RAIRO. INFORMATIQUE THÉORIQUE

L. BOASSON

Un langage algébrique particulier

RAIRO. Informatique théorique, tome 13, nº 3 (1979), p. 203-215

<http://www.numdam.org/item?id=ITA_1979__13_3_203_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 13, n° 3, 1979, p. 203 à 215)

UN LANGAGE ALGÉBRIQUE PARTICULIER (*)

par L. BOASSON (1)
Communiqué par J. BERSTEL

Résumé. — *On sait que la famille des langages algébriques constitue un cône rationnel principal. Elle contient donc un sous-cône maximal, noté Nge, constitué des langages non-générateurs. L'un des principaux problèmes (toujours ouvert) est de savoir si ce dernier est principal ou non. Le principal résultat touchant à cette question affirme que Nge contient strictement la famille Gre des langages de Greibach (= la clôture par substitution de la famille des langages linéaires et à-un-compteur). Pour mieux comprendre la structure de Nge, on peut chercher à savoir ce qui rend cette inclusion stricte, c'est-à-dire à trouver des familles de langages sur lesquelles Nge et Gre coïncident. Il a été conjecturé* [8] *que les langages IRS formaient une telle famille. Nous montrons qu'il n'en est rien, construisant par la même des langages d'un nouveau type qui sont dans Nge et pas dans Gre. De façon surprenante, notre contre-exemple à la conjecture ci-dessus fournit aussi deux réponses négatives à deux problèmes différents, le premier touchant à la notion d'adhérence* [9], *le second à celle d'index rationnel* [11].

Abstract. — *It is well-known that the family of context-free languages is a full principal AFL. It thus contains a largest full sub-AFL, denoted Nge, consisting of the languages which are not full generators. One of the major problems (still unsolved) is to know whether this AFL is principal or not. The main result in this direction states that Nge strictly contains the family Gre of the Greibach languages (= the substitution closure of the linear and one counter languages'family). To get a better understanding of how Nge is built, one possible way is to try to find what makes this inclusion strict; that is to find subfamilies on which Nge and Gre do coincide. It was conjectured* [8] *that the family of IRS languages was such a sub-family. We disprove this conjecture, getting thus a new kind of languages that lie between Nge and Gre. Surprisingly, our counter example gives also negative answers to two different questions, the first about the notion of adherence* [9], *the second dealing with rational indexes* [11].

I. INTRODUCTION

La famille Alg des langages algébriques a fait l'objet de nombreuses études depuis fort longtemps. Sans revenir ici ni sur les diverses raisons de cet état de fait, ni sur le cheminement qu'ont suivis les divers travaux sur ce sujet, notons que, le plus souvent, les problèmes concernant cette famille sont maintenant posés et traités dans le cadre des cônes rationnels ou des full AFL's. Le théorème de Chomsky-Schützenberger [12], par exemple, s'énonce facilement dans ce cadre : il affirme que le cône rationnel Alg est principal [12]. On sait alors que

(*) Reçu décembre 1977, révisée septembre 1978.
(1) Université de Picardie, Amiens et L. A. du C.N.R.S. « Informatique théorique et Programmation ».

Alg contient strictement un cône rationnel maximal Nge formé des langages algébriques qui ne sont pas générateurs de Alg. On sait aussi que ce cône est un full-AFL fermé par substitution [7]. L'une des questions ouvertes les plus importantes concernant ce cône est celle de sa principalité. (Il est conjecturé dans [7] que Nge est un cône non-principal). Or, il se trouve que l'on ne sait presque rien de cette famille de langages si ce n'est qu'elle contient strictement le cône Gre des langages de Greibach (= le plus petit cône fermé par substitution contenant les langages linéaires et à-un-compteur) [3].

Ce dernier cône est beaucoup mieux connu : on sait qu'il est non-principal [7]; on sait aussi que tout langage IRS (= ne contenant aucun langage rationnel infini) qu'il contient est quasi-rationnel [8]. (Rappelons qu'un langage est quasi-rationnel ou non-expansif s'il est dans la clôture par substitution du cône des langages linéaires). La considération simultanée de ce dernier fait et du contre-exemple donné dans [3] à l'égalité de Gre et Nge a conduit alors à la conjecture [8] : Si un langage IRS n'est pas générateur, il est quasi-rationnel.

Notons que cette conjecture semblait renforcée par les résultats plus particuliers concernant les langages parenthétiques [5] ou très simples [6] puisque ceux-ci établissent la véracité de cette conjecture dans ces deux cas précis.

Nous nous proposons ici d'étudier un langage L particulier qui soit un contre-exemple à cette conjecture. Il se trouve que le même langage L joue également le rôle de contre-exemple pour deux questions a priori totalement différentes :

— dans [9], on définit l'adhérence d'un langage L comme l'ensemble des mots infinis dont tout facteur gauche (fini) est facteur gauche d'un mot de L. Il se trouve que le cas particulier considéré ici établit que l'on ne peut pas espérer lire le caractère générateur d'un langage sur son adhérence;

— dans [11], on définit l'index rationnel d'un langage L et l'on propose la :

Conjecture : Si L est non-générateur, son index rationnel est polynomial. A nouveau, le même langage que celui répondant aux deux questions précédentes donne ici une réponse (encore négative !).

