RAIRO. INFORMATIQUE THÉORIQUE

A. RESTIVO

Mots sans répétitions et langages rationnels bornés

RAIRO. Informatique théorique, tome 11, nº 3 (1977), p. 197-202
<http://www.numdam.org/item?id=ITA_1977__11_3_197_0>

R.A.I.R.O. Informatique théorique/Theoretical Computer Science
(vol. 11, n° 3, 1977 p. 197 à 202).

MOTS SANS RÉPÉTITIONS ET LANGAGES RATIONNELS BORNÉS (1)

par A. RESTIVO (*)

Communiqué par J. F. Perrot

Résumé. — *On démontre que les trois conditions suivantes sont équivalentes pour un langage rationnel L:*

(i) *L est borné;*

(ii) *pour tout entier k, il existe un entier p tel que tout mot de L dont la longueur est au moins p contienne un facteur v^k, avec v non-vide;*

(iii) *L possède un nombre fini de facteurs itérants primitifs.*

I. INTRODUCTION

Un des chapitres les plus intéressants de la théorie combinatoire du monoïde libre concerne les « régularités » que peuvent présenter certains mots (sur un alphabet fini) lorsque leur longueur devient de plus en plus grande : par exemple, G. Th. Guilbaud a montré, en traduisant le célèbre thèorème de Van der Waerden sur les suites d'entiers, que tout mot suffisamment long sur un alphabet fini doit contenir une « cadence », i.e. la répétition d'une même lettre à intervalles réguliers [5]. Ainsi, lorsque l'alphabet X ne contient que deux lettres, on vérifie sans peine que tout mot f de longueur supérieure ou égale à 4 contient au moins deux facteurs égaux, i.e. $f = uv^2w$ (f contient un carré). Il est remarquable que ce phénomène ne se produise plus si X a au moins trois lettres, ni pour des puissances supérieures à deux. En effet A. Thue en 1912, dans un long article consacré à l'étude des propriétés des suites, finies ou infinies, de symboles pris dans un alphabet fini X, a établi en particulier le résultat suivant [7].

(*) Reçu en janvier 1977.

(1) Instituto di Matematica dell'Universita di Palermo et Unita di Ricerca GIFCO del CNR, Palermo.

Soit $L_k(X)$ l'ensemble des mots, sur l'alphabet X, ne contenant pas de puissance k-ième :

$$L_k(X) = \{ f \mid \forall u, v, w \in X^*, f = uv^k w \Rightarrow v = 1 \}$$

(1 *désigne ici le mot vide ou élément neutre de* X^*). *Alors, si* $Card(X) = 2$, $L_3(X)$ *est infini; si* $Card(X) = 3$, $L_2(X)$ *est infini.*

Ce résultat a été maintes fois redécouvert dans des contextes tout à fait différents comme la théorie des semigroupes et la dynamique symbolique. Pour une bibliographie complète et un raffinement du résultat on se reportera au travail récent de F. Dejean [1].

Dans une perspective toute différente on a beaucoup étudié sous le nom de *langages bornés*, des langages dont les mots se caractérisent par la répétition de certains facteurs. Plus précisément un langage L est borné s'il existe un ensemble fini de mots $w_1, w_2, \ldots, w_p$ tel que $L \subset w_1^* w_2^* \ldots w_p^*$. Chaque mot $f \in L$ est alors de la forme $f = w_1^{n_1} w_2^{n_2} \ldots w_p^{n_p}$ et il est univoquement déterminé par la suite $(n_1, n_2, \ldots, n_p)$ des répétitions des facteurs $w_1, w_2, \ldots, w_p$ respectivement.

Dans le présent article on établit une liaison entre le phénomène découvert par Thue et les langages bornés. On fait d'abord la remarque élémentaire suivante : étant donné un langage L borné quelconque, L ne contient pas de « mots sans répétitions » arbitrairement longs, i.e. il vérifie la condition

(B) $\qquad L \cap L_k$ est fini pour tout entier k.

