) is a prefix of $\mathbf{x}_\alpha$.

Let us define $A(\mathbf{x}_\alpha) = \bigcup_{n\geq 0}\{A_n\}$ and $B(\mathbf{x}_\alpha) = \bigcup_{n\geq 0}\{B_n\}$. Finally, for any integer $k \geq 1$ let $I_\alpha(k) = \{p \in \mathbb{Z}, p \geq k |$ there exists $u \in A(\mathbf{x}_\alpha) \cup B(\mathbf{x}_\alpha)$ such that $p = |u|\}$.

In what follows α and β are two positive irrational number.

The proof of the following lemma is left to the reader.

Lemma 12. *If $I_\alpha(k) = I_\beta(k)$ for some $k \geq 1$ then either $\alpha = \beta$ or $\alpha = \frac{1}{\beta}$.*

Theorem 13. *If $\alpha \neq \beta$ and $\alpha \neq \frac{1}{\beta}$, then $\mathcal{S}_\alpha$ is not isomorphic to $\mathcal{S}_\beta$.*

Proof. Suppose that $\mathcal{S}_\alpha$ is isomorphic to $\mathcal{S}_\beta$; we want to prove that either $\alpha = \beta$ or $\alpha = \frac{1}{\beta}$. By Lemma 12 we will prove that there exists an integer $k \geq 1$ such that $I_\alpha(k) = I_\beta(k)$. By [7, Proposition 9] it is easy to prove that s is bispecial in $\mathcal{S}_\alpha$ if and only if s is a palindrome prefix of $\mathbf{x}_\alpha$. By [7, Proposition 7] s is a palindrome prefix of $\mathbf{x}_\alpha$ if and only if $s \in A(\mathbf{x}_\alpha)(ba)^{-1} \cup B(\mathbf{x}_\alpha)(ab)^{-1}$. By Propositions 2 and 3, if $w \in MF(L(\mathcal{S}_\alpha))$ then there exists $u \in A(\mathbf{x}_\alpha) \cup B(\mathbf{x}_\alpha)$ such that $|w| = |u|$ and conversely if $u \in A(\mathbf{x}_\alpha) \cup B(\mathbf{x}_\alpha)$, $|u| \geq 2$, then there exists $w \in MF(L(\mathcal{S}_\alpha))$ such that $|w| = |u|$ (more precisely the bispecial prefix s of u of length $|u| - 2$ is the "central" factor of $|w|$) .

Hence, for $k \geq 2$, $I_\alpha(k)$ is the set of lengths $p \geq k$ of minimal forbidden words of $L(\mathcal{S}_\alpha)$ and then, for $n \geq k$, $F_{\mathcal{S}_\alpha}(n)$ is a zero-one function that coincides with the characteristic function of the set $I_\alpha(k)$. Therefore, in order to prove that $I_\alpha(k) = I_\beta(k)$, it sufficies to prove that for $n \geq k$, $F_{\mathcal{S}_\alpha}(n) = F_{\mathcal{S}_\beta}(n)$.

By definition of (A_n, B_n), the sequence of the distances between two consecutive integers in $I_\alpha(k)$ (resp. $I_\beta(k)$) (i.e. the distances between two consecutive integers n, m such that $F_{\mathcal{S}_\alpha}(n) = F_{\mathcal{S}_\alpha}(m) = 1$) is an increasing divergent sequence. This fact, joined with the hypothesis that, by Theorem 4, $F_{\mathcal{S}_\alpha}$ and $F_{\mathcal{S}_\beta}$ are linearly equivalent, implies that there exist $C \geq 0$, $k \geq 2$ such that for any $n \geq k$, if $F_{\mathcal{S}_\alpha}(n) = 1$ (resp. $F_{\mathcal{S}_\beta}(n) = 1$) then there exists m such that $|n - m| \leq C$ and $F_{\mathcal{S}_\beta}(m) = 1$ (resp. $F_{\mathcal{S}_\alpha}(m) = 1$).

Using the developments in continued fractions of α and β, one can derive by arithmetic computations that the constant C is equal to zero. This concludes the proof.

