Contemporary Mathematics
Volume 135, 1992

SYNCHRONIZING PREFIX CODES AND AUTOMATA AND THE ROAD COLORING PROBLEM

Dominique Perrin, Marcel-Paul Schützenberger
LITP
Institut Blaise Pascal
Paris

Abstract

We prove two new results concerning the existence of synchronizing words for prefix codes. Both results assert that any finite aperiodic maximal prefix code is equivalent to a synchronizing one under two equivalence relations to be defined more precisely below. One of these equivalence relations is that of tree isomorphism and is the subject of a conjecture, known as the road coloring conjecture, that is settled in the case corresponding to our hypotheses.

1. INTRODUCTION

The notion of a *synchronizing word* is a basic and elementary notion in automata theory. Given a finite deterministic automaton, a word x is called synchronizing if the state reached after processing the word x is independent of the initial state in which the automaton was started. This notion has been studied since the beginning of automata theory and appeared with E.F. Moore's "gedanken experiments". It also appears in many recent developments concerning automata (see e.g. Aho, 1988 or Eppstein, 1990). The term "synchronizing word" is however not universally in use and one may find instead *resolving block* (Adler, Marcus, 1979) or *reset sequence* (Eppstein, 1990).

From the abstract point of view, synchronizing words correspond, in the semigroup of transitions of the automaton, to elements of minimal possible rank. This algebraic formulation allows a generalization to nondeterministic automata (see Berstel, Perrin, 1985). From another viewpoint, the existence of synchronizing words guarantees an almost everywhere one-to-one correspondance between paths and their labels in an

1991 Mathematics Subject Classification. Primary 68Q70; Secondary 20M05.
This paper is in final form and no version will be submitted for publication elsewhere.

0271-4132/92 $1.00 + $.25 per page

appropriate measure space. It is this property which is of interest in the applications to coding since it guarantees stability against errors.

It is curious that such a simple notion gives rise to several unsolved problems. We mention two of them in this introduction : The Cerny-Pin conjecture and the road coloring conjecture.

The Cerny conjecture asserts that any n-state synchronizing automaton has a synchronizing word of length at most $(n-1)^2$. It is easy to prove the existence of a synchronizing word of length bounded by a cubic polynomial in n but no quadratic bound has yet been obtained. The simple example of the automaton of Figure 1.1 shows that the bound $(n-1)^2$ cannot be improved.

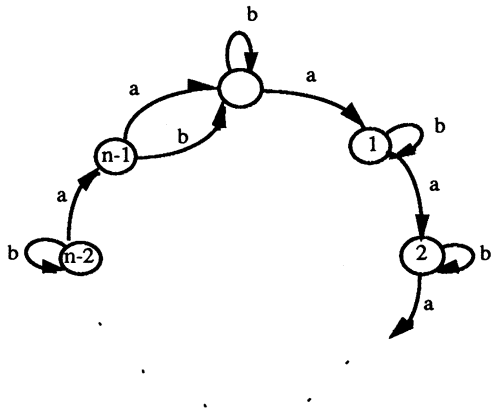

Figure 1.1. A worst case for Cerny's problem

The conjecture has been put in a more general form by Pin : if there is a word of rank d (as a mapping from the state set into itself) then there is one of length at most $(n-d)^2$. A bibliography on this problem can be found in (Berstel, Perrin, 1985). A recent result by A. Carpi (1988)shows that a cubic bound also holds in the case of unambiguous automata.

The road coloring problem is encountered in the study of isomorphism of symbolic dynamical systems (Adler, Goodwin, Weiss, 1977). It is conjectured that, except for the trivial case of periodicity, it is always possible to modify the labeling of the graph underlying a deterministic automaton to make it synchronizing. The name comes from the analogy

where the states of the automaton are cities and the edges roads connecting them. An appropriate coloring would allow a traveller to find his way to some city by following a fixed rule specifying the appropriate succession of colors, irrespective of his starting point. The conjecture is presently unsettled.

In this paper, we prove two new results on synchronizing words. Our results start with a prefix code instead of an automaton. Both notions are strongly related since, for any deterministic automaton, the set of first returns to a given state is a prefix code. However, our hypothesis are more easily formulated in terms of prefix codes.

We introduce an equivalence on prefix codes, called *the flipping equivalence*. It corresponds to isomorphism of the associated unlabeled trees.

Our first result is that any finite aperiodic maximal prefix code is flipping equivalent to a synchronizing one. The proof uses in a crucial way a theorem of Reutenauer (1985) giving a non-commutative factorization of the polynomial associated with a prefix code.

Our second result is a modification of the first one for another equivalence relation : the commutative equivalence which identifies words differing only in the relative ordering of their letters. The proof is quite similar to that of the previous result.

The first result settles, under our hypotheses, the road coloring problem. In terms of the original formulation, it settles it in the case of graphs satisfying the additional assumption that all vertices except one have exactly one entering edge. Such graphs are sometimes refered to as "renewal systems" in symbolic dynamics.

Our paper is organized as follows. In Section 2, we recall the definitions and results to be used later, especially the factorization theorem of Reutenauer. In Section 3, we discuss the case of an equivalence relation which is a common refinement of the two equivalence relations considered above. We reproduce a result of (Schützenberger, 1967) with part of its proof with the intention both of updating the statement and to prepare the study of the road coloring problem given in Section 6. In Section 4, we prove our main result concerning flipping equivalence. The corresponding result for commutative equivalence is proved in Section 5. Finally, in Section 6 we discuss the exact relationship of our results with the road

coloring problem.