Nous présentons cette étude en trois parties : la première donne quelques définitions commentées du langage étudié; la deuxième en établit quatre propriétés utilisées dans la troisième partie pour démontrer les trois résultats annoncés (Prop. 1, 2 et 3).

II. DÉFINITION DU LANGAGE L ÉTUDIÉ

Nous allons donner trois définitions du langage L étudié ici. La première est « synthétique », c'est-à-dire qu'elle décrit les mots de L; la seconde consiste à donner une grammaire algébrique engendrant L, la dernière donnant un automate à pile reconnaissant ce langage. Nous ne donnerons pas de preuves formelles de l'équivalence de ces trois définitions, d'abord parce que ces preuves sont faciles mais ennuyeuses, ensuite et surtout parce que nous n'aurons vraiment besoin que de la troisième définition. Les deux autres sont essentiellement destinées à faire mieux comprendre comment sont construits les mots de L.

Avant de pouvoir définir ce langage, nous avons besoin de quelques définitions spécifiques liées au langage classique E engendré par la grammaire $\langle S \to aSSc + d \rangle$. Ce langage est générateur de la famille des langages algébriques; il est sans facteur itérant. Ces propriétés ainsi que quelques autres faciles à établir sont prouvées dans [2]. C'est un langage simple reconnu par l'automate à pile $\mathscr{A}_E$ dont les règles sont :

$$\delta(a, q, S) = (q, cSS), \qquad \delta(c, q, c) = (q, 1), \qquad \delta(d, q, S) = (q, 1).$$

Si f est un facteur gauche d'un mot de E, on définit *le poids* de f, $p(f)$ comme étant la hauteur de la pile de $\mathscr{A}_E$ après lecture de f, soit encore $p(f) = |m|$ avec $\delta(f, q, S) = (q, m)$. De façon équivalente, on peut définir $p(a) = 2$, $\bar{p}(d) = \bar{p}(c) = 1$ et, étendant $\bar{p}$ en un morphisme de X^* dans N on vérifie facilement que $p(f) = \bar{p}(f) + 1$. Encore convient-il de bien noter que cette égalité n'a de sens que si f est bien un facteur gauche de E.

Exemples : $f = aadaddc$:

$$\delta(f, q, S) = (q, cSc), \qquad p(f) = 3, \quad \bar{p}(f) = 2;$$

$f = aada$:

$$\delta(f, q, S) = (q, cSccSS), \qquad p(f) = 6, \quad \bar{p}(f) = 5;$$

$f \in E$:

$$\delta(f, q, S) = (q, 1), \qquad p(f) = 0, \quad \bar{p}(f) = -1.$$

Considérant maintenant un mot f quelconque, nous définissons $e(f)$ comme le facteur gauche de f le plus long qui soit aussi facteur gauche d'un mot de E. Soit encore $f = e(f) f_2$ avec $f_2 = 1$ ou $f_2 = x f_2'$ et $e(f)x$ n'est pas facteur gauche d'un mot de E.

Remarquons que, puisque le mot vide 1 est facteur gauche de E, $e(f)$ est défini pour tout f. On peut alors définir

$$(a) \qquad L = \{ f \mid f = e(f) h \text{ et } p(e(f)) \leqq |h| \leqq 2p(e(f)) \}.$$

Comme si $f \in E$ $e(f) = f$ et $p(f) = 0$, on note que L contient E. Ainsi les mots de L sont soit des mots de E, soit des mots dont le facteur gauche maximal qui soit facteur gauche (propre) de E est suivi d'un mot de longueur comprise entre le poids de ce facteur gauche et le double de ce poids. De cette description, on déduit facilement

(b) Une grammaire engendrant L :

Celle-ci sera constituée de trois non-terminaux : S engendrant E, T engendrant $(L-E)=L_1$ et l'axiome A :

$$A \to S + T,$$

$$S \to a\,S\,S\,c + d \qquad \text{(engendrant } E\text{)},$$

$$T \to \underset{\cdot}{a}\,T\,\alpha_1 + a\,S\,T\,\alpha_2, \qquad \forall \alpha_1,\, \alpha_2 \in X^*, \quad 2 \leqq |\alpha_1| \leqq 4, \quad 1 \leqq |\alpha_2| \leqq 2,$$

$$T \to a\underset{\cdot}{c}\,\alpha_1 + a\,S\,\underset{\cdot}{c}\,\alpha_2 + a\,S\,S\,\underset{\cdot}{d}\,\alpha_3 + a\,S\,S\,\underset{\cdot}{a}\,\alpha_3,$$

$$\forall \alpha_1,\, \alpha_2,\, \alpha_3 \in X^*, \qquad 2 \leqq |\alpha_1| \leqq 5, \qquad 1 \leqq |\alpha_2| \leqq 3, \qquad 0 \leqq |\alpha_3| \leqq 1.$$

On vérifie facilement que la variation de hauteur de la pile du figurant à gauche de T est bien compensée par un mot à droite de T de longueur comprise entre cette variation et son double. Les règles terminales sont telles que figure explicitement la première lettre du mot f engendré qui ne soit dans $e(f)$. (Cette lettre est indiquée par un point dans les règles ci-dessus.)

