On syntactic groups

Dominique Perrin Giuseppina Rindone

Abstract

We prove that for any finite prefix code X with n elements, the non special subgroups in the syntactic monoid of X^* have degree at most $n-1$. This implies in particular that the groups in the syntactic monoid of X^* are all cyclic when X is a prefix code with three elements.

Résumé

Cet article présente un nouveau résultat sur les groupes G apparaissant dans les monoïdes syntaxiques d'ensembles de la forme X^* où X est un code préfixe fini. Nous prouvons que si X a n éléments, alors G est de degré au plus $n-1$ en tant que groupe de permutations. Nous situons ce résultat au sein des autres résultats connus sur les propriétés syntaxiques des codes finis.

1 Introduction

A lot of research has been done on describing the possible relations among a finite set of words. The case of sets with 2 or 3 words is almost completely understood while with as little as 4 words, many things remain unknown.

The subject presents various aspects, following the point of view adopted. A possibility is to fix a constraint satisfied by a set of words. This leads to the study of *word equations*. The solution of a word equation in n variables over an alphabet A is a set of n words on A. This relates families of n-sets of words as solutions of a given fixed equation. For example, the equation

$$xy = yz$$

1991 *Mathematics Subject Classification :* 20M05, 68R15.
Key words and phrases : variable-length codes, syntactic monoids, finite groups.

has as a solution

$$x = uv, \quad y = u, \quad z = vu.$$

On the opposite, one may fix the set of words and study the relations that they satisfy. The aim can be either to try to simplify the approach in open problems by limiting the cardinality or, on the contrary to try to list the special properties induced by the given cardinality.

In the first case, one meets the type of problem studied in particular in [8] which is the commutator problem raised by Conway: is the set of words commutating with a given rational language rational? the answer is not even known for a set with four words.

The other point of view leads to the study of properties shared by sets X of n words for small values of n. For example, it is known that for $n \leq 3$, the submonoid X^* is finitely presented [13]. The types of presentations have been explicitely described by Spehner in [21]. Actually, Markov proved in [13] that one can decide whether a submonoid of A^* is or not finitely presented.

In this paper, we prove a new result (Theorem 5) which represents a progress on finite sets of words under the restriction that they form a prefix code. It has consequences on properties of sets of words of small cardinality. The statement of this result will be given in the next section.

2 Finitely generated submonoids

In a pioneering paper, Lentin and Schützenberger [10] investigated the properties of sets $X = \{x, y\}$. Their main result is that X^* is star-free iff each word in $x^*y \cup xy^*$ is primitive. Recall that a set is called star-free if it can be obtained from the subsets of the alphabet by the operations of concatenation, union and absolute complement. It is known that a set is star-free if and only if its syntactic monoid is aperiodic. A monoid M is called aperiodic if there is an integer k such that $x^k = x^{k+1}$ for all $x \in M$ (see [6] for example). For example, if $X = \{ab, ba\}$, then X^* is star-free. Actually, one knows presently the answers of all questions concerning sets with two words (see below for more details).

Schützenberger introduced in [19] the following concept: a function f assigning to each finite set of words X an integer is said to be *absolutely bounded* if it can be bounded in terms of the number n of elements of X. For example, the cardinality of the syntactic monoid M of X^* is not absolutely bounded since M has d elements when $X = a^d$. He proved that the following functions are absolutely bounded:

- the number of conjugacy classes of groups in M.
- the order of the subgroups, apart from at most n conjugacy classes of cyclic groups.

The results of [19] rely on a combinatorial lemma on words. This lemma was further refined by Césari and Vincent [2] and by Duval [5] as the following statement, known as the critical factorization theorem. A word r is said to be a *repetition* of a pair (u, v) of words if

- r is a suffix of u or u is a suffix of r.
- r is a prefix of v or v is a prefix of r.

We say that the factorization (u, v) of a word x is *critical* if the length of the shortest repetition of (u, v) is the period of x. The critical factorization theorem can now be stated as follows (see [11] for a more precise formulation and [3] for a different proof).