2. PREFIX CODES

In all that follows we use the notation and terminology of (Berstel, Perrin, 1985). For the sake of readability we recall most of the definitions.

Let A be an alphabet. We denote by A^* the free monoid on the set A, which is the set of all finite sequences on A equipped with the concatenation as a product, the neutral element being the empty sequence, called the empty word. We denote by 1 the empty word and by $A^+ = A^* - 1$ the free semigroup on A. In general, we recall that a monoid is a set with a binary associative operation and a neutral element whereas a semigroup is the same but without the necessity of a neutral element.

A *prefix code* on A is a subset X of A^+ which contains no proper prefix of any of its elements. A prefix code can be identified with a labeled tree. Thus the prefix code $X = \{aa, ab, baa, bab\}$ on $A = \{a, b\}$ corresponds to the binary tree represented on Figure 2.1 with an obvious convention for the labeling using

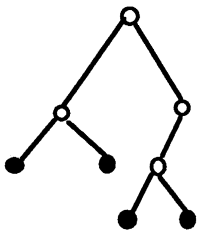

Figure 2.1. A prefix code

a for left and b for right. The words of X are in 1-1 correspondance with the leaves of the associated tree.

A *prefix* of X is a proper prefix of some word of X. The set of prefixes thus corresponds bijectively to the internal nodes of the associated tree.

For a subset X of A^*, we denote by X^* the submonoid generated by X. When X is a prefix code, X^* is free with basis X. This is the origin of the term "code" which refers in general to the uniqueness of parsing or deciphering.

A prefix code is said to be *maximal* when it is maximal under inclusion among the prefix codes on the alphabet A. It is easy to verify that a prefix

code X is maximal iff it is *right complete*, that is to say that for each word w in A^* one has

$$wA^* \cap XA^* \neq \emptyset \tag{2.1}$$

Equation (2.1) means that every word is comparable to a codeword for the prefix ordering. It is not difficult to prove that it is equivalent to the fact that each word w in A^* is a prefix of some word in X^*, i.e.

$$wA^* \cap X^* \neq \emptyset \tag{2.2}$$

In terms of trees, a prefix code is maximal iff the associated tree is a complete k-ary tree, where $k = Card(A)$.

We shall mainly discuss here *finite* prefix codes. We shall however occasionnally consider a much weaker condition defined as follows. A prefix code X on the alphabet A is called *thin* if there exists a word w in A^* that does not appear inside words of X, i.e. such that

$$A^*wA^* \cap X = \emptyset \tag{2.3}$$

A finite prefix code X is thin since only words of bounded length may appear inside words of X.

We now come to the definition of the objects of central interest to us. A word x is said to be *synchronizing* for a prefix code X if wx is in X^* for all words w in A^*. Hence x is synchronizing iff

$$A^*x \subset X^* \tag{2.4}$$

A prefix code X is called synchronizing if there exists a synchronizing word for X. A synchronizing prefix code is obviously maximal since Formula (2.4) is a uniformisation of Formula (2.2) It is also thin since no element of X contains x properly.

For instance, the prefix code $X = \{aa, ab, baa, bab, bb\}$ represented on Figure 2.2. (i) admits $x = baa$ as a synchronizing word as the reader may check by a little reasonning. On the contrary, the code $X = \{aa, ab, ba, bb\}$ of Figure 2.2 (ii) is not synchronizing and the same is true of any code in which all words have the same length not equal to one.

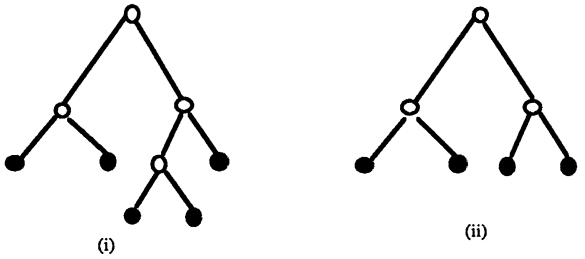

Figure 2.2. (i) A synchronizing prefix code and (ii) a non synchronizing one

We shall see below how to systematically look for synchronizing words.

We also need some terminology from automata theory. Let Q be a set. An *automaton* on Q is given by a function

$$\delta : Q \times A \to Q$$

This function defines a right action of A^* on Q. We denote this action by a dot, writing $q.a$ instead of $\delta(q, a)$.

Given an element $i \in Q$, the *stabilizer* of i is the set

$$Stab(i) = \{x \in A^* \mid i.x = i\}$$

It can be verified that $Stab(i)$ has the form $Stab(i) = X^*$ with X a prefix code, sometimes called the set of *first returns*. Conversely any prefix code can be obtained in this way. One may further assume that all elements q of Q play a role in the sense that there exist u, v in A^* such that $i.u = q$ and $q.v = i$. We say in this case that the automaton is *trim* or *irreducible*.

We define the *rank* $r(w)$ of a word w as the number of elements of Q reachable through w, i.e.

$$r(w) = Card\{q.w \mid q \in Q\}$$

A word $x \in X^*$ is clearly synchronizing iff $r(x) = 1$. In general, the *degree* of X denoted $d(X)$ is the minimal non-zero value of the ranks of the words of A^*. It can proved that it does not depend on the automaton used to obtain X (provided it is trim). Hence X is synchronizing iff $d(X) = 1$.

A finite prefix code can be obtained from a finite automaton. It is synchronizing iff its degree is equal to one. In this case, the automaton itself is also called synchronizing. For example, the prefix code of Figure 2.2 (i) corresponds the first return at node 1 in the automaton given on Figure 2.3.