Remark that if $\alpha = \frac{1}{\beta}$ then $\mathbf{x}_\beta$ is obtained from $\mathbf{x}_\alpha$ by exchanging letter a with letter b. Hence $\mathcal{S}_\alpha$ and $\mathcal{S}_\beta$ are trivially isomorphic.

Finally let us briefly discuss some consequences of our main result in the framework of formal languages. The fact that the growth of the function $F_L(n)$ is invariant under some natural transformation suggests that this function is a "good" tool to investigate the structure of the language L. The equivalence relation $\simeq$ introduced in this paper between functions from N to N induces an equivalence relation between languages as follows: $L_1 \simeq L_2$ if and only if $F_{L_1} \simeq F_{L_2}$. This gives a new classification for (factorial) formal languages. For instance locally testable factorial languages correspond to a class in this new hierarchy. In the same way, languages having a number of minimal forbidden words of length n which is a polynomial of same degree (exponential having same base, resp.) are in the same level of the hierarchy.

Remark that our theorem presents some analogies with a result of Maurer and Nivat in [13] concerning rational bijections of rational languages. We observe that the notion of "isomorphism", here considered for (factorial, extensible) languages associated to topologically isomorphic systems, is different from the notion of rational bijection in [13]. On the other hand the result of Maurer and Nivat gives a condition involving a function that counts, for any n, the number of words in the language of length n; whereas in our result we count the number of *forbidden* words. Moreover the equivalence relation between functions introduced in this paper is finer than that considered in [13].

References

1. R. Adler, D. Coppersmith, M. Hassner. *Algorithms for Sliding-Block Codes*, IEEE Trans. Inf. Theory, **IT-29** (1983), 5-22.
2. R. Badii. *Quantitave Characterization of Complexity and Predictability*, Physics Letters A **160** (1991), 372-377.
3. R. Badii. *Complexity and Unpredictable Scaling of Hierarchical Structures*, in "Chaotic Dynamics: Theory and Practice", T. Bountis Ed., Plenum Press, New York, 1992.
4. M. P. Béal. *Codage Symbolique*, Masson, 1993.
5. J. Berstel. *Fibonacci Words - a Survey*, in "The Book of L", G. Rozenberg, A. Salomaa eds., Springer Verlag 1986.
6. J. Cassaigne. *Complexité et Facteurs Spéciaux*, Actes des Journées Montoises, 1994.
7. A. de Luca, F. Mignosi. *Some Combinatorial Properties of Sturmian Words*, Theor. Comp. Science, **136** (1994), 361-385.
8. A. de Luca, L. Mione. *On Bispecial Factors of the Thue-Morse Word*, Inf. Proc. Lett., **49** (1994), 179-183.
9. S. Eilenberg. *Automata, Languages, Machines*, Vol. A, Academic Press, 1974.
10. G. A. Hedlund. *Endomorphisms and Autorphisms of the Shift Dynamical System*, Math. System Theory, **3** (1969), 320-375.
11. K. H. Kim, F. W. Roush. *Williams Conjecture is False for Reducible Matrices*, J. Amer. Math. Soc. **5** (1992), 213-125.
12. M. Lothaire. *Combinatorics on Words*, Addison-Wesley, Reading, MA 1983.
13. H. A. Maurer, M. Nivat. *Rational bijections of rational sets*, Acta Informatica **13** (1980), 365-378.
14. R. McNaughton, S. Papert. *Counter-Free Automata*, M.I.T. Press, MA 1970.
15. M. Morse. *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84-110.
16. D. Perrin. *Symbolic Dynamics and Finite Automata*, invited lecture in Proc. MFCS'95, Lect. Notes in Comp. Sci., **969**.
17. G. Rauzy, *Mots infinis en arithmétique* in "Automata on Infinite words", M. Nivat and D. Perrin eds., Lecture Notes in Comp. Science, **192**, Springer, Berlin 1984.
18. R. Williams. *Classification of shifts of finite type*, Annals of Math., **98** (1973), 120-153. Errata, Annals of Math., **99** (1974), 380-381.

This article was processed using the LaTeX macro package with LLNCS style