Theorem 1. *Any word has a critical factorization.*

It is easy to deduce from the previous result the following one. An X-factorization of a word w is a sequence $(s, x_1, \ldots, x_k, p)$ with s a suffix of X, each x_i in X and p a prefix of X such that $w = sx_1 \cdots x_k p$. Two X-factorizations $(s, x_1, \ldots, x_k, p)$ and $(s', x'_1, \ldots, x'_{k'}, p')$ are disjoint if $sx_1 \cdots x_i \neq s'x'_1 \cdots x'_j$ for all $i = 0, \ldots, k$, $j = 0, \ldots, k'$.

Theorem 2. *A word w of period exceeding the maximum of the lengths of the elements of X has at most $n = Card(X)$ disjoint factorizations in words of X.*

We recall some notions on groups in the syntactic monoid of a finitely generated submonoid X^*. To simplify the presentation, we suppose that X is prefix. Let Q be the set of states of the minimal automaton of X^*. We denote by 1 the initial state and final state. Let $\varphi : A^* \to M$ be the morphism from A^* onto the syntactic monoid M of X^*. We consider M as a monoid of applications of Q into itself. Let G be a subgroup in M. Let $I \subset Q$ be the common image of the elements of G. Thus G is a permutation group on the set I. As a permutation group, its degree is the number of elements of I. Such a group G is called a *syntactic group* of X.

The subgroup G is called *special* if the submonoid $\varphi^{-1}(G)$ is cyclic (recall that a monoid is called cyclic if it generated by one element). A special subgroup is cyclic but the converse is not true. The degree of special subgroups is not bounded in terms of the number of elements of X. Indeed, when $X = a^d$, M is a cyclic group of order d (plus a zero if A has more than one element).

The following statement is now a consequence of the critical factorization theorem. It was proved in [19] in a weaker form (with $2n$ instead of n).

Theorem 3. *The degree of non special syntactic groups is at most n.*

In [15], the first author proposed a generalization of the result concerning the degrees of the syntactic groups as follows. Let X be a finite set of words with n elements. Let $\varphi : A^* \to M$ be the morphism from A^* onto the syntactic monoid of X^*. Let G be a group in M of degree d. Let r be the rank of the subgroup generated by $\varphi^{-1}(G)$. The conjecture is the inequality

$$n - 1 \geq d(r - 1). \tag{1}$$

The case $r = 1$ in the conjecture corresponds to cyclic groups. Thus the case of special syntactic groups is taken into account. For non-cyclic groups, one has $r \geq 2$ and thus the conjecture implies that $n - 1 \geq d$. This what we are going to prove, under the additional hypothesis that the set X is a prefix code.

The conjectured inequality is related to Schreier's formula. Recall that, according to this formula (see [7]), a subgroup of finite index d in a free group of rank r has rank

$$n = d(r-1)+1$$

In some cases (as in Examples 1 to 6), the set X is the basis of a subgroup H of index d of the free group F on A and the above equality holds. Moreover, the syntactic monoid of X^* contains in these cases a group G isomorphic to the representation of F on the cosets of H. The degree of G is equal to the index of H in G. Thus, in this case, inequality (1) is an equality resulting from Schreier's formula.

For the sake of completeness, we recall the main result of [15], which gives a sufficient condition to obtain a group G as a syntactic group of a finite set X. Let G be a transitive permutation group on a set Q and let H be the subgroup fixing a given point $1 \in Q$. Let ψ be a morphism from A^* onto a finite group G. Let Z be the basis of the submonoid $\psi^{-1}(H)$. We may choose for each letter $a \in A$ a word y_a such that

(i) $\psi(y_a) = \psi(a)$.

(ii) Only a finite number of words of Z have y_a as a factor.

Let Y be the set of the words y_a for $a \in A$. Then

Theorem 4. *The set X formed of the words of Z which are a factor of Y^* is a finite set such that G is a syntactic group of X.*

We will see in Section 4 examples of application of this result.

3 Main result

We can now state the main result of this paper.