(c) Un automate à pile reconnaissant L :

$$\left.\begin{aligned} \delta(a,\, q,\, S) &= (q,\, c\,S\,S) \\ \delta(c,\, q,\, c) &= (q,\, 1) \\ \delta(d,\, q,\, S) &= (q,\, 1) \end{aligned}\right\} \text{ règles de } \mathscr{A}_E$$

$$\delta(x,\, q,\, y) = (t,\, 1) + (\bar{t},\, y), \qquad \forall (x,\, y) \in \{(a,\, c);\, (d,\, c);\, (c,\, S)\}.$$

Cette règle détecte la première lettre (si elle existe !) du mot f analysé qui ne soit pas dans $e(f)$. L'état t correspond alors aux lettres effaçant un symbole de pile, l'état $\bar{t}$ à celles effaçant « 1/2 symbole » :

$$\left.\begin{aligned} \delta(x,\, t,\, y) &= (t,\, 1) + (\bar{t},\, y) \\ \delta(x,\, \bar{t},\, y) &= (t,\, 1) \end{aligned}\right\} \forall (x,\, y) \in \{a,\, c,\, d\} \times \{c,\, S\}$$

On vérifiera facilement que

$$\delta(f,\, q,\, S) = (q,\, 1) \quad \Leftrightarrow \quad f \in E,$$

$$\delta(f,\, q,\, S) = (t,\, 1) \quad \Leftrightarrow \quad f \in (L-E) = L_1.$$

Ainsi on trouve que $L = \{ f \mid \delta(f,\, q,\, s) = (r,\, 1) \text{ avec } r = q \text{ ou } r = t \}$. Notons que l'on peut alors définir $e(f)$ comme le plus long facteur gauche de f, f_1, tel que $\delta(f_1,\, q,\, S) = (q,\, m)$. Ainsi, les preuves utilisant la définition donnée précédemment de $e(f)$ ne font-elles en fait appel qu'à cette dernière définition de L.

III. PROPRIÉTÉS DU LANGAGE L

Nous allons établir quatre propriétés du langage que nous venons de définir. Les trois premières permettent de répondre à une conjecture de S. Greibach [8], les deux dernières à un problème naturel soulevé par la notion d'adhérence donnée dans [9].

A. Le langage L est sans facteur itérant

Rappelons d'abord que si $\alpha\beta$ et $\alpha u \beta$ sont deux mots de E, alors $u=1$. Supposons alors qu'il existe trois mots non vides α, u et β tels que pour tout entier n, $\alpha u^n \beta$ soit un mot de L. On en déduit que, puisque $u \neq 1$, $\exists\, n_0$, $\forall\, n \in N$, $n \neq n_0$, $\alpha u^n \beta \in L_1 = (L - E)$.

Deux cas sont alors possibles :

(a) $\exists\, n_1$ tel que $e(\alpha u^{n_1}) \neq \alpha u^{n_1}$.

On en déduit que, pour tout m, $e(\alpha u^{n_1+m}) = e(\alpha u^{n_1})$ et donc $p(e(\alpha u^{n_1+m})) = p(e(\alpha u^{n_1})) = p_1$. Mais alors, on doit avoir $\forall\, m$, $m|u| + |\beta| \leqq 2\, p_1$ et $|u| = 0$. Ainsi, ce premier cas étant exclu, c'est que :

(b) $\forall\, n$, $e(\alpha u^n) = \alpha u^n$. A chaque valeur de n, on peut alors associer une factorisation de β en $\beta_{1,n}\, \beta_{2,n}$ telle que $e(\alpha u^n \beta) = \alpha u^n \beta_{1,n}$. Il existe donc au moins une factorisation de β en $\beta_1\, \beta_2$ telle que $P(\beta_1) = \{ n \mid e(\alpha u^n \beta) = \alpha u^n \beta_1 \}$ soit infini. Comme alors la longueur de β_2 est fixe, on sait que $p(\alpha u^n \beta_1)$ est borné par $|\beta_2|$. Il existe alors un poids p_1 tel que

$$P(\beta_1) = \{ n \mid n \in P(\beta_1) \text{ et } p(\alpha u^n \beta_1) = p_1 \} \quad \text{soit infini.}$$

Le nombre de mots de pile (sur C et S) de longueur p_1 étant fini; il existe alors au moins un tel mot m tel que

$$P_1(m) = \{ n \mid n \in P(\beta_1) \text{ et } \delta(\alpha u^n \beta_1, q, S) = (q, m) \} \quad \text{soit infini.}$$

Comme il existe alors un mot γ satisfaisant $\delta(\gamma, q, m) = (q, 1)$ on en déduit que : $\forall\, n \in P_1(m)$, $\alpha u^n \gamma \in E$, et, en vertu de notre remarque initiale $u = 1$.

