PUBLICATIONS DU DÉPARTEMENT DE MATHÉMATIQUES DE LYON

MAURICE BLAB

Quelques langages algébriques

Publications du Département de Mathématiques de Lyon, 1984, fascicule 6B
« Théorie des langages et complexité des algorithmes », , p. 1-11

<http://www.numdam.org/item?id=PDML_1984___6B_A1_0>

QUELQUES LANGAGES ALGEBRIQUES

par Maurice BLAB

RESUME. L'objet de ce texte est de présenter divers langages algébriques, en expliquant à quelles occasions ils ont été découverts ou construits. L'intérêt d'une telle présentation est essentiellement de montrer à quel point nous connaissons peu d'exemples de langages algébriques. Ce fait explique que, souvent, une conjecture est établie fausse en fabriquant un nouveau type de langage.

La première partie de ce texte est formée de brefs préliminaires concernant les transductions et les langages. La deuxième partie présente les langages algébriques classiques. Les troisième et quatrième parties sont dévolues aux langages moins classiques soit assez hauts dans la hiérarchie, soit au contraire, très bas.

I. PRELIMINAIRES.

L'ensemble des notions et résultats rappelés ici figurent dans l'ouvrage de J. Berstel [2] . Une transduction du monoïde libre X^* dans le monoïde libre Y^* est une application de X^* dans les parties de Y^*. Une transduction est dite rationnelle si elle peut s'écrire à l'aide de deux morphismes et d'un langage rationnel K :

$$f \in X^* \quad \tau(f) = \psi(\varphi^{-1}(f) \cap K).$$

Exemple : l'opération qui à un mot f associe l'ensemble de ses facteurs gauches est une transduction rationnelle. Au contraire, l'opération qui à un mot associe son image miroir est une transduction qui n'est pas rationnelle.

On sait que la composée de deux transductions rationnelles est une transduction rationnelle. Un cône rationnel est une famille de langages fermée par transduction rationnelle. Etant donnée une famille de langages $\mathscr{L}$ le cône rationnel engendré, noté $\mathscr{T}(\mathscr{L})$, est le plus petit cône rationnel contenant $\mathscr{L}$. Clairement

$\mathscr{T}(\mathscr{L}) = \{M \mid \exists\, L \in \mathscr{L}\ ,\ \tau$ transduction rationnelle telle que $\tau(L) = M\}$.

Un cône rationnel engendré par une famille réduite à un élément L est dit principal de générateurL. Ainsi,

$\mathscr{T}(L) = \{M \mid \exists\, \tau$ transduction rationnelle telle que $\tau(L) = M\}$.

On sait que tout cône rationnel contient la famille Rat des langages rationnels qui est donc le cône minimal. Ce cône est principal : tout langage rationnel en est un générateur.

La notion de transduction rationnelle permet d'introduire un préordre sur les langages défini par

$L \prec L'$ ssi $\exists\, \tau$ transduction rationnelle telle que $L = \tau(L')$

soit aussi $\mathscr{T}(L) \subseteq \mathscr{T}(L')$.

L'équivalence associée, dite équivalence rationnelle, est donnée par

$L \approx L'$ ssi $\mathscr{T}(L) = \mathscr{T}(L')$.

On peut alors énoncer une conjecture :

Conjecture 1 : Quels que soient les langages L et L' tels que $L \prec L'$ ou bien $L \approx L'$, ou bien il existe L" tel que $L \prec L'' \prec L'$ et $L \not\approx L'' \not\approx L'$.

Nous terminerons ces rappels par la notion de substitution. Etant donnés un langage L sur l'alphabet X et des langages M_x, $x \in X$, sur l'alphabet Y, la

substitution $\sigma : x \longrightarrow M_x$ est le morphisme de X^* dans les parties de Y^* définie par $\sigma(x) = M_x$. Si chaque M_x est dans la famille $\mathcal{M}$, σ est une substitution. On sait alors que si $\mathcal{L}$ et $\mathcal{M}$ sont deux cônes rationnels, $\mathcal{L} \circ \mathcal{M} = \{N \mid \exists L \in \mathcal{L}, \sigma \mathcal{M}\text{-substitution telle que } N = \sigma(L)\}$ est un cône rationnel. En outre si $\mathcal{L}$ et $\mathcal{M}$ sont principaux, $\mathcal{L} \circ \mathcal{M}$ est principal. On sait aussi que la cloture par substitution $\mathcal{L}^\sigma$ du cône $\mathcal{L}$ définie par $\mathcal{L}^\sigma = \mathcal{L} \cup \mathcal{L} \circ \mathcal{L} \cup \mathcal{L} \circ \mathcal{L} \circ \mathcal{L} \cup \ldots$ est un cône rationnel tel que soit $\mathcal{L} = \mathcal{L}^\sigma$ soit $\mathcal{L} \subsetneq \mathcal{L}^\sigma$ auquel cas $\mathcal{L}^\sigma$ est non principal.