Theorem 5. *Let $X \subset A^*$ be a prefix code with n elements. The degree of the non-special groups contained in the syntactic monoid of X^* is at most $n-1$.*

Proof. Suppose that M contains a non-special subgroup of degree d. We consider the set $\mathcal{J}$ of d-subsets of Q. Let $I \in \mathcal{J}$ be the set fixed by a non-special subgroup G. Since $\varphi^{-1}(G)$ is not cyclic, we can find two words u, v which are not powers of a common third one such that $I \cdot u = I \cdot v = I$. This implies that there exists a set $J \in \mathcal{J}$ and two distinct letters $a, b \in A$ such that $J.a$, $J.b$ are both elements of $\mathcal{J}$. Indeed since the words u, v are not powers of a third one, by taking appropriate powers of u, v if necessary, we may assume that the set $\{u, v\}$ is a prefix code. Then (see Figure 1), let r be the longest common prefix of u, v. We can write $u = rs$ and $v = rt$ for some non-empty words s, t that do not begin with the same symbol. We can choose $J = I \cdot r$. Thus, the elements of the set J are d states p of the automaton such that at least two transitions from state p by symbols a and b exist. Thus, the set of prefixes of X is a tree such that at least d nodes have at least two sons. This implies that it has at least $d+1$ leaves. Thus X has at least $d+1$ elements.

We may observe that the argument used in the above proof is related with the one used in [14] in which *positive* and *negative* nodes in the graph associated with the action on the cosets of a subgroup are introduced. The positive nodes are those with two outgoing edges and correspond to our nodes above with at least two sons.

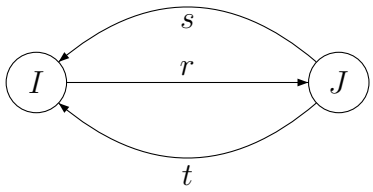

Figure 1: The action on d-subsets.

4 Examples

We present several examples of syntactic groups of finitely generated submonoids, the first two with a syntactic group which is the symmetric group S_3 and the third one with the dihedral group D_4.

Example 1. Let $A = \{a, b\}$ and $X = \{aaa, aaba, ab, baa\}$. The minimal automaton of X^* is represented on Figure 2.

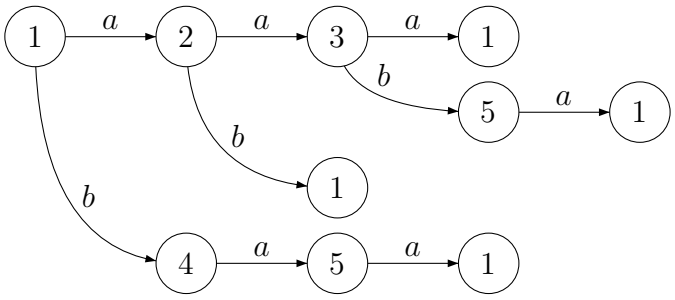

Figure 2: The minimal automaton of $\{aaa, aaba, ab, baa\}^*$

The action on 3-subsets is represented on Figure 3. Since a induces the cycle (123) and baa the transposition (23), M contains a group S_3. The node labeled 123 on Figure 3 has outdegree 2. Correspondingly, in the automaton of Figure 2, the states $1, 2, 3$ have outdegree 2.

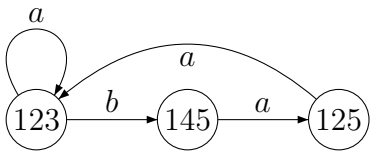

Figure 3: The action on 3-subsets

This example (and the two following ones) is actually a particular case of Theorem 4 with $\psi(a) = (123)$, $\psi(b) = (12)$, $y_a = a^4$, $y_b = a^3b$.

Example 2. Let $A = \{a, b\}$ and $X = \{aaba, ab, baa, baba\}$. The minimal automaton of X is represented on Figure 4. The action on 3-subsets is represented on Figure 5.

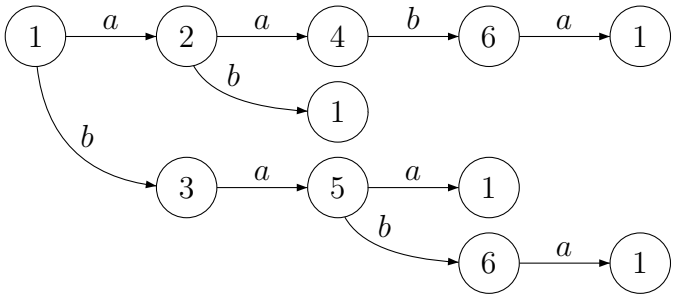

Figure 4: The minimal automaton of $\{aaba, ab, baa, baba\}^*$

The node labeled 125 has outdegree 2. Correspondingly, in the automaton of Figure 4, the states $1, 2, 5$ have outdegree 2.

