LANGAGES RECONNAISSABLES ET CODAGE PREFIXE PUR

Jean-Eric PIN

Université Paris VI et CNRS, Tour 55-65, 4è étage
4 Place Jussieu 75230 Paris CEDEX 05 - FRANCE

Le théorème des variétés d'Eilenberg établit une correspondance bijective entre certaines classes de langages reconnaissables, les variétés de langages, et certaines classes de monoïdes finis, les variétés de monoïdes. Il est donc naturel que les opérations sur les variétés de langages correspondent à des opérations sur les variétés de monoïdes. Expliciter cette correspondance pour diverses opérations constitue un problème fondamental qui intéresse à la fois la classification des langages reconnaissables et celle des monoïdes finis. Ce problème a été résolu pour les morphismes littéraux (ou strictement alphabétiques) et les substitutions inverses ([6], [7], [10], [15], [18]), pour le produit de concaténation [17] et pour le produit non-ambigu et ses variantes [12]. Dans chaque cas, fait remarquable, l'opération correspondante sur les variétés de monoïdes était déjà connue antérieurement, mais pour d'autres raisons.

On peut aussi partir d'une opération sur les variétés de monoïdes et tenter de déterminer l'opération qui lui correspond sur les variétés de langages. Dans ce domaine, les opérations du type $\underline{V} \to \underline{W}^{-1}\underline{V}$ et $\underline{V} \to \underline{V} * \underline{W}$ (où $\underline{W}$ est une variété de semigroupes dans le premier cas, de monoïdes dans le second cas) semblent particulièrement importantes. Ainsi l'opération $\underline{V} \to \underline{A}^{-1}\underline{V}$ (où $\underline{A}$ désigne la variété des semigroupes apériodiques) correspond au produit de concaténation et l'opération $\underline{V} \to \underline{LI}^{-1}\underline{V}$ (où $\underline{LI}$ désigne la variété des semigroupes localement triviaux) correspond au produit non-ambigu. L'opération correspondant à $\underline{V} \to \underline{Nil}^{-1}\underline{V}$ n'est en revanche pas encore connue mais sa connaissance permettrait probablement de fournir une nouvelle démonstration du théorème de Simon sur les langages testables par morceaux. Par ailleurs, des travaux récents de Straubing (à paraître) montrent que l'opération $\underline{V} \to \underline{V}*\underline{LI}$ joue un rôle important dans l'étude des hiérarchies du type de celle de Brzozowski [2].

Dans cet article, nous nous intéressons plus particulièrement à l'opération $\underline{V} \to \underline{V}*\underline{A}$, qui à une variété de monoïdes $\underline{V}$, associe la variété $\underline{V}*\underline{A}$ engendrée par les produits semidirects de la forme $M*N$ où M est un monoïde de $\underline{V}$ et où N est un monoïde apériodique. Cette opération, bien qu'étudiée depuis fort longtemps en liaison avec la théorie de la complexité des semigroupes (au sens de Krohn et Rhodes,

cf. [19]) est encore fort mal connue. Malgré ce handicap, nous décrivons complètement l'opération sur les variétés de langages qui lui est associée, et qui ne fait intervenir que des opérations élémentaires connues : le produit d'une lettre par un langage et les morphismes injectifs (ou codages) préfixes purs (i.e. les morphismes injectifs $\alpha : A^* \to B^*$ tels que le langage $A\alpha$ soit un code préfixe pur). A titre d'application, on en déduit une description de la variété des langages dont le monoïde syntactique est de complexité $\leqslant 1$. Ce dernier résultat laisse entrevoir de nouvelles méthodes d'approche pour résoudre la question suivante, ouverte depuis quinze ans ([3] et [19]) : peut-on décider si la complexité d'un monoïde est $\leqslant 1$?

Les démonstrations font appel soit à des méthodes récentes, telles que la simulation d'un automate par un code ([9], [10], [11]), soit à des méthodes plus anciennes mais récemment perfectionnées, par exemple le principe du produit semi-direct [16] ou l'utilisation des transductions pour l'étude des opérations [1]. Notons que la même technique permettrait d'étudier les opérations du type $\underline{V} \to \underline{V}*\underline{W}$ pour d'autres variétés $\underline{W}$: cf [13].