B. Le langage L est expansif

Nous commencerons par quelques remarques préliminaires.

Remarque 1 : $\forall\, h \in E\; p(ah) = 2$ et $p((ah)^n) = n + 1$.

Si l'on définit maintenant les mots $h_0 = d$, et $h_{i+1} = (ah_i)^N h_i c^N$ (où N est un paramètre qui sera précisé ultérieurement), on vérifie facilement par récurrence sur i:

Remarque 2 : $\forall\, i \geqq 0$, $h_i \in E$.

Remarque 3 : $p_i = \text{Max}\{p(g) \mid g \text{ est facteur gauche de } h_i\} = iN+1$.

La preuve se fait par récurrence sur i. Pour $i=0$, on a

$$p_0 = \text{Max}\{p(1), p(d)\} = 1.$$

Supposons alors que $p_i = iN+1$, nous pouvons évaluer p_{i+1} par examen des divers facteurs gauches de h_{i+1} :

$-g=(ah_i)^n ah'_i$ avec $h_i = h'_i h''_i$:

$$p(g) = n+1+p(h'_i) < N+1+p_i \qquad (n < N);$$

$-g=(ah_i)^n$:

$$p(g) = n+2 \qquad (n \leqq N);$$

$-g=(ah_i)^N h'_i$ avec $h'_i h''_i = h_i$:

$$p(g) = N+p(h'_i) \leqq N+p_i, \qquad p(h'_i) = p_i, \ p(g) = N+p_i;$$

$-g=(ah_i)^N h_i c^n$:

$$p(g) = N-n \qquad (n \leqq N).$$

Il en résulte immédiatement que $p_{i+1} = N+p_i = (i+1)N+1$.

Remarque 4 : $\forall i \geqq 0$, $|h_i| = 3(N+1)^i - 2$.

A nouveau, procédant par récurrence sur i, on vérifie que $|h_0| = 3-2=1$. Puis, comme $|h_i| = 2N+(N+1)|h_{i-1}|$, on obtient $|h_i| = 3(N+1)^i - 2$.

On fixe maintenant N comme étant l'entier associé au langage L par le lemme d'Ogden [10]. Nous allons alors montrer que L est expansif. Dans la première occurence de h_i dans le mot h_{i+1}, on marque les N dernières occurences de la lettre c. Soit

$$h_{i+1} = a(ah_{i-1})^N h_{i-1} \underline{c^N} (ah_i)^{N-1} h_i c^N,$$

où les occurences marquées sont soulignées. Nous allons montrer que la paire itérante fournie par le lemme d'Ogden doit avoir toutes ses itérées dans E. Celle-ci sera notée $(\alpha, u, \beta, v, \gamma)$ et l'on sait que soit α, u et β, soit β, v, γ contiennent des occurrences marquées. Il est bien clair qu'en vertu de la propriété A, on ne peut pas avoir u et v simultanément contenus dans le facteur marqué c^N. On peut alors supposer :

1. u est contenu dans le facteur distingué. Soit $\alpha = a(ah_{i-1})^N h_{i-1} c^{n_1}$ et $u = c^{n_2}$. Mais alors, comme $e(ah_i c) = ah_i$, on sait que $e(\alpha u^2 \beta v^2 \gamma) = e(ah_i c) = ah_i$. Il en résulte que l'on doit avoir $p(ah_i) \geqq |\beta v^2 \gamma| \geqq |\beta v \gamma| > |h_i|$ et donc $2 > 3(N+1)^i - 2$ ce qui est impossible. On est donc sûr que :

2. *C'est u qui est contenu dans le facteur marqué* (seuls les cas possibles sont en italique). C'est-à-dire que l'on a

$$\gamma = c^{n_2} (ah_i)^{N-1}\, h_i\, c^N\, (n_2 \neq 0),$$

$$v = c^n\, (n \neq 0),$$

$$\beta = \beta_1\, c^{n_1}\, (n_1 \neq 0) \qquad \text{avec} \quad n_1 + n + n_2 = N.$$

Nous allons alors éliminer successivement divers cas possible. Supposons d'abord :

2 *a*. Il existe un entier n_0 tel que $e(\alpha u^{n_0}) \neq \alpha u^{n_0}$. Alors pour tout entier $m \geqq 0$, $e(\alpha u^{n_0+m}) = e(\alpha u^{n_0})$; il en résulte que la longueur de $v^{n_0+m}\gamma$ doit être bornée par $2\, p(e(\alpha u^{n_0}))$, ce qui implique $v = 1$ et donc $n = 0$ On en déduit donc que :

2 *b*. *Pour tout n*, $e(\alpha u^n) = \alpha u^n$.