II. LANGAGES ALGEBRIQUES CLASSIQUES :

Une grammaire algébrique est donnée par un triplet $G = \langle X, V, P \rangle$, où X est l'alphabet terminal, V l'alphabet non terminal ($X \cap V = \emptyset$) et P un ensemble fini de règles de la forme $(v \longrightarrow m)$ avec $v \in V$ et $m \in (X \cup V)^*$. Les dérivations sont séquentielles et le langage engendré est noté $L_G(\sigma)$; c'est $\{f \in X^* \mid \sigma \xrightarrow{*} f\}$.

Les langages de Dyck :

Sur l'alphabet $\hat{Z}_n = Z_n \cup \overline{Z}_n$, avec $Z_n = \{a_1, a_2, \ldots, a_n\}$ et $\overline{Z}_n = \{\overline{a}_1, \overline{a}_2, \ldots, \overline{a}_n\}$, on définit D_n^* comme la classe du mot vide 1 dans la congruence engendrée par les n relations $a_i \overline{a}_i = 1 \quad 1 \leqslant i \leqslant n$. D_n désigne les mots premiers, c'est-à-dire les mots de D_n^* qui n'ont aucun facteur gauche propre dans D_n^*.

Connus sous le nom de langages de Dyck, les langages D_n^* sont algébriques. Ainsi $D_n^* = L_G(\sigma)$ avec G donnée par $\langle \sigma \to a_i \, \sigma \, \overline{a}_i \, \sigma + 1, \; 1 \leqslant i \leqslant n \rangle$. De même, les langages D_n sont algébriques : $D_n = L_G(\sigma)$ avec $G = \langle \sigma \to a_i \, S \overline{a}_i \; ; \; S \to a_i \, S \overline{a}_i S + 1 \quad 1 \leqslant i \leqslant n \rangle$.

Similairement, on définit les langages $\hat{D}_n^*$ et $\hat{D}_n$: $\hat{D}_n^*$ est la classe de 1 dans la congruence engendrée par les 2n relations $a_i\bar{a}_i = \bar{a}_i a_i = 1$; $\hat{D}_n$ est à nouveau l'ensemble des mots premiers. Ces langages sont algébriques.

Tous ces langages sont classiques et jouent un rôle important en théorie des langages algébriques en raison du

Théorème de Chomsky-Schützenberger : L est algébrique sur X si et seulement si il existe un entier n, un morphisme ψ de $\hat{Z}_n^*$ dans X^* et un langage rationnel K sur $\hat{Z}_n$ tels que

$$L = \psi (D_n^* \cap K).$$

Le même résultat vaut avec $\hat{D}_n^*$ au lieu de D_n^* .

Comme par ailleurs, on construit facilement un morphisme φ_n tel que $D_n^* = \varphi_n^{-1}(D_2^*)$ (et $\hat{D}_n^* = \varphi_n^{-1}(\hat{D}_2^*)$) , on en déduit

Fait : $Alg = \mathscr{T}(D_2^*)$

et , $n \geqslant 2$, $D_n^* \approx D_n \approx \hat{D}_n^* \approx \hat{D}_n \approx D_2^*$.

Nous verrons plus loin que D_1^* , D_1 , $\hat{D}_1^*$ et $\hat{D}_1$ ont un statut différent. Alg est donc un cône rationnel principal de générateur D_2^* .

Un peu moins connu, et pourtant fort utile, est le langage E des expressions arithmétiques, il est défini comme la classe de d dans la congruence engendrée par la relation adbdc = d. On a alors $E = L_G(S)$ avec $G < S \to aSbSc + d >$ et $\mathscr{T}(E) = Alg$, soit $E \approx D_2^*$. Notons qu'en effaçant b ou d ou les deux, on trouve trois nouveaux générateurs de Alg.