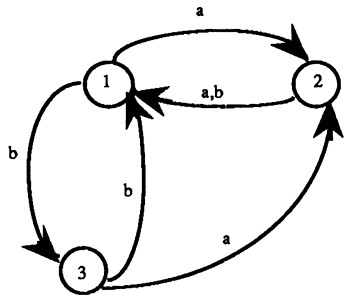

Figure 2.3. A finite automaton

The search for a synchronizing word is easily done with a finite automaton. It reduces to a search in the graph obtained by considering the action of the letters on the subsets of the state set. A synchronizing word is one that is the label of a path from the set of all states to a singleton set. The graph corresponding to the automaton of Figure 2.3 is represented on Figure 2.4 with only part of the edges represented. It allows one to find easily the synchronizing word $x = baa$.

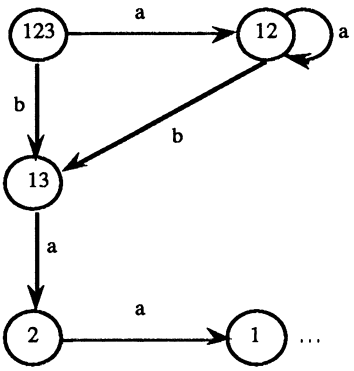

Figure 2.4. The action on subsets(partial drawing)

It is of course not true that, conversely a prefix code obtained from a finite automaton is itself finite, since there may be cycles in the graph of

the automaton that do not use the special state i. One may show however that the code is thin when the automaton is finite (see Berstel, Perrin, 1984).

The *period* of a prefix code X is the gcd of the lengths of its elements. It is known that, *for a maximal prefix code of finite degree, the period is a divisor of the degree* (see Berstel, Perrin, 1984 p. 242). This implies that a prefix code can be synchronizing only when it is of period 1. The study of synchronizing automata or codes deals with the problem of finding additional conditions ensuring that the converse implication holds.

Several other properties relating the degree to other parameters are also known. A useful one is the following : For a finite maximal prefix code X, *the degree is a divisor of each of the integers n such that $a^n \in X$ for $a \in A$* (see Berstel, Perrin, 1984 p. 117).

We will use on several occasions non-commutative polynomials and series. We recall here the basic notions on this subject. A systematic exposition can be found in (Cohn, 1985) or (Berstel, Reutenauer, 1988).

We denote by $\mathbb{Z} << A >>$ the ring of series with coefficents in $\mathbb{Z}$ and non-commutative variables in A and by $\mathbb{Z} < A >$ the corresponding ring of polynomials. For a serie S, we denote by (S, w) the value of S on the word w, also called the coefficient of w in S. We shall write

$$S = \sum_{w \in A^*} (S, w)w$$

The *support* of a series S, denoted $supp(S)$ is the set of words w such that $(S, w) \neq 0$. A serie is a polynomial iff its support is finite.

We shall not distinguish between a subset X of A^* and its characteristic series, writing therefore

$$X = \sum_{x \in X} x$$

We denote by $|P|$ the *degree* of a polynomial P, which is the maximum of the lengths $|w|$ of the elements w in its support. We also denote by $\widehat{P}$ the homogeneous component of P of maximum degree. Therefore

$$(\widehat{P}, w) = \begin{cases} (P, w) & \text{if } |w| = |P| \\ 0 & \text{otherwise} \end{cases}$$

For u in A^* and S in $\mathbb{Z} << A >>$ we denote $u^{-1}S$ the series defined by

$$(u^{-1}S, w) = (S, uw)$$

with the symmetric definition for Su^{-1}.

We shall use several times the fact that the set of homogeneous polynomials is a *free* subsemigroup of $\mathbb{Z} < A >$.

We now show the interplay between codes and polynomials.

Let X be a maximal prefix code on A and let P be the set of its proper prefixes including the empty prefix. We have the equality

$$X - 1 = P(A - 1) \tag{2.5}$$

in which 1 denotes the empty word.

Formula (2.5) expresses a factorisation property. It is easy to derive from the equality between sets

$$PA + 1 = X + P$$

expressing the fact that a prefix followed by a letter is either still a prefix or is a word of X.

A much deeper factorisation property was given by Reutenauer(1985). We state it below in its simplified version concerning prefix codes although his result is more general and holds for general codes.

THEOREM 2.1 (REUTENAUER). — *Let X be a finite maximal prefix code on the alphabet A. There exists two polynomials $L, D \in \mathbb{Z} < A >$ such that*

$$X - 1 = L(d + (A - 1)D)(A - 1) \tag{2.6}$$

where d denotes the degree of X.

A proof of the result is presented in the book of (Berstel, Reutenauer, 1988). It is important to see that when X is not synchronizing, i.e. when $d > 1$, the central factor in the right handside of (2.6) is non trivial. In fact, assuming that the constant term of L is 1, the constant term of D must be $d - 1$, which implies $D \neq 0$.

Also comparing (2.5) and (2.6), we obtain the equality

$$P = L(d + (A - 1)D) \tag{2.7}$$

which expresses a factorisation of the polynomial of prefixes of X.

Equality (2.6) can be rewritten

$$X - 1 = L(A - 1)(d + D(A - 1)) \tag{2.8}$$

By inverting both sides and using the identity $X^* = (1 - X)^{-1}$ we obtain

$$A^* = (d + D(A - 1))X^*L \tag{2.9}$$

We conjecture that for any finite maximal prefix code X of degree d there exist a finite collection of d disjoint maximal prefix codes $T_i (1 \leq i \leq d)$ and a set L such that

$$A^* = (\sum_{i=1}^{d} T_i)X^*L \tag{2.10}$$

Such an equality implies the existence of a factorization like (2.9) since, letting $T_i - 1 = U_i(A - 1)$ we have

$$A^* = (d + (\Sigma U_i)(A - 1))X^*L \tag{2.11}$$

It implies the stronger property that the polynomials L, D in (2.6) can be chosen to have positive coefficients. It also implies that the degree of X is at least equal to d according to the following observation.

