INFORMATIQUE THÉORIQUE ET APPLICATIONS

MARTINE LEONARD

A property of biprefix codes

Informatique théorique et applications, tome 22, nº 3 (1988), p. 311-318

<http://www.numdam.org/item?id=ITA_1988__22_3_311_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 22, n° 3, 1988, p. 311 à 318)

A PROPERTY OF BIPREFIX CODES (*)

by Martine LEONARD (1)

Communicated by J. E. PIN

Abstract. – *Let X be a recognizable biprefix code on A^*. A word w of A^* is full with respect to X if for any factorization $w=w_1 w_2$, w_1 (resp. w_2) is a suffix (resp. prefix) of some word of X^*. Given X, we present a necessary and sufficient condition for the set of full words with respect to X to be finite.*

Résumé. – *Soit X un code bipréfixe reconnaissable sur A^*. Un mot w de A^* est plein pour X si, pour toute factorisation $w=w_1 w_2$, w_1 (resp. w_2) est un suffixe (resp. préfixe) d'un mot de X^*. Nous présentons une condition nécessaire et suffisante pour que, pour un X donné, l'ensemble des mots pleins soit fini.*

INTRODUCTION

The property which we present here is used in an algorithm constructing finite biprefix codes [6]. Biprefix codes were first studied by M. P. Schutzenberger in 1956 and Gilbert and Moore in 1959 [5]. These codes have remarkable properties which are presented and developped in the "Theory of codes" of J. Berstel and D. Perrin [1]. It is agreed that investigating biprefix codes will help in studying more general (neither prefix nor suffix) codes. Y. Cesari proposed two effective constructions of finite maximal biprefix codes [2, 3]. The first one was implemented in 1975 by C. Precetti and has some disadvantages because each code can be obtained several times. We have implemented the second one in 1985 and then, we have enumerated all finite maximal biprefix codes up to degree 5 on a two letter alphabet. We have found that there are 5 086 783 such codes [6]. It is essential when constructing finite maximal biprefix codes to know whether the set of full words is finite or not. Here, we give a method to answer this question.

(*) Received in July 1987, revised in November 1987.
(1) Faculté des Sciences, Laboratoire d'Informatique, B.P. n° 67, 76134 Mont-Saint-Aignan.

BASIC DEFINITIONS

From now, we consider a finite alphabet.

Let A be a finite alphabet and let A^* be the set of all words on A.

We denote by 1 the empty word and by $A^+ = A^* - 1$ the free semigroup.

A subset X of A^* is a *code* on A^* if each word of A^* has at most one factorization in words of X, i. e. if $x_1 \ldots x_n = x'_1 \ldots x'_m$, x_i, $x'_j \in X$, $\forall i = 1, \ldots, n$, $\forall j = 1, \ldots, m$ implies $n = m$ and $x_1 = x'_1, \ldots, x_m = x'_m$.

Let u and w be in A^*, u is a *prefix* (resp. *suffix*) of w if u is a left (resp. right) factor of w.

By convention, a prefix (resp. a suffix) u of a word w is *proper* if it is different from w. Let X be a subset of A^* and let X^* be the submonoid which it generates i. e. the set of products of words of X. We denote by XA^- the set $\{w \in A^* / w A^+ \cap X \neq \varnothing\}$ of the proper prefixes of all words in X and dually by $A^- X$ the set of the proper suffixes of all words in X.

A subset X is a *biprefix code* if it does not contain any proper prefix or suffix of its words or equivalently if $X \cap XA^- = \varnothing$ and $X \cap A^- X = \varnothing$.

Example 1 :

$$A = \{a, b\}$$

$$X = \{ab, baa, babb\} \text{ is a biprefix code on } A^* :$$

$$XA^- = \{1, a, b, ba, bab\} \Rightarrow X \cap XA^- = \varnothing$$

$$A^- X = \{1, a, aa, abb, b, bb\} \quad \Rightarrow \quad X \cap A^- X = \varnothing$$

X is a *maximal code* if it is not properly included in another code and X is a *maximal biprefix code* if X is biprefix and it is not properly included in another biprefix code. For example, for each $n > 0$, A^n is a maximal biprefix code, where A^n is the set of all the words of A^* whose length is n. A^n is called the *uniform code*.

An *X-interpretation* of w is a triple (v, x, u), where $v \in A^- X$, $x \in X^*$, $u \in XA^-$, such that $vxu = w$.