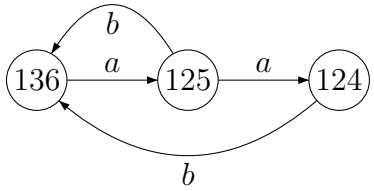

Figure 5: The action on 3-subsets

Example 3. Let $A = \{a, b\}$ and $X = \{aa, abaaba, abab, baab, baba\}$. Let $2 = 1.a$, $3 = 1.ba$, $4 = 2.ba$. Then $\{1, 2, 3, 4\}$ is stabilized by ba and aba which generate a group D_4.

5 Sets with two or three words

We can say that everything concerning sets with two words is known. First, except when x, y are power of a single word (i.e. when X^* is commutative), the set $X = \{x, y\}$ is code with bounded delay. This is easy to prove by induction on $|x| + |y|$.

Also, any code with two elements is obtained by a composition of prefix and suffix codes (and consequently included in a finite maximal code) [16].

The situation is already more complicated with three words. For example, there exists three elements sets which are not obtained by composition of prefix or suffix codes [4], a question raised in [16]. The simplest example is

$$X = \{a, aba, babaab\}$$

which is however included in a finite maximal code, namely the code Y given by

$$Y - 1 = (1 + b + baba(1 + a + b))(a + b - 1)(1 + ba).$$

This finite maximal code belongs to the family described in [1], Exercise VIII.1.4, p. 419. A simpler way to show that Y is a maximal code is to use the following lemma, easy to verify.

Lemma 1. *Let Z be a finite maximal code such that $Z - 1 = P(A-1)Q$ for two subsets $P, Q \subset A^*$. Let $x \in Z$ and $U, R \subset A^*$ be such that*

(i) *the set U is prefix closed and R is the prefix code such that $R - 1 = U(A-1)$.*

(ii) *$Q \subset 1 + R$.*

Then

$$Y = 1 + (P + xU)(A-1)Q.$$

is a finite maximal code

The above example corresponds to the case $P = 1 + b$, $Q = 1 + ba$, $x = baba$, $U = 1 + a + b$.

It is not known whether any code with three words is contained in a finite maximal code. Derencourt proposes in his paper [4] the example of

$$X = \{a, aba, baababababaab\}$$

as a possible counterexample. Since then, he was able [1] to complete this code in the code Y given by

$$Y = 1 + P(a + b - 1)Q$$

with

$$\begin{aligned} P &= 1 + ab + ba^2bab(1 + a + ab + aba + aab) \\ Q &= 1 + b + (1 + a + b)abab \end{aligned}$$

Theorem 5 has the following consequences for prefix codes with two or three words. First, it shows directly that the syntactic groups of two elements prefix codes are all special. Indeed, a non special group has to be of degree one, i.e. trivial. This is actually already known and true for all codes, prefix or not. Moreover, there can be at most two non-trivial special subgroups, corresponding to imprimitive words in the set $\{x, y, xy^+\}$ (see [9]).

Theorem 6. *The syntactic groups of any three-element prefix code are cyclic.*

This is indeed true since the syntactic groups are either special, and thus cyclic, or they are of degree at most two and again cyclic.

A sophisticated statement is that all syntactic groups of a prefix code with at most five elements are all solvable. Indeed, such a syntactic group G is either cyclic or of degree at most 4. But a group of degree 4 is included in the symmetric group of degree 4, which is solvable (see [7] for example). Thus G is solvable.

[1] personal communication of Michel Latteux, june 2003

6 Possible generalizations

There are two directions in which Theorem 5 can possibly be generalized. One is to suppress the hypothesis that the set is a prefix code. However, we do not know of an example of a set of minimal cardinality $d+1$ giving rise to syntactic group of degree d which is not prefix or suffix. In all the examples shown before, the set X is biprefix. An example of a minimal set giving rise to a syntactic group $\mathbb{Z}/2$ and which is not biprefix is given in the following example.

Example 4. Let

$$X = \{abb, bab, bb\}.$$

The set X is not suffix since bb is a suffix of abb. The minimal automaton of X^* is shown on Figure 6. The group $\mathbb{Z}/2$ is a syntactic group since the set $\{1, 4\}$ is a

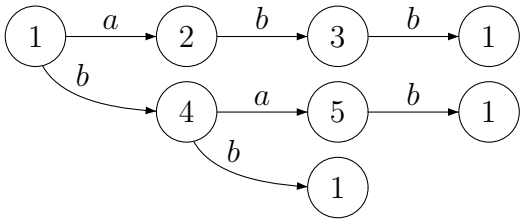

Figure 6: The minimal automaton of $\{abb, bab, bb\}^*$

fixpoint of the action of b and abb.