1. Préliminaires

Nous nous limiterons à quelques rappels. Pour plus de détails, le lecteur pourra consulter les livres de Berstel [1], Eilenberg [4] ou Lallement [5].

Une variété de monoïdes (semigroupes) est une classe de monoïdes (semigroupes) fermée par division et par produit direct fini. En particulier, la variété des monoïdes (semigroupes) apériodiques (ou "group-free" en anglais) sera notée $\underline{A}$, et celle des groupes, $\underline{G}$.

Une variété de langages $\mathcal{V}$ associe à chaque alphabet A une algèbre de Boole $A^*\mathcal{V}$ telle que :

(1) Pour tout alphabet A, si $a \in A$ et si $L \in A^*\mathcal{V}$, alors $a^{-1}L$, $La^{-1} \in A^*\mathcal{V}$

(2) Si $\varphi : A^* \to B^*$ est un morphisme de monoïdes, $L \in B^*\mathcal{V}$ entraine $L\varphi^{-1} \in A^*\mathcal{V}$

Le théorème des variétés d'Eilenberg assure l'existence d'une correspondance bijective entre variétés de monoïdes et variétés de langages. Dans la suite, le terme "variété correspondante" fera toujours référence à ce théorème.

Si $\underline{V}$ et $\underline{W}$ sont deux variétés de monoïdes, on note $\underline{V}*\underline{W}$ la variété engendrée par les produits semidirects $M*N$ avec $M \in \underline{V}$ et $N \in \underline{W}$. Soit $\underline{V}$ une variété de semigroupes. Un morphisme de monoïdes $\varphi : M \to N$ est un $\underline{V}$-morphisme si pour tout sous-semigroupe N' de N élément de $\underline{V}$, $N'\varphi^{-1} \in \underline{V}$. Si $\underline{W}$ est une variété de monoïdes, on note $\underline{V}^{-1}\underline{W}$ la variété engendrée par les monoïdes M tels qu'il existe un $\underline{V}$-morphisme $\varphi : M \to N$ avec $N \in \underline{W}$. En particulier, Straubing [17] a montré que si $\mathcal{V}$ est la variété de langages qui correspond à $\underline{V}$, $\underline{A}^{-1}\underline{V}$ est la variété de monoïdes qui correspond à la fermeture de $\mathcal{V}$ par produit de concaténation.

Un code préfixe est une partie P de A^+ telle que $u, uv \in P$ entraine $v = 1$.

Un morphisme $\alpha : A^* \to B^*$ est appelé codage préfixe si $A\alpha$ est un code préfixe : dans ce cas, α est nécessairement un morphisme injectif. Un code (préfixe) P est dit pur si $u^n \in P^*$ entraine $u \in P^*$. On démontre qu'un code fini P est pur si et seulement si le monoïde syntactique de P^* est apériodique.

On notera U_1 le monoïde $\{0,1\}$ muni de la multiplication usuelle des entiers et MoN le produit en couronne du monoïde M par le monoïde N.

Enfin, on rappelle qu'un morphisme $\eta : A^* \to M$ reconnait un langage L s'il existe une partie P de M telle que $L = P\eta^{-1}$. On dit aussi dans ce cas que M reconnait L.

2. Résultat principal

Dans tout ce qui suit, $\underline{V}$ est une variété de monoïdes et $\mathcal{V}$ est la variété de langages correspondante. Le résultat qui suit donne une description de la variété correspondant à $\underline{V*A}$.

Théorème 2.1 La variété de langages correspondant à $\underline{V*A}$ est la plus petite variété $\mathcal{W}$ satisfaisant les conditions suivantes

(1) Pour tout alphabet A, $A^*\mathcal{W}$ contient les langages de la forme $L\alpha$ où $L \in B^*\mathcal{V}$ et où $\alpha : B^* \to A^*$ est un codage préfixe pur.