Example 2 :

$$A = \{a, b\}$$

$$X = \{ab, baa, babb\}, \qquad w = babbabb$$

$$(b, ab \,.\, babb, 1) \text{ is an } X\text{-interpretation of } w.$$

A *point* of w is a pair $(r, s) \in A^* \times A^*$ such that $rs = w$. We say that an X-interpretation (v, x, u) of w *passes through the point* (r, s) of w if there exist $y, z \in X^*$ with $w = rs$, $r = vy$, $s = zu$ and $x = yz$.

In example 2, $(b, ab.babb, 1)$ passes through the points $(b, abbabb)$, $(bab, babb)$ and $(babbabb, 1)$ of w.

A word w is *full with respect to* a subset X of A^* if, for each point of w, there exists an X-interpretation passing through this point. In example 2, w is not full because we cannot find an X-interpretation passing through the point $(ba, bbabb)$ because ba is not a suffix of any word in X^*.

Definitions and results on full words are developped in [1], [3] and [6].

Results

Let $X \subset A^*$ be a biprefix code.

We denote by $FW(X)$ the set of words which are full with respect to X and by $S(X)$ the set of words of A^* any prefix of which is a suffix of some word in X^*.

PROPOSITION [6]: *Let X be a biprefix code on A^*. Then $FW(X)$ is included in $S(X)$.*

Proof: Let w be a word which is full with respect to X and let p be a prefix of w. Consider q in A^* with $pq = w$. There exists an X-interpretation (v, x, u) passing through the point (p, q) of w, i.e., there exist $y, z \in X^*$ with $p = vy$ and $q = zu$. By definition, v is a suffix of some word in X and y is in X^*, therefore $p = vy$ is a suffix of some word in X^* and w is in $S(X)$. ■

Example 3: The following example shows that the converse is not true. Let $A = \{a, b\}$ and $X = \{aab, abb\}$. Then X is a finite biprefix code. The word $w = bb$ is in $S(X)$ because the set $\{1, b, bb\}$ of prefixes of w is included in the set of suffixes of X^*. But we cannot find any X-interpretation passing through the point $(1, bb)$ of w neither through the point (b, b) because b and bb are not prefixes of any word in X.

THEOREM [6]: *Let X be a recognizable biprefix code on A^*. Then $FW(X)$ is finite if and only if $S(X)$ is finite.*

Proof: We know that $FW(X) \subset S(X)$. Therefore, if $S(X)$ is finite then $FW(X)$ is also finite.

Now, assume that $S(X)$ is infinite. We first show that $S(X)$ is recognizable. Since X is recognizable, then X^* is recognized by some finite not necessarily deterministic automaton $\mathscr{A}$ whose set of states is Q. Without loss of generality, we may assume that the initial and terminal states are reduced to a unique state which we denote by 1. We obtain an automaton $\mathscr{B}$ recognizing $S(X)$ by applying the usual subset construction [4] restricted to the subsets of Q which contain the special state 1. More precisely, a triple (P_1, a, P_2) is a transition of $\mathscr{B}$ iff P_1 and P_2 are both subsets of Q containing 1, a is a letter of A and P_2 is the set of all the states of Q which are accessible from P_1 by the letter a. Moreover, the initial state of $\mathscr{B}$ is the set Q and all states of $\mathscr{B}$ are terminal. Now we clain that $\mathscr{B}$ actually recognizes $S(X)$:

LEMMA: *$\mathscr{B}$ is a finite automaton recognizing $S(X)$.*

Proof: As usual, we denote by $|\mathscr{B}|$ the behavior of $\mathscr{B}$, i. e. the set of all the words recognized by $\mathscr{B}$. Let us first make two observations:

(1) Because every state of $\mathscr{B}$ is terminal, all prefixes of $|\mathscr{B}|$ belong to $|\mathscr{B}|$.

(2) There exists a path labelled by w in A^* leading from P_1 to P_2 in $\mathscr{B}$ iff for every state p_2 of P_2 there exists a state $p_1 \in P_1$ and a path labelled by w leading from p_1 to p_2 in $\mathscr{A}$, and there exists a state $p_2 \in P_2$ such that $p_2 = 1$.

The equality $|\mathscr{B}| = S(X)$ is easily checked by induction on the length of $w \in A^*$. Indeed, for $w = 1$, the result is trivially true. Assume that the result is true for some $w' \in A^*$ and let $a \in A$ and $w \in A^+$ such that $w = w' a$. If $w \in |\mathscr{B}|$ then by (2), there exist $p \in Q$ and a path labelled by w leading from p to 1 in $\mathscr{A}$.

This means that w is a suffix of some word in X^*. By (1) this is true for each prefix of w. Therefore w is in $S(X)$.