Parmi les langages algébriques, on a très tôt particularisé la famille Lin des langages linéaires : une grammaire est linéaire si chaque membre droit de règle contient au plus une variable. Pour cette sous-famille, on dispose aussi d'un théorème de Chomsky-Schützenberger dans lequel D_n^* est remplacé par le langage symétrique S_n défini sur $\hat{Z}_n$ par $S_n = L_G(S)$ avec $G = \langle S \to a_i S \bar{a}_i + 1 \ , \ 1 \leqslant i \leqslant n \rangle$. A nouveau, on peut encoder S_n dans S_2 , i.e. $S_n = \varphi_n^{-1}(S_2)$, d'où il résulte

Fait : $n \geqslant 2 \quad S_n \approx S_2$ et $\mathcal{T}(S_2) = \text{lin}$.

Ici encore S_1 est à part. On sait enfin que Lin $\subsetneq$ Alg.

Si l'on revient maintenant aux langages laissés de côté, on vérifie facilement que

$D_1 = a_1 D_1^* \bar{a}_1$ et donc $D_1 \approx D_1^*$

$\hat{D}_1 = D_1 \cup \bar{D}_1$ et donc $D_1 \approx \hat{D}_1$

On sait aussi que $\mathcal{T}(D_1^*)$ est un sous-cône strict de Alg : c'est la famille des langages à un compteur Rocl. Cette famille est celle des langages algébriques reconnus par un automate à pile dont l'alphabet de pile est réduit à un seul symbole, ce qui fait que celle-ci est un compteur. Il est connu que $\mathcal{T}(S_2)$ et $\mathcal{T}(D_1^*)$ et $\mathcal{T}(\hat{D}_1^*)$ sont incomparables. On sait aussi que $\mathcal{T}(S_1) \subseteq \mathcal{T}(D_1^*) \cap \mathcal{T}(S_2)$.

Question ; A-t-on $\mathcal{T}(S_1) = \mathcal{T}(D_1^*) \cap \mathcal{T}(S_2)$?

Cette liste de langages classiques serait incomplète sans le "hardest context-free language" ou version non-déterministe du langage de Dyck. Ce langage est défini sur l'alphabet $\hat{Z}_n \cup \{[,],+\}$. On appelle bloc un mot de

la forme $[u_1 + u_2 + u_3 + \ldots + u_p]$; ainsi, l'ensemble des blocs est-il le langage rationnel $[(\hat{Z}_n \cup \{+\})^*]$. On appelle choix dans un bloc un facteur u de Z_n^* du bloc (facteur maximal sur $\hat{Z}_n$). Alors, $f \in H_n$ si et seulement si $f = f_1 f_2 f_3 \ldots f_r$ ou chaque f_i est un bloc et il existe un choix u_i dans chaque bloc f_i tel que $u_1 u_2 u_3 \ldots u_r \in D_n^*$.

Il est facile de voir que les langages H_n sout tous algébriques. On sait aussi que, comme dans le cas des langages de Dyck, $H_n \approx H_2$ $n \geqslant 2$. Enfin, ces langages ont été introduits car ils engendrent la famille Alg sans utiliser de morphismes directs :

Théorème de Shamir-Greibach : L est algébrique si et seulement si il il existe un morphisme φ tel que $L - \{1\} = \varphi^{-1}(H_2)$.

Avant de clore ce paragraphe, nous allons présenter quatre sous-familles classiques de Alg.

Les langages quasirationnels:

Cette famille est obtenue en construisant $Lin^{\sigma} = QRat$, la cloture par substitution de la famille des langages linéaires. C'est un cône non principal (strictement) inclus dans Alg. Il est formé des langages engendrables par les grammaires algébriques non expansives, i.e. dont aucune variable S ne peut se dériver en uSvSw avec u,v,w des mots.

Les langages à compteur itéré :

Cette famille est obtenue similairement à partir de Rocl : $Rocl^{\sigma} = Ict$. Ces langages sont ceux reconnus par un automate à pile dont les configurations sont dans le borné $y_1^* y_2^* y_3^* \ldots y_p^*$ si $\{y_1, y_2, y_3, \ldots, y_p\}$ est l'alphabet de pile.

Les langages de Greibach :

Cette famille est le plus grand sous-cône de Alg connu que l'on puisse décrire à partir de familles plus petites. Elle est obtenue par cloture par substitution de Lin U Rocl : $(\text{Lin U Rocl})^{\sigma}$ = Gre. C'est un cône rationnel non principal.

