

LE CYLINDRE DES LANGAGES LINEAIRES N'EST PAS PRINCIPAL

L. BOASSON et M. NIVAT

La présente communication répond à une question posée par S. Greibach [1,4]: existe-t-il un langage linéaire dont tout langage linéaire soit image dans un homomorphisme inverse? La réponse est non ainsi qu'il est suggéré dans [4].

Rappelons qu'un langage linéaire est un langage algébrique (ou "context-free") engendré par une grammaire algébrique linéaire (i.e. dont les membres droits de règles contiennent au plus une lettre non-terminale). Par définition, une famille de langages \mathcal{L} constitue un cylindre si elle est fermée par homomorphisme inverse et intersection par les langages rationnels (ou réguliers). Notons qu'un cylindre fermé par homomorphisme est un cône rationnel [1]. Un cylindre \mathcal{C} est principal s'il existe un langage L tel que le plus petit cylindre contenant L soit précisément \mathcal{C} . Le but de cette communication est d'établir le

Théorème: La famille Lin des langages linéaires constitue un cylindre qui n'est pas principal.

Notons que la famille Lin, qui constitue un cône rationnel, est principale en tant que cône et pas en tant que cylindre, alors qu'au contraire la famille des langages algébriques est principale non seulement en tant que cône rationnel, mais aussi en tant que cylindre [4]. La question reste ouverte de savoir si le cylindre des langages à compteur est principal ou non.

Nous donnons ci-dessous les principaux éléments nécessaires à l'établissement de la preuve du théorème annoncé.

A - Les langages \hat{S}_n :

Sur l'alphabet $Z_n = Y_n \cup \bar{Y}_n$ où $Y_n = \{a_i \mid i=1, \dots, n\}$ et $\bar{Y}_n = \{\bar{a}_i \mid i=1, \dots, n\}$, on définit de façon classique le langage symétrique S_n

$$S_n = \{a_{i_1} a_{i_2} \dots a_{i_p} \bar{a}_{i_p} \dots \bar{a}_{i_2} \bar{a}_{i_1} \mid p \geq 1 \ a_{i_j} \in Y_n\}.$$

Désignant par R_n le langage rationnel $\{a_i \bar{a}_i \mid i=1, \dots, n\}^+$, on construit le langage S_n'' à partir de S_n en substituant dans ce dernier langage à chaque occurrence de $a \in Y$ le langage rationnel aR_n et à chaque $\bar{a} \in \bar{Y}$ le langage rationnel $(\bar{a}R_n \cup \bar{a})$. Notons que, puisque S_n est un langage linéaire et que l'opération ainsi définie est une substitution rationnelle, S_n'' est également linéaire. Il en va alors de même du langage $S_n' = R_n \cdot S_n''$.

Considérant alors deux nouveaux symboles $\#$ et α , on peut définir le langage

$$\hat{S}'_n = \{ \# u_1 \alpha v_1 \omega v_1 \# \dots \# u_p \alpha v_p \omega v_p \mid p \geq 1, u_i, w_i \in (Z_n \cup \{\alpha\})^* \text{ et } v_1 \dots v_p \in S'_n \}$$

Il est facile de voir que \hat{S}'_n est un langage linéaire (comme S'_n) et donc que le langage \hat{S}_n défini ci-dessous l'est aussi:

$$\hat{S}_n = \{ g \mid g = \# \alpha a_1 \alpha \# f \in \hat{S}'_n \text{ et } \# f \in K_n \} \text{ où}$$

K_n est le langage rationnel $(\{ \# \alpha, \alpha \} \bar{V}(Y^2 \cup Y \cup \{1\}))^*$.

L'intérêt des langages \hat{S}_n vient du

Lemme 1: Le langage L sur l'alphabet X est linéaire propre si et seulement si il existe un entier n et un homomorphisme ϕ de X^* dans Z_n^* tel que

$$L = \phi^{-1} \hat{S}_n.$$

Ce lemme est établi en étudiant d'assez près la forme normale de Greibach [3] des grammaires linéaires (la preuve est ensuite analogue à celle de [4] pour les langages algébriques). Cette étude des règles linéaires unilatères droites (i.e. de la forme $v \rightarrow v'_m$ où v' est un non-terminal) suggère un traitement simultané de règles linéaires unilatères

et permet de démontrer qu'à tout langage linéaire L on peut associer un langage linéaire pur L' (i.e. engendré par une grammaire linéaire sans règle unilatère) et une substitution rationnelle propre σ telle que $L = \sigma(L')$. Il en résulte que le langage symétrique S_2 est un générateur fidèle [1] du cône rationnel Lin et, en vertu d'un résultat de [1] la

Proposition: Si L est un générateur du cône rationnel Lin, $L \cup \{1\}$ en est un générateur fidèle.

B - Le résultat principal:

Nous commençons par une remarque d'ordre général: le langage L' sur l'alphabet X_L , est élément du cylindre engendré par le langage L sur l'alphabet X_L si et seulement si il existe un langage rationnel K (sur l'alphabet de L') et un homomorphisme ϕ du monoïde libre X_L^* , dans X_L^* tels que $L' = \phi^{-1} L \cap K$. Ceci provient de ce que si $L' = \phi^{-1}(L \cap R)$, on a $L' = \phi^{-1} L \cap \phi^{-1} R$ et du fait que la composition de deux homomorphismes est un homomorphisme.