Conversely, assume that $w \in S(X)$, with $w = w' a$, $a \in A$. Because of the definition of $S(X)$, $w' \in S(X)$. By induction, $w' \in |\mathscr{B}|$. There exist $P' \subseteqq Q$ with $1 \in P'$ such that $Q . w' = P'$. As w is in $S(X)$, there exists $p' \in P'$ such that a path labelled by the letter a leads from p to the state 1 in $\mathscr{A}$. Thus, there exists $P \subseteqq Q$ such that $Q . w = P$ with $1 \in P$ and $w \in |\mathscr{B}|$.

Let us return to the proof of the theorem.

We have just shown in the lemma that the infinite set $S(X)$ is recognized by a finite automaton. Then, there exist $u, v, w \in A^*$, $v \neq 1$, such that $uv^* w \subseteqq S(X)$.

Let n be an integer and r be a prefix of v^n. ur is a prefix of $uv^n w$ and $uv^n w \in S(X)$. Then ur is a suffix of some word of X^* and r too, thus v^* is included in $S(X)$.

We are now going to show that $v^* \subset FW(X)$. We first define a bijection f from the set of all prefixes of v onto itself in the following way:

Let p be a prefix of v. We set $f(p)=p'$, where according to whether p has a suffix in X or p is a suffix of some word in X we have:

either

(1) $\exists x \in X$ such that $p' x = p$, i. e. $p \in A^* X$

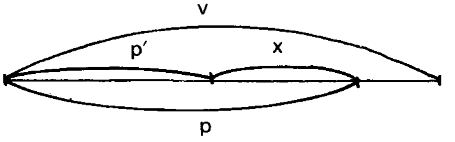


Figure 1.

or

(2) $\exists n \geqq 0$ and $q' \in A^*$ such that $v = p' q'$ and $q' v^n p \in X$, i. e. that $p \in A^- X$.

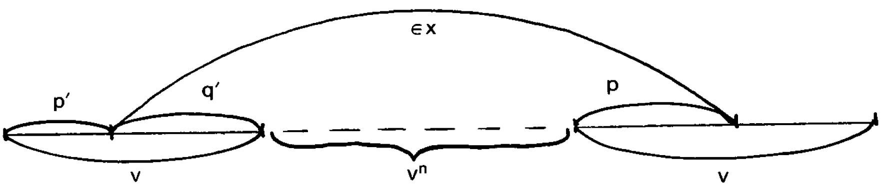


Figure 2.

We have first to show that, for each prefix p of v, p' exists and is unique.

If $p \in A^* X$, there exist a prefix p' of v and $x \in X$ such that $p = p' x$ and p' is unique because X is biprefix.

Assume now that $p \in A^* - A^* X$. We are going to show that there exist $n \geqq 0$, $q', p' \in A^*$ such that $v = p' q'$ and $q' v^n p \in X$. Assume that p' does not exist. As $v^* \subset S(X)$, then for each $n \geqq 0$, $v^n p$ is a suffix of some word of X, i. e. $\forall n \geqq 0$ there exists $w_n \in A^*$ such that $w_n v^n p \in X$.

Let $\mathscr{A} = (Q, I, T)$ be an automaton recognizing X^* with $I = T = \{i\}$, i. e. i is the unique initial and terminal state of $\mathscr{A}$. Since $w_n v^n p \in X$, there exists $q \in Q$ such that $i . w_n v^n p = q v^n p = i$. Because of the finiteness of Q, we can find $n, m \geqq 0$ with $n < m$ such that:

$$i . w_n v^n p = q . v^n p = i$$
$$i . w_m v^m p = q . v^m p = i.$$

Then $w_n v^n p \in X$ is a prefix of $w_n v^m p \in X$, leading to a contradiction because X is prefix. Therefore p' exists. Moreover p' is unique because X is biprefix.

We have now to show that f is a bijection.

Assume that there exist three prefixes p, q, p' of v such that $f(p)=f(q)=p'$. It is easy to see that in each case (see fig. 3, 4, 5) we can find two words x

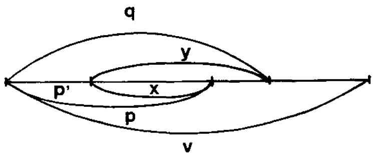


Figure 3. – **If p and q satisfy (1).**

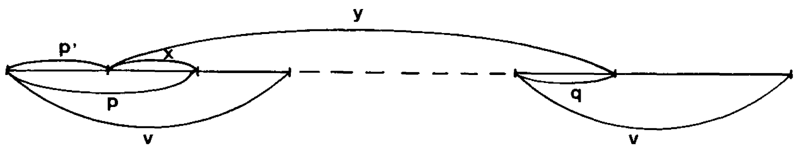


Figure 4. – **If p satisfies (1) and q satisfies (2).**

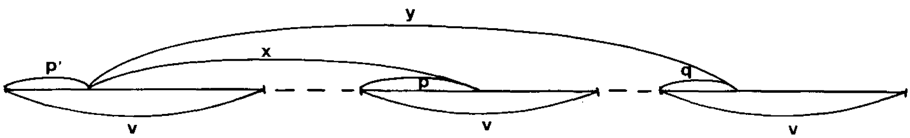


Figure 5. – **If p and q satisfy (2).**

and y in X such that x is a prefix of y which is impossible because X is biprefix. Thus f is one-one and is defined everywhere. Since the set of prefixes of v is finite, f is bijective.