PROPOSITION 2.3.. — *Let X be a finite maximal prefix code on the alphabet A such that*

$$X - 1 = L(A - 1)R \tag{2.12}$$

with L, R two subsets of A^*. If R the disjoint union of d maximal prefix codes, then X is of degree at least equal to d.

Proof : We first show that each element of R is a suffix of an element of X. Let indeed r be in R an let $l \in L$ be chosen of length $|L|$. Then, for any letter a in A, lar has coefficient at least one in $X + LR$. Since l is of maximal length, this implies that either r is a suffix of X or it is a suffix of some other element of R. This proves the property by ascending induction on $|r|$.

We now consider an automaton on Q such that X is the set of first returns to a state i. We will show that any word in A^* has at least d states in its range. Let $w \in A^*$ be longer than $|X|$. Then w has d prefixes $t_1, \ldots, t_d$ in R. Since each t_k is a suffix of an element of X, there is a state q_k such that $q_k.t_k = i$.

For each $k = 1, \ldots, d$, let r_k be the state defined by

$$r_k = q_k.w$$

We will verify that all r_k are distinct and this will prove the claim. Let indeed k, ℓ be such that $r_k = r_\ell$. Since we may concatenate w on the right by any word we may suppose that $r_k = r_\ell = i$. Let $w = t_k x_k = t_\ell x_\ell$. Then x_k, x_ℓ are in X^* since they stabilize i. But then the word w has two distinct factorizations in the product RX^*L namely (t_k, x_k, ε) and $(t_\ell, x_\ell, \varepsilon)$, (see Figure 2.5). This is a contradiction since (2.12) is equivalent to the equation

$$A^* = RX^*L$$

□

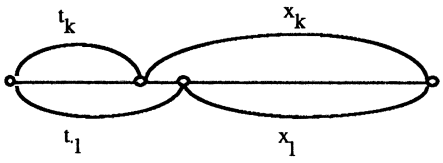

Figure 2.5. Two parsings of w

It is known that a factorization like Eq. (2.10) holds for biprefix codes with $L = 1$ (see Berstel, Perrin, 1985). In the general case, the answer is not known. It is a particular case of a more general conjecture on codes known as the factorization conjecture (ibid. p. 423).

3. LENGTH DISTRIBUTIONS

For a subset X of A^*, the sequence of numbers $\alpha = (\alpha_n)_{n \geq 0}$ given by

$$\alpha_n = Card(X \cap A^n)$$

is called the *length distribution* of X. We also denote

$$f_X(t) = \sum_{n \geq 0} \alpha_n t^n$$

the corresponding generating series. We denote by ρ_X or ρ_α the radius of convergence of the series $f_X(t)$. Let $q = Card(A)$. Since $\alpha_n \leq q^n$, we have $\rho_X \geq 1/q$.

When X is thin, we have $\rho_X > 1/q$ (see Eilenberg, 1974 p. 230 or Berstel, Perrin 1985 p. 67).

A sequence $(\alpha_n)_{n\geq 0}$ is the length distribution of a prefix code on a q-letter alphabet iff it satisfies the inequality

$$\sum_{n\geq 1} \alpha_n q^{-n} \leq 1 \tag{3.1}$$

Inequality (3.1) goes back to C. Shannon and it is at times referred to as *Kraft Inequality*. When X is a thin maximal prefix code, we have

$$\sum_{n\geq 1} \alpha_n q^{-n} = 1 \tag{3.2}$$

Indeed, by Equality (2.5) we have

$$f_X - 1 = f_P(qt - 1) \tag{3.3}$$

Since X is thin, P is thin and therefore $\rho_P > 1/q$. Evaluating both sides of (3.3) at $t = 1/q$ gives the desired equality. Conversely, we have the following

THEOREM 3.1 (SCHÜTZENBERGER, 1967). — *If α is a sequence satisfying Equality (3.2) and $\rho_\alpha > 1/q$, it is the enumerating sequence of some thin maximal prefix code. Moreover, the code can also be chosen synchronizing provided the sequence $(\alpha_n)_{n\geq 0}$ satisfies the additional requirement that the integers α_n are relatively prime.*

We shall reproduce here the part of the proof of this result needed for the purpose of a discussion presented in Section 6.

It is not difficult to see that the first part is true. Indeed, if $\rho_\alpha > 1/q$, we may build a prefix code X with length distribution α such that some word w does not appear within any word of X. Then X is thin and maximal. We may always choose $w = a^k$ for some letter a in A. Then X satisfies the following inclusion

$$A^*w \subset X^*a^*$$

and to choose X synchronizing, we only need a word $x \in X^*$ such that

$$a^*x \subset X^*$$

One may show that except for a trivial case where the sequence α_n is ultimately equal to one, we may rearrange the words of X in a length-preserving way so that for some $b \in A$ and some integer $n \geq 0$ the prefix code $Y = X \cap (a^* \cup a^*ba^*)$ satisfies

$$Y = a^n \cup \{y_0, y_1, \cdots, y_{n-1}\} \tag{3.6}$$

where each $y_i = a^i b a^{\lambda_i - i - 1}$ is a word of length λ_i satisfying

$$i + 1 \leq \lambda_i \leq n \tag{3.7}$$

and there is an integer t with $1 \leq t \leq n$ such that $\lambda_i = n$ iff $i \geq n - t$ and finally the number λ_i are relatively prime.