Les langages Non-Générateurs :

On peut définir le plus grand sous-cône strict de Alg, puisque Alg est un cône principal. Ce plus grand sous-cône est formé des langages algébriques tels que $\mathcal{T}(L) \subsetneq \text{Alg}$. C'est donc la famille des langages non-générateurs notée Nge. On ne sait à peu près rien de ce cône si ce n'est qu'il est fermé par substitution.

Conjecture 2 : Nge est un cône rationnel non principal.

On notera que si la conjecture 1 est vraie, alors la conjecture 2 l'est aussi. C'est autour de cette conjecture qu'ont été batis les langages du paragraphe suivant. De façon duale, on peut se demander s'il existe un plus petit sous-cône de Alg contenant strictement Rat.

Conjecture 3 : Quelque soit le cône rationnel $\mathcal{L}$ de langages (algébriques), soit $\mathcal{L}$ = Rat, soit il existe un cône rationnel $\mathcal{L}'$ tel que $\mathcal{L} \supsetneq \mathcal{L}' \supsetneq \text{Rat}$.

On notera qu'à nouveau, si la conjecture 1 est vraie, la conjecture 3 l'est aussi. Nous présenterons dans le paragraphe IV deux langages utiles à l'étude de cette dernière conjecture.

III. LANGAGES NON-GENERATEURS :

Au vu de ce qui précède, il apparait naturel d'espérer que Nge = Gre. Il n'en est rien. Le premier langage de Ng-Gre est assez facile à décrire. Sur $\hat{Z}_n$ On définit la substitution $s(a_i) = a_i$ $s(\bar{a}_i) = \bar{a}_i \, \bar{Z}_n^*$ $1 \leqslant i \leqslant n$.

$\Delta_n = s(D_n^*)$ est alors algébrique et satisfait $\Delta_n \in$ Nge-Gre ($n \geqslant 2$). On voit facilement que $\Delta_n \in$ Nge ; en revanche $\Delta_n \notin$ Gre exige une preuve technique très longue. Par ailleurs, on sait que Δ_{n+1} domine strictement Δ_n . On voit facilement que $\Delta_1 \in$ Rocl [3] .

Un second langage dans Nge-Gre a été contruit plus tard [7] . Il est plus facile de montrer que celui-ci est dans Nge-Gre. Il est défini sur $\{a,b,c,d,\#\}$ ainsi

$$\#(E) = \{f\#^p \mid f \in E \text{ et } p \geqslant 1\} \cup \{f\#^p \mid |f| \geqslant p \ f \in \{a,b,c,d,\}^*\}.$$

Ces deux langages contiennent des langages rationnels infinis. En outre, Greibach a prouvé qu'un langage à un compteur (i.e. du cône $\mathscr{T}(D_1^*)$) ne contenant aucun rationnel infini (IRS en anglais) était quasirationnel, ce qui donne : Gre $\cap$ Irs = QRat [9].

Greibach a conjecturé alors que tout langage ne contenant aucun rationnel infini était soit générateur (comme E), soit quasirationnel. Cette conjecture a été établie dans deux cas particuliers : les langages parenthétiques [5] et les langages très simples [6] . Cependant, elle est fausse [4] .
Sur $X = \{a,b,c,d\}$, on considère la congruence d = adtdc qui permet de définir E. Etant donné un mot f, on définit e(f) comme le plus long facteur gauche de f qui soit préfixe d'un mot de E. A ce facteur gauche est associé un mot irréductible red(e(f)) dont la longueur est dite norme de e(f), notée $\| e(f) \|$. Il est facile de voir que $\| e(f) \|$ est la hauteur de la pile après lecture de e(f) par l'automate standard reconnaissant E. On définit alors le langage particulier :

$$L_p = \{f \in X^* \mid f = e(f)g \text{ et } \| e(f) \| < |g| < 2\| e(f) \| \} .$$

On montre alors que L_p appartient à (Nge-Gre) $\cap$ Irs.