Cette remarque permet alors de prouver l'existence de langages non bornés comme conséquence immédiate du théorème de Thue (Ginsburg et Spanier ont prouvé le même résultat de manière directe [3]).

Il n'est pas généralement vrai que tout langage qui vérifie la condition (B) soit borné. Notre résultat principal établit une telle propriété dans le cas des langages rationnels.

THÉORÈME 1 : *Le langage rationnel L est borné si, et seulement si, $L \cap L_k$ est fini pour tout entier k.*

REMARQUE 1 : La condition du théorème 1 peut être formulée ainsi : un langage rationnel L est borné si, et seulement si, pour tout entier k, il existe un entier p tel que, si un mot $f \in L$ est de longueur $|f| > p$, alors $f = uv^k w$ pour quelques $u, v, w \in X^*$ avec $v \neq 1$.

REMARQUE 2 : Le théorème 1 n'est pas plus vrai si on ne suppose pas L rationnel. Soit $L = \{ fx^{|f|} \mid f \in \{ x, y \}^* \}$. L est algébrique, n'est pas borné et vérifie la condition (B).

Dans le cours de la preuve du théorème 1, les phénomènes d'itération dans les langages rationnels jouent un rôle fondamental et on obtient de nouvelles caractérisations des langages rationnels bornés. On établit en particulier que les langages rationnels ayant un nombre fini de facteurs itérants primitifs sont exactement les langages rationnels bornés. On montre aussi que la cardinalité de l'ensemble des facteurs itérants primitifs d'un langage rationnel L, si elle est finie, est bornée par l'index de la congruence syntactique induite par L. Ce résultat permet alors de donner une nouvelle procédure pour décider si un langage rationnel est borné.

II. DÉFINITIONS ET RÉSULTATS PRÉLIMINAIRES

Dans la suite on dénotera par X un alphabet fini et par X^* le monoïde libre engendré par X. Les éléments de X seront appelés *lettres*, les éléments de X^* *mots*; on dénotera par le symbole $|f|$ la longueur du mot $f \in X^*$ et par le symbole 1 le mot vide (élément neutre) de X^*.

Si L est une partie de X^*, la congruence syntactique σ_L induite par L et la congruence droite ρ_L induite par L sont définies sur les mots de X^* respectivement de la manière suivante :

$$v \, \sigma_L \, v' \quad \text{si et seulement si} \quad \forall u, w \in X^* \; \{ uvw \in L \Leftrightarrow uv'w \in L \}$$
$$v \, \rho_L \, v' \quad \text{si et seulement si} \quad \forall w \in X^* \; \{ vw \in L \Leftrightarrow v'w \in L \}.$$

Un langage L est *rationnel* si, et seulement si, σ_L (resp. ρ_L) est d'index fini. L'index de la congruence droite ρ_L donne en plus le nombre d'états de l'automate minimal acceptant L.

Une des propriétés les plus intéressantes des langages rationnels concerne les phénomènes d'itérations décrits par le lemme suivant.

LEMME 1 : *Soit L un langage rationnel et r l'index de la congruence droite ρ_L induite par L. Si $w_1 w_2 \ldots w_r \in L$, alors il existe deux entiers distincts i et j tels que :*

$$w_1 \ldots w_i (w_{i+1} \ldots w_j)^* w_{j+1} \ldots w_r \subseteq L.$$

Dans le cours de la preuve des résultats on utilise encore le lemme technique suivant.

LEMME 2 : *Étant donné deux alphabets finis X et Y et un monomorphisme α de X^* dans Y^*,*

$$f \in L_k(X) \Rightarrow \alpha f \in L_{mk}(Y)$$

avec $m = \sup \{ |\alpha x| \mid x \in X \}$.