The above conditions are satisfied in particular when $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$ and the λ_i are relatively prime.

The following lemma therefore completes the proof of Theorem 3.1.

LEMMA 3.2. — *If Y satisfies the above conditions, there exists a word y in Y^* such that $a^*y \subset Y^*$*

Proof : We denote $Q = \{0, 1, \cdots, n-1\}$ and we define an action on Q by

$$i.a = (i + 1) \mod n$$
$$i.b = (i - \lambda_i + 1) \mod n$$

The corresponding automaton is such that $Y^* = Stab(0)$. Let M be the transition monoid of the automaton, which is the monoid of all mappings from Q into Q obtained by the action of all words. For each d with $1 \leq d \leq n$, let

$$I_d = \{n - d, \cdots, n - 2, n - 1\}$$

and let M_d be the monoid

$M_d = \{m \in M \mid Q.m = I_d \text{ and } i.m = i \text{ for all } i \in I_d\}$.

We want to prove that M_1 is not empty. This implies our conclusion since a word z defining an element of M_1 satisfies

$$a^*(za) \subset Y^*$$

In the sequel we do not distinguish between a word and the element of M that it defines. We first observe that for all $i \in Q$ we have

$$i.ba^{n-1} \geq i$$

with equality iff $i \in I_t$. Thus ba^{n-1} has a power which belongs to M_t. This proves that M_t is not empty. We shall now prove that if M_s with $1 < s \leq t$ is not empty, then some M_r with $1 \leq r < s$ is not empty. For q in Q, we denote $[q]_s$ the integer in $\{1, 2, \cdots, s\}$ congruent to q mod. s. Let m be an element of M_s.

Case 1. There exists a p with $1 \leq p \leq n$ such that $(n-p).m \neq n - [p]_s$. We choose the smallest p satisfying this inequality. Then $p > s$ by the definition of M_s. Let

$$m' = ma^{n+s-p}m$$
$$J_s = I_s - (n - [p]_s)$$

Since $Q.m = I_s.m$, we have $Q.m' = I_s.m'$. For all s' with $1 \leq s' < s$ we have

$$\begin{aligned}(n-s').m' &= (n-s').a^{n+s-p}m \\ &= (n + (s-s') - p).m \\ &= n - [p + s']_s\end{aligned}$$

and

$$\begin{aligned}(n-s).m' &= (n-s).a^{n+s-p}m \\ &= (n-p).m \\ &\neq n - [p]_s\end{aligned}$$

Hence $Q.m' = J_s$. If $[p]_s = 1$, the element $m_1 = m'a$ belongs to M_{s-1}. Otherwise, let $[p]_s = k > 1$, and let

$$m_2 = m'(a^{n-1}m)^{s-k}$$

We have, for $1 \leq s' < s$,

$$(n-s').m' = n - [k + s']_s$$

and for $1 \leq s' \leq s$

$$(n-s').(a^{n-1}m) = (n-s'-1).m = n - s' - 1$$

whence

$$(n-s').(a^{n-1}m)^{s-k} = n - s' - k$$

and finally

$$(n - s').m_2 = n - s' .$$

Hence m_2 belongs to M_{s-1} whence the desired conclusion in this case also.

Case 2. For all p with $1 \leq p \leq n$ we have $(n-p).m = n - [p]_s$. We first suppose that s does not divide n. Let then $n = n's + d$ with $1 \leq d \leq s$. For all d' with $1 \leq d' \leq d$ we have

$$\begin{aligned}(n - d').a^d m &= (d - d').m \\ &= n - d'\end{aligned}$$

Hence $a^d m$ fixes pointwise the set I_d. Also for $d < d' \leq s$ we have

$$\begin{aligned}(n - d').a^d m &= (n + d - d').m \\ &= n + d - d'\end{aligned}$$

Hence some power of $a^d m$ belongs to M_d.

We are finally left with the case where s divides n. It is easy to see that this implies that $p.m \equiv p$ mod.s for all $p \in I$ and $m \in M$. Since the λ_i are relatively prime, this implies $s = 1$, a contradiction. □

An additional problem concerning length distributions is the following. When a prefix code X is the stabilizer of a state in a finite automaton, then the series $f_X(t)$ is rational (see Eilenberg, 1974). It is not completely known under which conditions the converse holds, i.e. under which additional assumptions Theorem 3.1 holds with the additional conclusion that the prefix code is a stabilizer in a finite automaton. See (Perrin, 1989) for a partial answer.

4. FLIPPING EQUIVALENCE

We introduce a transformation on prefix codes called *flipping*. It is defined as follows. Let X be a prefix code on the alphabet A. Let $a, b \in A$ be two letters and let u be a proper prefix of X. Let

$$X = X' + uaR + ubS$$

with X', R, S prefix codes. One has in fact $R = (ua)^{-1}X, S = (ub)^{-1}X$.

The prefix code

$$Y = X' + uaS + ubR$$

is said to be the image of X under an *elementary flip*. A flip is a composition of elementary flips. The flipping transformation defines an equivalence called the *flipping equivalence*. We denote

$$X \sim Y$$

two prefix codes X, Y which are flipping equivalent.

The flipping transformation is of course a very natural and simple one on the trees associated with prefix codes. Indeed, an elementary flip is just an exchange of two subtrees rooted at the sons of some vertex.