De ce dernier résultat, on peut déduire comment se passe le calcul de l'automate $\mathscr{A}_E$ lisant αu^n :

$$\delta(\alpha, q, S) = (q, m).$$

On peut alors factoriser m en $m_1\, m_2$ avec

$$\delta(u, q, m_2) = (q, \nu).$$

(On indique ainsi que la lecture de u n'utilise aucun symbole de pile pris dans m_1.)

Ainsi $\delta(\alpha u, q, S) = (q, m_1\, \nu)$.

On sait aussi que $\delta(\alpha u^2, q, S)$ est défini dans $\mathscr{A}_E$ et donc que $\delta(u, q, m_1\, \nu)$ est défini. Or m_2 est le (seul) mot de pile le plus court tel que $\delta(u, q, m_2)$ soit défini dans $\mathscr{A}_E$. Il en résulte que $m_1\, \nu$ admet m_2 comme facteur droit. Deux cas sont alors possibles :

1° $m_1 = m_1'\, m_1''$ et $m_2 = m_1''\, \nu$. On sait alors que

$$\delta(u, q, m_1'\, m_1''\, \nu) = \delta(u, q, m_1'\, m_2) = (q, m_1'\, \nu).$$

Mais alors $p(\alpha u) < p(\alpha u^2)$ et, de façon générale

$$p(\alpha u^n) < p(\alpha u^{n+1})$$

ce qui contredit le fait que

$$e(\alpha u^n) = \alpha u^n, \qquad \forall n \geqq 1.$$

2° C'est donc que $\nu = \nu'\, m_2$ et

$$\delta(u, q, \nu) = \delta(u, q, \nu'\, m_2) = (q, \nu'\, \nu) = (q, \nu'^2\, m_2).$$

Ainsi on sait que

$$\delta(\alpha, q, S) = (q, m_1, m_2),$$

$$\delta(u, q, m_2) = (q, \nu'\, m_2),$$

et donc

$$\delta(\alpha u^n, q, S) = (q, m_1 \nu'^n m_2), \qquad \forall n \geqq 1.$$

Supposons alors :

2 b 1. Il existe n_0 tel que $e(\alpha u^{n_0} \beta) \neq \alpha u^{n_0} \beta$.

Rappelons que l'on sait évidemment que $e(\alpha u \beta) = \alpha u \beta$.

Ainsi, l'automate $\mathscr{A}_E$ ne se bloque pas en calculant

$$(1) \qquad \delta(\beta, q, m_1 \nu' m_2)$$

et il se bloque en calculant

$$(2) \qquad \delta(\beta, q, m_1 \nu'^{n_0} m_2).$$

De (1) et (2) on déduit que $\beta = \beta_1 \beta'$ avec $\delta(\beta', q, m_2) = (q, 1)$. (Sinon, β ne provoquera jamais de lecture de symboles de $m_1 \nu'$ et (2) ne conduira jamais à un blocage de $\mathscr{A}_E$.) On a alors

$$(1') \qquad \delta(\beta', q, m_1 \nu') \quad \text{est défini,}$$

$$(2') \qquad \delta(\beta', q, m_1 \nu'^{n_0}) \quad \text{se bloque.}$$

A nouveau, pour les mêmes raisons $\beta' = \beta_2 \beta''$ avec $\delta(\beta_2, q, \nu') = (q, 1)$ et

$$(1'') \qquad \delta(\beta'', q, m_1) \quad \text{est défini,}$$

$$(2'') \qquad \delta(\beta'', q, m_1 \nu'^{n_0 - 1}) \quad \text{ne l'est pas.}$$

De (1″) on déduit que $m_1 = m_2 m_3$, où m_3 est le mot le plus court tel que $\delta(\beta'', q, m_3)$ soit défini. Si (2″) ne vaut pas pour $n_0 = 2$ c'est que $m_1 \nu'$ admet aussi m_3 comme facteur droit et comme ci-dessus, on doit avoir m_3 facteur droit de ν' ce qui contredit (2″). Il en résulte que (2) vaut pour $n_0 = 2$.

Mais alors $p(e(\alpha u^2 \beta)) \leqq 2 p_i$ et $|\gamma| < |h_i|$ impliquent que $|h_i| < 4 p_i$ ce qui est impossible. On sait donc :

2 b 2. *Pour tout* n, $e(\alpha u^n \beta) = \alpha u^n \beta$.

Reprenant les calculs de l'automate $\mathscr{A}_E$ lisant $\alpha u^n \beta \Lambda$ on sait que

$$\left.\begin{array}{l} \delta(\alpha, q, S) = (q, m_1 m_2) \\ \delta(u, q, m_2) = (q, \nu' m_2) \\ \delta(\beta, q, m_3 m_2) = (q, 1). \end{array}\right\} \quad \text{avec,} \quad \nu' = \nu'' m_3.$$

Soit alors

$$\delta(\alpha, u^n \beta, q, S) = (q, m_1 \nu'^{n-1} \nu''), \qquad \forall n \geqq 1.$$

Deux cas sont alors possibles :

2 b 2 a. Il existe un entier n_0 tel que $e(\alpha u^{n_0}\beta v^{n_0}) \neq \alpha u^{n_0}\beta v^{n_0}$.