All minimal sets that we know are bases of subgroups of the free group. More precisely the set X is the basis of a subgroup H of the free group on A and the syntactic group G is the permutation group induced by the action of A on the cosets of H.

We warn the reader that the converse is not always true. Indeed there exist sets $X \subset A^+$ which are the basis of a subgroup of finite index of the free group on A such that X^* is aperiodic. An example of this situation is given below.

Example 5. The set $X = \{aab, baa, bab, bba\}$ is the basis of a subgroup of index 3 of the free group on $\{a, b\}$. However, one may verify by computing the action of the symbols on the subsets of the set of states of the automaton of X^* represented on Figure 7 that the submonoid X^* is aperiodic.

However, the sets of minimal size giving rise to syntactic groups seem to be closely related with bases of subgroups.

The other generalization of Theorem 5 is to study the role of the rank of the group for $r \geq 3$.

In all the examples of Section 4 the rank of the subgroup $H = \varphi^{-1}(G)$ is always 2. The subgroup H is actually equal to the free group on $A = \{a, b\}$. Taking an

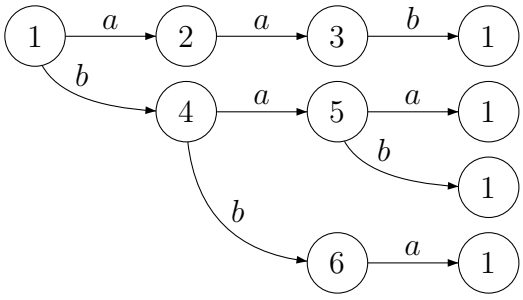

Figure 7: The minimal automaton of $\{aab, baa, bab, bba\}^*$

alphabet with three symbols, one obtains examples of sets with n elements having a syntactic group of degree d for which the formula $n - 1 = 2d$ holds. For example, if

$$X = \{aa, ab, ac, ba, ca\}$$

the group $\mathbb{Z}/2$ is a syntactic group.

It is more difficult to find examples in which the syntactic group G itself has rank 3, which forces $\varphi^{-1}(G)$ to have also rank 3. The following example is from [17]

Example 6. Let X be the set formed of the following words listed in alphabetic order.

aa
$abaaba$
$abacaabaca$
$abacabaabacaba$
$abacabac$
$acaabacaba$
$acabaabaca$
$acabacaaba$
$acabacab$
$baab$
$bacaabac$
$bacabaabacab$
$bacabaca$
$caabacab$
$cabaabac$
$cabacaab$
$cabacaba$

The set X is a set of $n = 17$ words on an alphabet of 3 symbols such that the group $(\mathbb{Z}/2)^3$ of degree 8 and rank $r = 3$ is a syntactic group. The set is of minimal cardinality since $n - 1 = d(r - 1)$. Actually, X is the basis of a subgroup of index 8 of the free group on $\{a, b, c\}$.

The tree corresponding to the prefix code X in the classical way has 2 nodes of degree 3 (ϵ and a) and 12 nodes of degree 2 (aba, ba, ca,...) consistently with the number of leaves which is $3 \times 2 + 12 - 1 = 17$. Actually, these nodes correspond to the states appearing in the action on 8-subsets. The situation is pictured on Figure

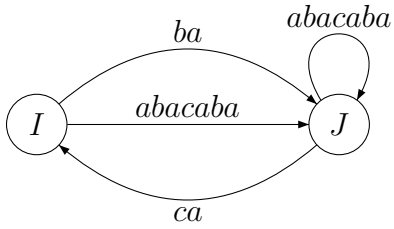

Figure 8: The action on 3-subsets

8 with

$$\begin{aligned} I &= \{\epsilon, a, ca, aca, baca, abaca, cabaca, acabaca\} \\ J &= \{\epsilon, a, ba, aba, caba, acaba, bacaba, abacaba\} \end{aligned}$$

The contents of nodes I and J correctly gives the count of 2 nodes of degree 3 (corresponding to $I \cap J$) and 12 nodes of degree 2 (corresponding to $I \cup J - (I \cap J)$). Unfortunately, we are not able to prove that the argument works in general.