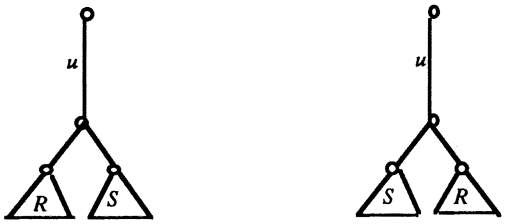

Figure 4.1. An elementary flip

We have represented on Figure 4.2 an equivalence class of the flipping equivalence. Actually two unlabeled complete binary trees correspond to flipping equivalent maximal prefix codes iff they are isomorphic, as one may easily verify.

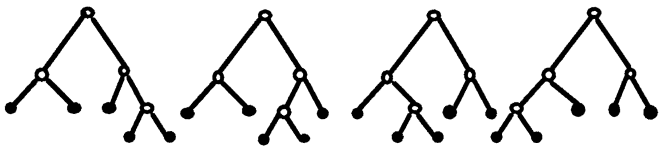

Figure 4.2. A flipping equivalence class

The flipping equivalence preserves some of the properties of prefix codes. First of all, two flipping equivalent pregix codes have the same length distribution (the converse implication is however not true). As a consequence, equivalent prefix codes have the same period. Also, two equivalent prefix codes are simultaneously maximal or not maximal.

We are going to prove the following result. It is, in the case of finite prefix codes, a reinforcement of Theorem 3.1

THEOREM 4.1. — *The flipping equivalence class of any finite maximal prefix code of period 1 contains at least one synchronizing prefix code.*

The proof relies on two lemmas. In the first lemma we start with a Reutenauer's factorization (2.8)

$$X - 1 = L(A-1)(d + (A-1)D)$$

and consider $R = (d+(A-1)D)$. The homogeneous component of highest degree is, when $D \neq 0$

$$\widehat{R} = A\widehat{D}$$

The lemma shows that, except for the periodic case, the homogeneous polynomial $\widehat{R}$ is not equal to A^n. The condition given is the lemma is of course also sufficient.

LEMMA 4.2. — *If X is a finite maximal prefix code of period p such that*

$$X - 1 = L(A-1)R$$

where $\widehat{R} = A^n$ for some $n \geq 1$, then R is a polynomial in A dividing $1 + A + \ldots + A^{p-1}$.

Proof.Let $E = (A-1)R$. We first show that E is a polynomial in A. Let us suppose by induction on $m < n$ that

$$E = E' + \sum_{i=m+1}^{n} s_i A^i \tag{4.1}$$

with $| E' | \leq m$. Let g be in the support of $\widehat{L}$ and let h be a word of length m. For all words k of length $n - m$ we have $ghk \in supp(\widehat{L}\widehat{E}) \subset supp(X)$ and thus $ghk \in X$. Since X is prefix, we have $(LE, gh) = 0$.

But, by Formula (4.1) we have

$$(LE, gh) = (L, g)(E', h) + \sum_{i=m+1}^{n} r_{t+m-i} s_i \tag{4.2}$$

where r_i is the coefficient in L of the prefix of length i of g and $t = | g |$. Since $(LE,gh) = 0$, we deduce from (4.2) the Formula

$$(E', h) = -(1/(L, g)) \sum_{i=m+1}^{n} r_{t+m-i} s_i$$

It shows that (E', h) does not depend on the word h but only on its length m and proves that Equality (4.1) is true for $m-1$. Thus we have proved by induction that E is a polynomial in A, i.e.

$$E = \sum_{i=0}^{n} s_i A^i$$

Let x be a word of X of length q. Let r, s be the polynomials in the variable z

$$r(z) = \sum_{i=0}^{q} r_i z^i \quad s(z) = \sum_{i=0}^{n} s_i z^i$$

where r_i is the coefficient in L of the prefix of length i of x. We have for each integer m such that $0 < m < q$ the equality similar to (4.2)

$$\sum_{i+j=m} r_i s_j = 0$$

since, X being prefix, the coefficient of the prefix of length m of x in LE is zero. We therefore have

$$z^q - 1 = r(z)s(z)$$

and the lemma is proved. □

We now prove a second lemma. It shows that, in the non periodic case, we may use the flipping transformation to destroy the possibility of a non trivial factorization of the polynomial $X - 1$. For a finite maximal prefix code X, we denote by $e(X)$ the integer defined by

$$e(X) = \max\{e \geq 0 \mid X - 1 = L(A-1)R, e = \mid R \mid\}$$

Thus, $e(X) > 0$ iff X has a non-trivial factorization. Consequently, $e(X) = 0$ implies that X is synchronizing.

LEMMA 4.3. — *Let X be a finite maximal prefix code such that*

$$X - 1 = L(A-1)R \tag{4.3}$$

with $\mid R \mid = n \geq 1$ and $\widehat{R} \neq A^n$ Then there exists a prefix code X' flipping equivalent to X such that

$$e(X') < e(X)$$

Proof.Let $E = (A - 1)R$. We first note that Eq. (4.3) implies that $\widehat{X} = \widehat{L}A\widehat{R} = \widehat{L}\widehat{E}$. Therefore the homogeneous polynomials $\widehat{L}, \widehat{E}$ are unambiguous, i.e. have 0-1 coefficients. Let $g \in \widehat{L}$ and let Y be the finite maximal prefix code $Y = g^{-1}X$. We have $\widehat{Y} = \widehat{E}$. Since $\widehat{Y} \neq A^{n+1}$, there exists a prefix code Y' flipping equivalent to Y such that $\widehat{Y} \neq \widehat{Y}'$. Let X' be the maximal prefix code defined by the equality

$$X' - gY' = X - gY \tag{4.4}$$

We have $X' \sim Y'$. Consider a non trivial factorization

$$X' - 1 = L'E' \tag{4.5}$$

and suppose by contradiction that $| E | \leq | E' | < | X |$. Since $g\widehat{Y}' \subset \widehat{X}' = \widehat{L}'\widehat{E}'$, the set $\widehat{L}'$ contains a prefix g' of g. Let $g = g'h$ (see Figure 4.3).