De la même façon que ci-dessus, on vérifie facilement que si tel est le cas, 2 b 2 a vaut pour $n_0=2$ et ce cas est écarté comme le cas 2 b 1. On en déduit :

2 b 2 b. *Pour tout entier* n, $e(\alpha\, u^n\, \beta v^n)=\alpha\, u^n\, \beta v^n$.

Mais alors si $p_n=p(\alpha\, u^n\, \beta v^n)$, on doit avoir $p_n > 1/2\,|\gamma|$ et donc p_n est borné. Donc, $p(\alpha\, u^n\, \beta v^n)$ doit en fait être constant et $\alpha\, u^n\, \beta v^n \gamma \in E$.

Il en résulte immédiatement que le facteur h_{i-1} souligné ci-dessous n'admet que des paires itérantes dans E :

$$h_{i+1}=a(a\,h_{i-1})^N\,\underline{h_{i-1}}\,c^N(a\,h_i)^{N-1}\,h_i\,c^N.$$

Or la preuve classique que le langage E est expansif procède ainsi : si E était quasi-rationnel, h_i étant un mot de E, ce langage serait quasi-rationnel d'ordre au moins i. Reproduisant ici cette preuve, sachant que h_{i-1} est dans E, on trouve que si L était quasi-rationnel, le fait que h_{i+1} soit un mot de L implique que L soit d'ordre au moins $i-1$.

C. Le langage L n'est pas générateur

Pour prouver cette propriété, nous utilisons le :

THÉORÈME DE BEAUQUIER [1] : *Un langage algébrique A est générateur si et seulement si il existe* 6 *mots* $\sharp_1, \alpha, \beta, \gamma, \delta$ *et* $\sharp_2$ *et un langage rationnel K tels que*

$$A \cap K = E_0(\langle S_0 \to \sharp_1 S\, \sharp_1;\ S \to \alpha S \beta S \gamma + \delta \rangle).$$

(Cette notation signifiant que E_0 est le langage engendré par la grammaire donnée entre chevrons d'axiome S_0.)

Supposons donc que L soit générateur et désignons par $\sharp_1$, α, β, γ, δ $\sharp_2$ les 6 mots donnés par le théorème de Beauquier.

1. On doit avoir $e(\sharp_1\, \alpha^n(\alpha\delta\beta)^p\,\delta\gamma^p)=\sharp_1\quad \alpha^n(\alpha\delta\beta)^p\,\delta\gamma^p=h_{n,\,p}$ (pour tout n, $p \geqq 0$).

En effet, sinon la longueur des mots f tels que $h_{n,\,p} f \in L$ serait bornée et il en irait de même des mots f tels que $h_{n,\,p} f \in L \cap K$ ce qui n'est pas (si l'on suppose L générateur).

2. Le poids $p(\sharp_1\, \alpha^n(\alpha\delta\beta)^p\,\delta\gamma^p)$ doit croître strictement avec n. En effet, sinon ce poids étant borné, on retrouve la même contradiction que celle du cas b de la propriété A.

Choisissant alors p assez grand pour qu'il existe n tel que $\sharp_1\, \alpha^n(\alpha\delta\beta)^p\,\delta\gamma^p(\gamma)^*\,\sharp_2 \subseteq K$, il vient :

$$g_{i,\,n}=\sharp_1\,\alpha^n(\alpha\delta\beta)^p\,\delta\gamma^p\gamma^{ir}\,\sharp_2 \in K\ \forall_{i,\,n}.$$

Or, certainement il existe un choix de n et i tel que le mot obtenu soit dans L. En effet, si le mot $g_{i,n} \notin E$, c'est qu'il existe une factorisation de γ en $\gamma_1 \gamma_2$ telle que $e(g_{i,n}) = h_{n,p} \gamma_1$ (quand on lit $\alpha\delta\beta\delta\gamma$ dans $\mathscr{A}_E$, on sait que γ efface la pile mémorisée après lecture de $\alpha\delta\beta\delta$ et que si l'on lit $\alpha\delta\gamma$ l'automate $\mathscr{A}_E$ doit se bloquer).

Posant alors $p(e(g_{i,n})) = p_n$, on cherche i tel que

$$p_n \leqq |\gamma_2| + ir|\gamma| + |\#_2| \leqq 2p_n$$

soit

$$p_n - s \leqq ir|\gamma| \leqq 2p_n - s.$$

Comme p_n croît strictement avec n, on trouvera au moins un tel i dès que

$$1 + \frac{p_n - s}{r|\gamma|} < \frac{2p_n - s}{r|\gamma|} \qquad \text{soit} \quad \frac{p_n}{r|\gamma|} > 1.$$

Ainsi ce mot $g_{i,n} \in L \cap K$ et $g_{i,n} \notin E_0$.