Nous terminerons ce bref panorama par un résultat récent : tout cône clos par substitution, strictement contenu dans Alg, est contenu dans un cône principal de non-générateurs. Ainsi, Nge n'est-il sûrement pas cloturé par substitution d'un cône rationnel plus petit. Ce résultat est établi à partir d'une construction générale d'un langage $L^{\uparrow}$ déduit d'un langage L tel que

$$\mathcal{T}(L^{\uparrow}) = \text{Alg} \iff \mathcal{T}(L) = \text{Alg et } \mathcal{T}(L^{\uparrow}) \supseteq \mathcal{T}^{\sigma}(L).$$

Si L est défini sur X, $L^{\uparrow}$ est défini sur $X \cup \hat{Z}_1$ ($X \cap \hat{Z}_1 = \emptyset$). Etant donné un mot sur $X \cup \hat{Z}_1$, ou bien la projection sur $\hat{Z}_1$ n'est pas dans D_1, auquel cas le mot est dans L, ou bien elle est dans D_1 et pour appartenir à L il doit satisfaire : entre deux parenthèses associées, le sous-mot obtenu en ne gardant que les lettres de X qui ne sont pas elles-mêmes entre parenthèses est dans L.

Exemple : $f = axaxx\bar{a}axy\bar{a}yaz\bar{a}\bar{a}$ projection $aa\bar{a}aaaa\bar{a} \in D_1$.

Aux parenthèses externes correspond le sous-mot $x(axx\bar{a})(axy\bar{a})y(az\bar{a})$, soit xy.

Aux parenthèses intérieures correspondent les sous-mots xx, xy et z.

Ainsi $f \in L^{\uparrow}$ si et seulement si $xx, xy, z \in L$.

V. PETITS LANGAGES ALGEBRIQUES :

Si l'on cherche des condidats à être générateurs d'un cône rationnel minimal, c'est-à-dire tel que $\mathcal{T}(L)$ contienne strictement Rat sans qu'aucun cône ne puisse être placé entre $\mathcal{T}(L)$ et Rat, on s'aperçoit vite que $\mathcal{T}(L)$ ne doit contenir aucun langage borné non rationnel. On connait fort peu de tels langages algébriques. Ils sont tous décrits par leurs complémentaires.

Le langage de Goldstine [8] :

Sur l'alphabet $\{a,b\}$, on définit $\overline{G} = \{aba^2ba^3b \ldots a^nb \mid n \geqslant 1\}$ et $G = (X^*/\overline{G}) \cap X^*b$. Le langage G est algébrique et $\mathcal{T}(G)$ ne contient aucun langage borné non rationnel. On sait que G n'est pas minimal : une hiérarchie infinie décroissante de cônes rationnels a été construite à partir de G [1].

Le langage de Paterson :

Sur l'alphabet $\{a,b\}$, on définit $\overline{P} = \{(a^nb)^n \mid n>1\}$ et $P = (X^*/\overline{P}) \cap X^* b$. A nouveau, P est algébrique et $\mathcal{T}(P)$ ne contient aucun langage borné non rationnel.

VI. CONCLUSION :

Il ressort clairement de ce texte qu'il y a peu de langages algébriques connus : s'il ne sont pas tous là ceux qui manquent sont en effet très voisins de l'un de ceux présentés. Ainsi, avant de chercher à établir un résultat il convient non seulement de vérifier celui-ci sur les langages connus donnés ci-dessus, mais aussi de tenter de construire un nouveau type de langage algébrique qui contredise la conjecture étudiée. On a ainsi toutes les chances de faire progresser notre connaissance de la famille Alg.

BIBLIOGRAPHIE

[1] AUTEBERT J.M. et BEAUQUIER J. et BOASSON L et LATTEUX M. , Very Small Families of Algebraic Non Rational Languages. In formal language theory R.V. Book ed., A.P. (1980), p. 89-108.

[2] BERSTEL J., Transductions and Context-Free Languages. Teubner VEr. (1979).

[3] BOASSON L., The inclusion of the substitution closure of linear and one-counter languages in the largest Sub-AFL of CFL's is proper. Infor. Proc. Letter, 2 (1973), p. 135-140.

[4] BOASSON L., Language algébrique particulier. Rairo 13 (1979) p. 203-215.

[5] BOASSON L.et NIVAT M., Parenthesis generators, 17th Annual IEE symp. Found. of Comp. Science Houston (1976) p. 253-257.

[6] FROUGNY C., Langages très simples générateurs. RAIRO 13, 1 (1979), p. p. 69-86.

[7] GABARRO J., Une application des notions de centre et index rationnel à certains langages algébriques. RAIRO 16, 4 (1982), p. 317-329.

[8] GOLDSTINE J., Substitution and bounded languages. JCSS 6 (1972) p. 9-29.

[9] GREIBACH S., One-counter languages and the IRS condition, JCSS 10 (1975), p. 237-247.

§§§§§§§§§§