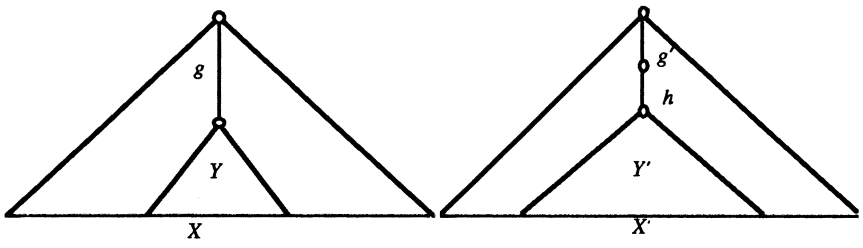

Figure 4.3. The codes X and X'

Let $F = h^{-1}E', H = g'^{-1}L - h$ and let L_1, E_1', L_1' be defined by the following equalities

$$L = L_1 + g'H + g'h$$
$$E' = E_1' + hF$$
$$L' = L_1' + g'$$

Substituting into (4.3) and (4.5), we obtain

$$\widehat{X} = (\widehat{L}_1 + g'\widehat{H} + g'h)\widehat{E} \tag{4.6}$$

$$\widehat{X}' = \widehat{L}'(\widehat{E}_1' + h\widehat{F}) \tag{4.7}$$

By restricting Equality (4.4) on both sides to the words of maximal length beginning with g', we derive

$$\widehat{E}'_1 = \widehat{H}\widehat{E} \tag{4.8}$$

And by restricting (4.4) to the words of maximal length that do not begin by g' we obtain

$$\widehat{L}'_1(\widehat{E}'_1 + h\widehat{F}) = \widehat{L}_1\widehat{E} \tag{4.9}$$

Substituting in (4.9) the value of $\widehat{E}'_1$ given by (4.8) we have

$$\widehat{L}'_1 h\widehat{F} = (\widehat{L}_1 - \widehat{L}'_1\widehat{H})\widehat{E} \tag{4.10}$$

Since $\mid \widehat{F} \mid = \mid \widehat{E} \mid$, we deduce from (4.10) that $\widehat{F} = \widehat{E}$.

This contredicts the hypothesis $\widehat{Y} \neq \widehat{Y}'$ since on one hand $\widehat{Y} = \widehat{E}$ and on the other hand $\widehat{F} = \widehat{Y}'$.□.

We can now complete the proof of Theorem 4.1. We use an induction on the integer $e(X)$. The property is true when $e(X) = 0$ since then X itself is synchronizing. When $e(X) \geq 1$, we have $X - 1 = L(A - 1)R$ with $\mid R \mid = n \geq 1$. If $\widehat{R} = A^n$, then by Lemma 4.2, R divides $1 + A + \ldots + A^{p-1}$ with p the period of X. Hence, $p \geq n + 1 \geq 2$ in contradiction with the hypothesis $p = 1$. Therefore, $\widehat{R} \neq A^n$ and by Lemma 4.3, there exists an X' flipping equivalent to X such that $e(X') < e(X)$ whence the property by induction.

5. COMMUTATIVE EQUIVALENCE

There is another equivalence on prefix codes which is also a refinement of the length distxribution equivalence. This equivalence, called *commutative equivalence* is of more general interest since it applies to all subsets of the free monoid. We first recall its definition.

Two words u, v in A^* are said to be commutatively equivalent if for all a in A the number of occurrences of a in u is equal to the number of occurrences of a in v. We denote this equivalence by the symbol $\equiv$. Two subsets X, Y of A^* are said to be commutatively equivalent if there is a one-to-one mapping f from X onto Y such that for all x in X, one has $f(x) \equiv x$. We again denote $X \equiv Y$. Figure 5.1 represents two commutatively

equivalent maximal prefix codes. Actually their equivalence class does not contain other prefix codes (compare with Figure 4.2)

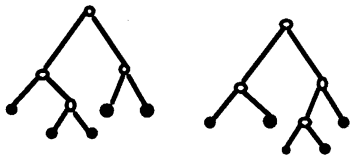

Figure 5.1. Two commutatively equivalent prefix codes

We will basically use the same arguments as in the preceding section to prove the following result.

THEOREM 5.1. — *The commutative equivalence class of any finite maximal prefix code of period 1 contains at least one synchronizing prefix code.*

The proof goes along the same lines as the proof of Theorem 4.1. We choose X such that the integer $e(X)$ is minimal in its commutative equivalence class. Suppose, by contradiction, that X is not synchronizing. Then we have

$$X - 1 = L(A - 1)(d + D(A - 1))$$

with $\mid D \mid \geq 2$. By Lemma 4.2 we have $\widehat{D} \neq A^n$ since otherwise X would be of period $p \geq 2$. Consequently, there exists a word h such that for some pair of letters a, b in A we have

$$(ha)^{-1}\widehat{D} \neq (hb)^{-1}\widehat{D}$$

Let $U = (ha)^{-1}\widehat{D}, V = (hb)^{-1}\widehat{D}$. Let $g \in \widehat{G}$ and $Y = g^{-1}X$. We have $\widehat{Y} = A\widehat{D}$ and therefore

$$Y = W + ahbU + bhaV$$

Let

$$Y' = W + ahbV + bhaU$$

Since $ahb \equiv bha$, we have $Y \equiv Y'$. Let X' be the prefix code commutatively equivalent to X defined by

$$X' - gY' = X - gY$$

Then, one may use the same proof as in Lemma 4.3 to show that $e(X') < e(X)$, a contradiction. This proves Theorem 5.1

To close this section, we mention the fact that the commutative equivalence is the object of an important open problem about codes. It is indeed conjectured that any finite maximal code is commutatively equivalent to a prefix code (see Berstel, Perrin, 1985).