This example is again a case where Theorem 4 applies with Z^* the set of words on $\{a, b, c\}$ with an even number of a, b and c, and the words $y_a = (abac)^2a$, $y_b = (abac)^2aba$, $y_c = (abac)^3aba$.

References

[1] Jean Berstel and Dominique Perrin. *Theory of Codes*. Academic Press, 1985.

[2] Yves Césari and Max Vincent. Une caractérisation des mots périodiques. *C. R. Acad. Sci. Paris Sér. A-B*, 286(24):A1175–A1177, 1978.

[3] Maxime Crochemore and Wojciech Rytter. *Text Algorithms*. The Clarendon Press Oxford University Press, New York, 1994.

[4] Denis Derencourt. A three-word code which is not prefix-suffix composed. *Theoret. Comput. Sci.*, 163(1-2):145–160, 1996.

[5] Jean-Pierre Duval. Périodes et répétitions des mots du monoïde libre. *Theoret. Comput. Sci.*, 9(1):17–26, 1979.

[6] Samuel Eilenberg. *Automata, Languages and Machines*, volume B. Academic Press, 1976.

[7] Marshall Hall. *Theory of Groups*. Chelsea, 1976.

[8] Juhani Karhumäki. Combinatorial and computational problems on finite sets of words. In *Machines, computations, and universality (Chişinău, 2001)*, volume 2055 of *Lecture Notes in Comput. Sci.*, pages 69–81. Springer, Berlin, 2001.

[9] Evelyne Le Rest and Michel Le Rest. *Sur les Relations entre un Nombre Fini de Mots*. PhD thesis, 1979.

[10] A. Lentin and M. P. Schützenberger. A combinatorial problem in the theory of free monoids. In *Combinatorial Mathematics and its Applications (Proc. Conf., Univ. North Carolina, Chapel Hill, N.C., 1967)*, pages 128–144. Univ. North Carolina Press, Chapel Hill, N.C., 1969.

[11] M. Lothaire. *Combinatorics on words*. Cambridge University Press, Cambridge, 1997.

[12] M. Lothaire. *Algebraic combinatorics on words*. Cambridge University Press, Cambridge, 2002.

[13] Al. A. Markov. On finitely generated subsemigroups of a free semigroup. *Semigroup Forum*, 3(3):251–258, 1971/1972.

[14] J. Meakin and P. Weil. Subgroups of free groups: a contribution to the Hanna Neumann conjecture. In *Proceedings of the Conference on Geometric and Combinatorial Group Theory, Part I (Haifa, 2000)*, volume 94, pages 33–43, 2002.

[15] Dominique Perrin. Sur les groupes dans les monoïdes finis. In *Noncommutative structures in algebra and geometric combinatorics (Naples, 1978)*, pages 27–36. CNR, Rome, 1981.

[16] A. Restivo, S. Salemi, and T. Sportelli. Completing codes. *RAIRO Inform. Théor. Appl.*, 23(2):135–147, 1989.

[17] Giuseppina Rindone. Sur les groupes syntaxiques d'un langage. *RAIRO Inform. Théor.*, 19(1):57–70, 1985.

[18] Giuseppina Rindone. Construction d'une famille de codes associés à certains groupes finis. *Theoret. Comput. Sci.*, 54(2-3):165–179, 1987.

[19] M.-P. Schützenberger. A property of finitely generated submonoids of free monoids. In *Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976)*, pages 545–576. North-Holland, Amsterdam, 1979.

[20] Marcel-Paul Schützenberger. Sur les sous-groupes de rang fini d'un groupe libre. *C. R. Acad. Sci. Paris Sér. A-B*, 290(5):A207–A208, 1980.

[21] J.-C. Spehner. Les présentations des sous-monoïdes de rang 3 d'un monoïde libre. In *Semigroups (Proc. Conf., Math. Res. Inst., Oberwolfach, 1978)*, volume 855 of *Lecture Notes in Math.*, pages 116–155. Springer, Berlin, 1981.

Institut Gaspard Monge
Université de Marne-la-Vallée
F-77454 Marne-la-Vallée cedex
email : perrin@univ-mlv.fr