We now show that, for each integer $n \geqq 0$, the word v^n is a full word with respect to X.

Let n be an integer and (u, w) be a point of v^n.

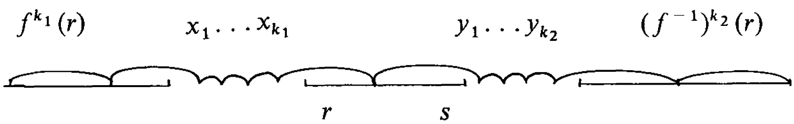
There exist $m \leqq n$, a prefix r of v, a suffix s of v such that

$$v=rs, \qquad u=v^m r, \qquad w=sv^{n-m-1}.$$

There exist $k_1, k_2 \geqq 0$, $x_1, x_2, \ldots, x_{k_1}, y_1, y_2, \ldots, y_{k_2} \in X$ such that, by iteration of the application f,

$$(f^{k_1}(r), x_1 x_2 \ldots x_{k_1} \cdot y_1 y_2 \ldots y_{k_2}, (f^{-1})^{k_2}(r))$$

is an X-interpretation of v^n:



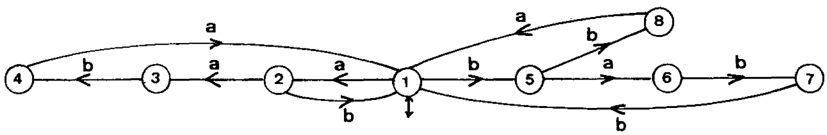
Then, for each $n \geqq 0$, v^n is in $FW(X)$ and thus $FW(X)$ is infinite. ■

As a conclusion, checking whether $FW(X)$ is finite amounts to checking whether $S(X)$ is finite. Since $S(X)$ is recognized by a finite automaton $\mathscr{B}$, it is infinite if and only if $\mathscr{B}$ contains a cycle. It thus suffices to construct $\mathscr{B}$ as explained above and to check whether it has a cycle

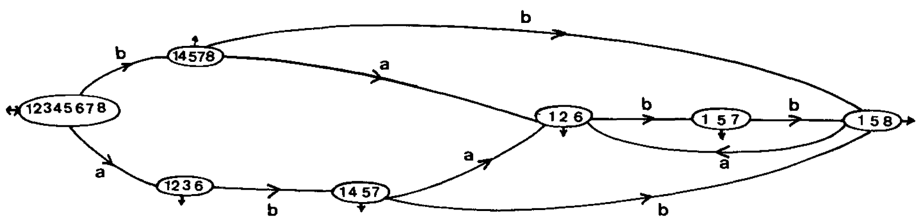
Example 4: Let $A=\{a, b\}$, $X=\{ab, aaba, babb, bba\}$.

X is a finite biprefix code because no word in X is a proper prefix or suffix of another word of X.

Automaton $\mathscr{A}$:



Automaton $\mathscr{B}$:



$\mathscr{B}$ contains a unique cycle and its label is bba. Then $(bba)^* \subset FW(X)$ and $FW(X)$ is infinite.

The algorithm which allows to obtain $\mathscr{B}$ is described in [6] where an estimate of its cost is given.

REFERENCES

1. J. Berstel and D. Perrin, *Theory of Codes*, Academic Press, 1985.
2. Y. Cesari, *Sur un algorithme donnant les codes bipréfixes finis*, Math. Syst. Theory, Vol. 6, 1972, pp. 221-225.
3. Y. Cesari, *Propriétés Combinatoires des codes bipréfixes complets finis*, in *Actes de la Septième École de Printemps d'Informatique Théorique*, Jougne, 1979, D. Perrin ed., LITP.
4. S. Eilenberg, *Automata, Languages and Machines*, Academic Press, Vol. A, 1974, p. 32.
5. E. N. Gilbert and E. F. Moore, *Variable-length Binary Encodings*, Bell. Syst. Techn., J38, 1959, pp. 933-967.
6. M. Léonard, *Thèse de Doctorat ès Sciences*, Université de Rouen, 1985.