6. THE ROAD COLORING PROBLEM

We finally discuss the road coloring problem mentioned in the introduction and we relate it to the results of the previous sections.

Let $\mathcal{A}$ be a finite automaton given by a function

$$\delta : Q \times A \to Q$$

The underlying graph of $\mathcal{A}$ is the directed graph having Q as set of vertices and an edge (p, q) iff there is an $a \in A$ such that $\delta(p, a) = q$. It is therefore the graph obtained from the familiar diagram associated with the automaton after removing the labels of the edges. It has the property that all its vertices have the same outdegree, in fact equal to the number of symbols in A.

A graph is said to be *road colorable* if it is the underlying graph of some synchronizing automaton.

Recall that a graph is called aperiodic if there is an integer n such that the n-th power of its adjacency matrix has all its elements strictly positive. This is of course equivalent to the graph being strongly connected and the g.c.d of the cycle lengths being equal to one.

The conjecture formulated in (Adler, Goodwin, Weiss, 1977) is the following : *any aperiodic graph with all vertices of the same outdegree is road colorable.*

We reformulate Theorem 4.1 as follows to obtain a solution of this conjecture in a particular case.

THEOREM 6.1. — *Any aperiodic graph such that*
(*i*) *all vertices have the same outdegree*
(*ii*) *all vertices except one have indegree one*
is road colorable.

Proof. We define the *renewal automaton* of a finite maximal prefix code X to be the automaton having the set P of prefixes of X as set of states and

the transition function defined by $\delta(p,a) = pa$ if $pa \in P$ and $\delta(p,a) = 1$ otherwise. The graph underlying the renewal automaton of X is therefore obtained from the unlabeled tree associated with X by merging all leaves with the root. Clearly, an elementary flip does not affect this graph.Hence, when X and X' are flipping equivalent, the underlying graphs of their renewal automata are the same and the result follows from Theorem 4.1.□

The road coloring conjecture is known to be true in some other particular cases. One of them (O'Brien, 1981) is that of graphs satisfying the additional assumptions

(i) there are no multiple edges

(ii) there is a simple cycle of prime length.

Another case, proved by Friedman (1990) is that of graphs containing a simple cycle of length prime to the weight of the graph. The weight of a graph is defined to be the sum of the components of an integer Perron left eigenvector chosen with its components relatively prime.

Some further particular cases have been investigated by A. Mahieux (1986).

In the paper of (Adler et al., 1977) the following result is proved : let G be an aperiodic graph with constant outdegree. Let M be the adjacency matrix of G and let n be an integer such that M^n has all its coefficients positive. For $k > 0$, let $G^{(k)}$ denote the graph having as vertices the paths of length h in G and edges the pairs (s,t) with $s = (s_1,\ldots,s_k), t = (s_2,\ldots,s_k,s_{k+1})$. Then $G^{(2n)}$ is road colorable. In terms of symbolic dynamics, this means that the system of finite type associated with G is conjugate to one that is road colorable. This result can actually also be proved using the construction of (Schützenberger, 1967) reproduced in Theorem 3.1. Indeed a splitting of the states of the graph will allow to label the cycles in such a way as to obtain a set of first returns containing the words described by Equations (3.3.6-7)

REFERENCES

[1] Adler R.L., Goodwin L.W., Weiss, B.,1977, Equivalence of topological Markov shifts,*Israel J. Math.* t.**27**, p. 49-63.

[2] Adler R.L., Marcus, B., 1979, Topological entropy and equivalence of dynamical systems, *Memoirs AMS*, **219**.

[3] Aho A., Dahbura, A., Lee, D., Uyar, M., 1988, An optimization technique for protocol conformance test generation based on UIO sequences and rural chinese postman tours, in *Protocol Specification, Testing and Verification VIII*, S. Aggarwal and K. Sabnani eds., North Holland.

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

[5] Berstel J., Reuteneauer, C., 1988, *Rational Series and their Languages*, Springer.

[6] Carpi, A., 1988, On synchronizing unambiguous automata, *Theoret. Comput. Sci.*,**60**, p.285-296.

[7] Cohn, P.M., 1985, *Free Rings and their Relations*, Academic Press (2nd edition).

[8] Eilenberg, S., 1974, *Automata, Languages and Machines*, Vol. A, Academic Press.

[9] Eppstein, D., 1990, Reset sequences for monotonic automata, *SIAM J. Comput.*, **19**, p. 500-510.

[10] Friedman, J., 1990, On the road coloring problem, *Proc. Amer. Math. Soc.* **110**, 1133-35.

[11] Mahieux, A., 1986, unpublished manuscript.

[12] O'Brien, G.L., 1981, The road coloring problem, *Israel J. Math*, t. **39**, p. 145-154.

[13] Perrin, D., 1989, Arbres et séries rationnelles, *C.R. Acad. Sci. Paris*, **309**, 713-716.

[14] Reutenauer, C.,1985, Noncommutative factorisation of variable-length codes *J. Pure applied Algebra* t.*36*, p.167-186.

[15] Schützenberger, M.P.,1967, On synchronizing prefix codes, *Information and Control*, t. **11**, p.396-401.