D. Les langages L et E ont mêmes adhérences

Cette dernière propriété résulte immédiatement de la définition suivante [9] : $\mathrm{Adh}(L) = \{u \in X^\omega \mid \text{Tout facteur gauche de } u \text{ est facteur gauche de } L\}$ (Rappelons que X^ω désigne l'ensemble des mots infinis sur l'alphabet X).

On peut donc écrire [9] : $u \in \mathrm{Adh}(L) \Leftrightarrow u_1 X^* \cap L$ est infini pour tout facteur gauche u_1 de u. Notons $FG(L)$ l'ensemble $\{u \mid \exists v, uv \in L\}$, si $u_1 \in FG(L) - FG(E)$, on vérifie que $u_1 X^* \cap L$ est fini. On en déduit que $\mathrm{Adh}(L) \subseteqq \mathrm{Adh}(E)$ et comme $L \supseteqq E$, $\mathrm{Adh}(L) \supseteqq \mathrm{Adh}(E)$.

IV. COROLLAIRES ET CONCLUSIONS

En rapprochant les propriétés A, B et C de L, on a :

PROPOSITION 1 : *Il existe des langages expansifs sans facteurs itérants qui ne sont pas générateurs.*

En rapprochant les propriétés C et D de L, on a :

PROPOSITION 2 : *Il existe des langages non générateurs ayant même adhérence qu'un générateur.*

Enfin, on peut énoncer la :

PROPOSITION 3 : *Il existe des langages algébriques non générateurs d'index rationnel au moins exponentiel.*

La définition précise de l'index rationnel $g_L(n)$ d'un langage L est donnée dans [11]. Rappelons brièvement que, Rat_n désignant la famille des langages rationnels reconnus par un automate fini (déterministe ou non) ayant n états, on pose

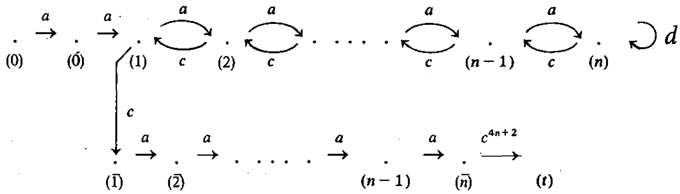
$$g_L(n) = \mathrm{Max}\{\,\mathrm{Min}\{\,|f|\;|\,f \in L \cap K\}\,|\,K \in \mathrm{Rat}_n\}.$$

Nous allons ici simplement donner une famille d'automates assurant que pour chaque n, $g_L(n)$ est supérieur à une fonction exponentielle de n, la propriété C donnant alors l'énoncé de la proposition 3 qui contredit une conjecture de [11].

On trouvera dans [11] une famille d'automates $\mathscr{A}_n$ établissant que E est d'index rationnel au moins exponentiel. Schématiquement $\mathscr{A}_n$ est décrit ainsi :

Si $K_n = \{f \,|\, (1)f = (1)\}$, on vérifie immédiatement :
$E \cap K_1 = \{d\}$.
– $E \cap K_n = \{h_n\}$ avec $h_n = ah_{n-1}h_{n-1}c$, $\forall n \geqq 2$.
– $|h_n| = 2^n + 2^{n-1} - 2$.
– Tout facteur gauche de h_n est de poids au plus $2n+1$.
– Si $R_n = \{f \,|\, (1)f \text{ est défini}\}$, on a $FG(E) \cap R_n = FG(h_n)$. En particulier, le seul facteur gauche de E allant de l'état (1) à (1) est le mot h_n lui-même.

On construit alors la famille d'automates $\mathscr{B}_n$ décrits ci-dessous :



$\mathscr{B}_n$ admet donc $6n+2$ états. Par ailleurs si K_n désigne maintenant $\{f \,|\, (0)f = (t)\}$, on remarque immédiatement $f \in K_n \Rightarrow f$ se termine par $ca^n c^{4n+1}$ et commence par a^2.

Ainsi si $f \in K_n \cap L$, $f \in (L - E)$ soit encore $e(f) \neq f$. Mieux, on peut même préciser que $f = aagca^n c^{4n+1}$ et $e(f) = e(aagca^n)$, puisque ac n'est jamais facteur d'un mot de E. En outre $(1)g = (1)$.

Supposons alors :

– $e(aagca^n) \neq aagca^n$. Or, si $aagc$ est facteur gauche de E, il en va de même de $aagca^n$. On a donc $e(aagca^n) = aag_1$, où g_1 est un facteur gauche du mot g. On sait alors que $(1)\,g_1$ est défini aussi bien dans $\mathscr{A}_n$ que dans $\mathscr{B}_n$ et, de nos remarques initiales, il résulte que g_1 est facteur gauche de h_n. Il est donc de poids au plus $2n+1$. Ainsi $p(aag_1) \leqq 2n+5$ et comme $|g_2 ca^n c^{4n+2}| < 5n+3$, $aagca^n c^{4n+1} \notin L$.