Preuve : La preuve est par contradiction; c'est-à-dire, on vérifie que, si $\alpha f \notin L_{mk}(Y)$, alors $f \notin L_k(X)$. On suppose alors qu'il existe trois mots u, v, $w \in Y^*$ tels que

$$\alpha f = u v^{mk} w.$$

Soit alors

$$f = x_1 x_2 \ldots x_p \quad \text{avec} \quad x_i \in X$$
$$\alpha f = a_1 a_2 \ldots a_p \quad \text{avec} \quad a_i = \alpha x_i$$

Dans le monoïde libre Y^* on a l'égalité :

$$uv^{mk}w = a_1 a_2 \ldots a_p$$

On considère l'ensemble I des index i pour lesquels il existe un entier r et une factorisation $v = v_1 v_2$ de v (qui dépendent en général de i) tels que

$$uv^r v_1 = a_1 a_2 \ldots a_i$$

L'ensemble I a une cardinalité Card $(I) > \frac{|v^{mk}|}{m} = \frac{mk\,|v|}{m} = k\,|v|$. A deux éléments différents de I correspondent généralement des factorisations différentes de v. Mais à au moins s éléments, avec $s \geqslant \frac{\text{Card}\,(I)}{|v|} > k$, correspond la même factorisation $v = v_1 v_2$. Soit I' cet ensemble,

$$I' = \{ i_1, i_2 \ldots, i_s \} \qquad \text{avec} \quad i_1 < i_2 \ldots < i_s.$$

On a alors

$$a_{i_j} \ldots a_{i_{j+1}} = v_2 v^{r_j} v_1 = (v_2 v_1)^{r_j + 1} \qquad (j = 1, \ldots s - 1).$$

et on vérifie sans peine que la condition que α soit un monomorphisme entraîne que $r_j = r$ ne dépend pas de j. On peut ainsi écrire :

$$a_{i_1} \ldots a_{i_s} = (a_{i_1} \ldots a_{i_2})^{s-1}$$

et donc

$$\alpha f = a_1 \ldots a_{i_1 - 1}(a_{i_1} \ldots a_{i_2})^{s-1} a_{i_s + 1} \ldots a_p.$$

Il vient

$$f = x_1 \ldots x_{i_1 - 1}(x_{i_1} \ldots x_{i_2})^{s-1} x_{i_s + 1} \ldots x_p$$

avec $s - 1 \geqslant k$. C'est-à-dire $f \notin L_k(X)$.

III. PHÉNOMÈNES D'ITÉRATION ET LANGAGES BORNÉS

Dans ce paragraphe on établit des liaisons entre les phénomènes d'itération dans un langage rationnel, la condition (B) et la propriété d'un langage d'être borné. D'abord nous donnons la définition suivante.

DÉFINITION : Étant donné un langage L, un mot $v \in X^*$ est un *facteur itérant* de L s'il existe deux mots u, $w \in X^*$ tels que $L \cap uv^*w$ est infini.

Dans un langage rationnel L, si v est un facteur itérant de L, il y a une infinité, de puissances de v qui sont aussi des facteurs itérants de L. On considère alors l'ensemble des facteurs itérants de L qui sont aussi des mots primitifs (un mot d'un monoïde libre est *primitif* s'il n'est pas une puissance propre d'un autre mot du monoïde libre) : d'après notre définition, tout facteur itérant de L est puissance d'un facteur itérant primitif de L. La cardinalité de cet ensemble joue un rôle fondamental dans les problèmes que nous considérons.

THÉORÈME 2 : *Soit L un langage rationnel tel que* $L \cap L_k$ *est fini pour tout entier k. Alors le nombre des facteurs itérants primitifs de L est borné par l'index de la congruence syntactique* σ_L *induite par L.*

Preuve : La preuve est par contradiction. Soit s l'index de la congruence syntactique σ_L enduite par L, et supposons que L ait un nombre de facteurs itérants primitifs plus grand que s; alors deux parmi eux, v et h, sont dans la même classe de σ_L. Si v est un facteur itérant, alors il existe $u, w \in X^*$ tels que $L \cap uv^*w$ soit infini. $v\sigma_L h$ entraîne que dans chaque mot de L, qui contient v comme facteur, on peut substituer h à v et on obtient encore un mot de L.