Nous considérons alors sur l'alphabet $X_p = \{x_1, x_2, \dots, x_p\}$ les langages suivants:

- les rationnels $R = X_1^+ X_2^+ \dots X_p^+$ et $\bar{R} = X_p^* - R$.
- les langages linéaires (où Z et t sont deux nouveaux symboles)

$$A_p = \{ Z^{n_1} X_1^{n_1} X_2^{n_2} \dots X_p^{n_p} t X_p^{n_p} \dots X_2^{n_2} X_1^{n_1} Z^{n_1} \mid n \geq 1, n_i \geq 1 \text{ pour } i=1, \dots, p \}$$

$$A'_p = \{ Z^{n_1} X_1^{n_1} X_2^{n_2} \dots X_p^{n_p} t X_p^{n_p} \dots X_2^{n_2} X_1^{n_1} Z^{m_1} \mid n, m \geq 1, n \neq m, n_i \geq 1 \text{ pour } i=1, \dots, p \}$$

$$L_p = A_p \cdot R$$

$$L'_p = A'_p \cdot \bar{R}$$

$$L''_p = L_p \cup L'_p.$$

Supposant alors que le langage L''_p est élément du cylindre engendré par le langage S, et donc que $L''_p = \phi^{-1} S \cap \bar{K}$ où \bar{K} est un langage rationnel, on établit d'abord

qu'il existe un entier r tel que

$$\bar{K} \ni (Z^r)^+ (x_1^r)^+ \dots (x_p^r)^+ t (x_p^r)^+ \dots (x_1^r)^+ x_p^*$$

(Cette preuve est identique à celle classique qui permet d'établir que si un langage rationnel K contient tous les mots de la forme $a^n b^m$ où $n \neq m$, il existe un entier r tel que K contienne $(a^r)^+ (b^r)^+$).

Il en résulte alors facilement que si $L_p'' = \phi^{-1} S \cap \bar{K}$, il existe un homomorphisme λ tel que $L_p = \lambda^{-1} S \cap K$ où K est le langage rationnel $K = Z^+ x_1^+ x_2^+ \dots x_p^+ t x_p^+ \dots x_2^+ x_1^+ Z^+ x_p^*$.

Le lemme essentiel à la preuve de notre théorie est alors le suivant:

Lemme 2: Pour $p > n^2$, il n'existe pas d'homomorphisme λ tel que
 $L_p = \lambda^{-1} \hat{S}_n \cap K$ où K est le langage rationnel donné ci-dessus.

L'idée de la preuve de ce lemme est la suivante. Supposant que $L_p = \lambda^{-1} \hat{S}_n \cap K$, on s'assure que, pour tout $i=1, \dots, p$ λx_i contient au moins une occurrence de $\#$. Puis on considère comment la contrainte d'égalité des nombres de Z à gauche de t et à droite de t peut apparaître par λ : ceci impose que dans les mots de \hat{S}_n dont proviennent les mots de L_p les facteurs itérants situés à droite de la dernière occurrence de Z (i.e. les occurrences de x_i) proviennent de facteurs itérants de mots de \hat{S}_n , et donc d'éléments de la forme $\bar{a}_i a_j$. On trouve alors que si $p > n^2$, on pourra échanger deux tels facteurs itérants ce qui est impossible dans L_p .

La preuve du théorème est alors immédiate: supposons que Lin soit un cylindre admettant le langage H comme générateur principal. Alors, puisque $H \in Lin$, il existe en vertu du lemme 1 un entier n et un homomorphisme ϕ tel que $H = \phi^{-1} \hat{S}_n$ et donc, il existe un entier n_0 tel que \hat{S}_{n_0} soit générateur principal du cylindre Lin . Mais alors L_p'' est, pour tout p , élément du cylindre engendré par \hat{S}_{n_0} ; il résulte donc de nos remarques préliminaires que pour tout p , il existe un homomorphisme λ tel que $L_p = \lambda^{-1} \hat{S}_{n_0} \cap K$, ce qui contredit le lemme 2.

Pour conclure, notons qu'un tel résultat négatif n'établit pas qu'il n'y ait aucun langage linéaire de complexité "maximale". En effet, bon nombre d'opérations ne modifiant pas la classe de complexité ne sont pas prises en compte ici (par exemple, l'intersection par des langages linéaires déterministes). En outre, la preuve du résultat suggère deux possibilités: elle fournit un système générateur du cylindre Lin , la famille des \hat{S}_n et il semble donc que si l'on sait reconnaître tous ces langages "raisonnablement", il en ira de même de tout langage linéaire. En outre, il apparaît qu'à défaut d'un générateur du cylindre, la substitution du langage symétrique S_2 dans \hat{S}_2 fournit un langage qui doit contenir dans le cylindre qu'il engendre toute la famille Lin (sans qu'il soit lui-même linéaire). Ainsi, la complexité des langages linéaires ne serait-elle pas plus "grande" que celle de ce dernier langage.

- [1] BOASSON L. et M. NIVAT: Sur Diverses Familles de Langages Fermées par Transductions Rationnelles.
Acta Informatica, 2, 180-188 (1973).
- [2] GINSBURG S.: The Mathematical Theory of Context-Free Languages.
Mc Graw Hill (1966).
- [3] GREIBACH S.: Jump pda's, Deterministic Context-Free Languages, Principal AFDL's and Polynomial Time Recognition.
Proc. of 5th Annual ACM Symposium on Theory of Computing, Austin, (Texas), 20-28 (1973).
- [4] GREIBACH S.: The Hardest Context-Free Language.
Siam J. of Comput., 2, n^o4, 304-330 (1975).

L. BOASSON

U.E.R. de Mathématiques
Université de Picardie

Ad. Postale: 5, Allée Georges Rouault
75020 - PARIS (F).

M. NIVAT

U.E.R. de Mathématiques
Université Paris 7

Ad. Postale: 8 rue Portalis
75009 - PARIS (F).