On doit donc nécessairement avoir :

— $e(aagca^n) = aagca^n$. Ainsi, en particulier aag est-il facteur gauche de E, ce qui implique, puisque $(1)g=(1)$ que $g=h_n h_n$. Il est alors facile de voir que $aah_n h_n ca^n$ est de poids $2n+1$ et que donc $h'_n = aah_n h_n ca^n c^{4n+2} \in L$. On sait donc que $L \cap K_n = \{ h'_n \}$ avec

$$|h'_n| = |h_{n+1}| + 5n+3 = 2^{n+1} + 2^n + 5n+1$$

et donc

$$g_L(6n+2) < 2^{n+1}.$$

Remarque : En vertu des résultats de [11], cette dernière propriété implique que L soit expansif. L'intérêt de la preuve directe donnée auparavant est de donner beaucoup plus de précisions sur la structure des mots de L alors qu'il est probable qu'une étude approfondie de ce langage sera fort instructive. C'est pourquoi, il semble utile de maintenir cette première preuve un peu technique.

Conclusion : On peut remarquer que l'on peut associer au langage E_0 engendré par la grammaire $\langle S \to aSbSc + d \rangle$ un langage L_0 de la même façon que L est associé à E. On trouvera un nouveau langage satisfaisant les propositions 1, 2 et 3. En outre, Adh (L_0) sera cette fois identique à Adh (E_0), et si l'on rapproche ce fait du théorème de Beauquier cité ci-dessus, on constate qu'il est vraiment impossible de caractériser un générateur par son adhérence.

La proposition 1 affirme que L est un contre-exemple à la conjecture de S. Greibach énoncée dans l'introduction. On s'assurera facilement que L est non-ambigu mais non déterministe. Le problème reste ouvert de savoir si l'on peut trouver un langage déterministe IRS non-générateur. Rappelons que l'on sait que l'on ne peut trouver un tel langage qui soit très simple [6] ou parenthétique [5]. A cet égard, il faut bien noter que le contre-exemple simple de [4] n'est pas correct puisqu'il s'agit du langage B engendré par $\langle S \to aSSa + aSSb + b \rangle$ que l'on montre être générateur en posant

$$\alpha = aab, \qquad \beta = aa, \qquad \gamma = baa, \qquad \delta = b,$$

et en considérant $B \cap K$, où K est le rationnel quasi-local « habituel » [1].

Notons enfin, pour terminer, que l'on peut associer de la même façon un langage L' au langage symétrique S_2 et que L' sera un langage IRS non générateur du cône des langages linéaires. En revanche, appliquée à un langage comme $D_1'^*$, la construction proposée ne donne rien.

L'auteur tient à remercier J. Messerschmidt dont la lecture attentive et les nombreux commentaires ont permis une meilleure présentation des résultats.

BIBLIOGRAPHIE

1. J. Beauquier, *Générateurs algébriques non-ambigus*, in *Automata, Languages and Programming*, Third International Colloquium, S. Michaelson and R. Milner, éd., Edinburgh University Press, 1976, p. 66-73.
2. L. Boasson, *Cônes rationnels et familles agréables de langages*, Thèse de 3e cycle, Université Paris-VII, 1971.
3. L. Boasson, *The Inclusion of the Substitution Closure of Linear and One Counter Languages in the Largest Full-Sub AFL of the CFL is Proper*, Information Processing Letters, vol. 2, 1973, p. 135-140.
4. L. Boasson, *Un langage algébrique non-générateur*, in 3rd G. I. Conference. Lecture Notes in Comp. Science, n° 48, Springer-Verlag, 1977, p. 145-148.
5. L. Boasson et M. Nivat, *Parenthesis Generators*, 17th Annual Symposium on F.O.C.S. (ex. SWAT), 1976, p. 253-257.
6. C. Frougny, *Langages très simples générateurs*, R.A.I.R.O., Informatique théorique, vol. 13, 1979, p. 69-86.
7. S. Greibach, *Chains of Full. AFL's*, Math. Systems Theory, vol. 4, 1970, p. 231-242.
8. S. Greibach, *One Counter Languages and the IRS Condition*, J. Comp. System. Science, vol. 10, 1975, p. 237-247.
9. M. Nivat, *Sur les ensembles de mots infinis engendrés par une grammaire algébrique*, R.A.I.R.O., Informatique théorique, vol. 12, 1978, p. 259-278.
10. W. Ogden, *A Helpful Result for Proving Inherent Ambiguity*, Math. Systems Theory, vol. 2, 1967, p. 191-194.
11. L. Boasson, B. Courcelle et M. Nivat, *A New Complexity Measure for Languages*, in *A Conference on Theoretical Computer Science*, Waterloo, 1977, p. 130-138.
12. A. Salomaa, *Formal Languages*, Academic press, 1973.