On a alors que $u \{ v, h \}^n w \subset L$ pour n arbitrairement grand. Puisque v et h sont des mots primitifs, on sait qu'ils engendrent un sous-monoïde libre de X^*, c'est-à-dire que l'homomorphisme α de $\{ x, y \}^*$ dans X^* défini par $\alpha x = v$ et $\alpha y = h$ est un monomorphisme. En considérant alors un mot f sur $\{ x, y \}$ arbitrairement long sans cube $f \in L_3(\{ x, y \})$, il vient :

– $\alpha f \in L_{3m}(X)$ avec $m = \sup \{ |v|, |h| \}$
– $u\alpha f w \in L_p$ avec $p = \sup \{ 3m, |u|, |w| \}$.

Donc $L \cap L_p$ contient des mots arbitrairement longs, contrairement à l'hypothèse.

Le théorème suivant établit une liaison entre la cardinalité de l'ensemble des facteurs itérants primitifs d'un langage rationnel et la condition d'être borné.

THÉOREME 3 : *Un langage rationnel ayant un nombre fini de facteurs itérants primitifs est borné.*

Preuve : Soit V l'ensemble des facteurs itérants primitifs de L. Par hypothèse on sait que Card (V) est fini. Chaque mot $f \in L$ peut s'écrire sous la forme suivante

$$f = a_1 v_1^{n_1} a_2 v_2^{n_2} \dots a_k v_k^{n_k} a_{k+1}$$

avec $v_i \in V$ et $a_i \notin X^* v^+ X^*$ pour tout $v \in V$ (où v^+ dénote l'ensemble des puissances positives de v). On peut encore supposer que cette décomposition de f comme produit de facteurs itérants (v_i) et facteurs qui ne contiennent pas de facteurs itérants (a_i) soit la plus compacte, i.e. on suppose que pour tout i, $j (1 \leqslant i < j \leqslant k) a_i v_i^{n_i} a_{i+1} \dots a_j v_j^{n_j} \notin v^*$ quel que soit $v \in V$. On vérifie alors que, quel que soit $f \in L$, on a $k \leqslant r$, où r est le nombre d'états de l'automate minimal acceptant L. En effet, si $k > r$, pour le lemme 1, il existe deux index i, j, tels que le mot $a_i v_i^{n_i} a_{i+1} \dots a_j v_j^{n_j}$ est un facteur itérant, i.e. ce mot appartient à v^* pour quelque $v \in V$, contre l'hypothèse. On a alors $k \leqslant r$. Puisque les facteurs a_i, qui ne contiennent pas des facteurs itérants et qui interviennent dans la factorisation de chaque mot de L, sont de nombre fini et aussi les facteurs itérants, on a que L peut être obtenu comme union et produit fini d'ensembles finis et sous-ensemble A de la forme $A \subset v^*$ avec $v \in X^*$, c'est-à-dire L est borné.

La preuve du théorème 1 est une conséquence des théorèmes 2 et 3. En plus on obtient le résultat suivant.

COROLLAIRE : *Un langage rationnel est borné si, et seulement si, l'ensemble de ses facteurs itérants primitifs est fini. De plus, la cardinalité de cet ensemble est bornée par l'index de la congruence syntactique induite par le langage.*

BIBLIOGRAPHIE

1. F. DEJEAN, *Sur un Théorème de Thue*. Journal of Combinatorial Theory (A), 13, 1972, 99-99.
2. S. EILENBERG, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.
3. S. GINSBURG and E. H. SPANIER, *Bounded ALGOL-like Languages*, Trans. American Math. Soc., 113, 1964, 333-368.
4. S. GINSBURG and E. H. SPANIER, *Bounded Regular Sets*, Proc. American Math. Soc., 17, 1966, 1043-1049.
5. G. TH. GUILBAUD et J. GARDELLE, *Cadences*, Mathématiques et Sciences Humaines, 9, 1964, 31-38.
6. A. RESTIVO, *A Characterization of Bounded Regular Sets*, 2nd GI Conference on Automata Theory and Formal Languages, Lecture Notes in Computer Science, 33, 1975, 239-244.
7. A. THUE, *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen*, Skr. Vid. Kristiania I Mat. Naturv. Klasse 1, 1912, 1-67.