(2) Pour tout alphabet A, si $L \in A^*\mathcal{W}$ et $a \in A$, alors $aL \in A^*\mathcal{W}$

De façon moins formelle, on peut dire que la variété de langages correspondant à $\underline{V*A}$ est la plus petite variété contenant les langages $L\alpha$ où L est un langage de $\mathcal{V}$ et α un codage préfixe pur et qui soit fermée pour l'opération $L \to aL$ où a est une lettre.

La preuve s'appuie sur les résultats suivants :

Proposition 2.2 Soit $L \subset A^*$ un langage et a une lettre de A. Si le monoïde M reconnait L, le monoïde $(MxU_1)oU_1$ reconnait aL.

Proposition 2.3 Soit $\alpha : A^* \to B^*$ un codage préfixe pur. Si le monoïde M reconnait L, alors L est reconnu par MoN, où N est un monoïde apériodique ne dépendant que de α.

La méthode de démonstration de ce type de résultats est exposée en détail en [14] ou en [13]. L'idée principale consiste à considérer l'opération étudiée (ici $L \to aL$ et $L \to L\alpha$) comme l'inverse d'une transduction. Il est à noter que le produit de Schützenberger de deux monoïdes, qui est la construction "classique" pour l'étude du produit de concaténation, ne permet pas d'obtenir la proposition 2.2.

Le principe du produit semidirect, que nous énonçons ci-dessous, est une version simplifiée du "wreath-product principle" de Straubing [16], qui est lui-même le premier énoncé complet d'un résultat utilisé antérieurement dans des cas

particuliers, notamment par A. Meyer et Brzozowski-Simon.

Soit $\eta : A^* \to M*N$ un morphisme de A^* dans un produit semidirect de monoïdes et soit $\pi : M*N \to N$ la projection canonique. On a alors, en posant $B = NxA$ et $\varphi = \eta\pi$:

Proposition 2.4 (Principe du produit semidirect) Si L est reconnu par η, alors L est réunion de langages de la forme $X \cap Y\sigma^{-1}$ où $X \subset A^*$ est reconnu par N, $Y \subset B^*$ est reconnu par M et où $\sigma : A^* \to B^*$ est la fonction séquentielle définie par $1\sigma = 1$ et $(a_1a_2\ldots a_n)\sigma = (1,a_1)(a_1\varphi, a_2)\ldots((a_1a_2\ldots a_{n-1})\varphi, a_n)$

Démonstration Il suffit d'établir le résultat dans le cas où $L = (m_o,n_o)\eta^{-1}$ pour un certain $(m_o,n_o) \in M*N$. Selon l'usage, nous noterons additivement le monoïde M et par juxtaposition l'action de N sur M. Soit $\alpha : B = NxA \to M$ défini par $(t,a)\alpha = tm$ où m est la première composante de $a\eta$ (i.e. $a\eta = (m,n)$ pour un certain $n \in N$). Posons $X = m_o\varphi^{-1}$ et $Y = n_o\alpha^{-1}$. X est donc reconnu par N et Y par M. Il vient $X \cap Y\sigma^{-1} = \{a_1\ldots a_k \in A^* \mid (a_1\ldots a_k)\varphi = n_o$ et $(a_1\ldots a_k)\sigma\alpha = m_o\}$ Or, en posant $a_i\eta = (m_i,n_i)$, on a successivement

$$(a_1\ldots a_k)\sigma\alpha = ((1,a_1)(a_1\varphi,a_2)\ldots((a_1\ldots a_{k-1})\varphi, a_k))\alpha$$
$$= (1,a_1)\alpha+(a_1\varphi,a_2)\alpha \ldots+((a_1\ldots a_{k-1})\varphi,a_k)\alpha$$
$$= m_1 + n_1 m_2 +\ldots+ n_1\ldots n_{k-1}m_k$$

et $(a_1\ldots a_k)\varphi = n_1n_2\ldots n_k$

On en déduit $((a_1\ldots a_k)\sigma\alpha , (a_1\ldots a_k)\varphi) = (a_1\ldots a_k)\eta$ d'où finalement $X \mathbin{\textcircled{\scriptsize *}} Y\sigma^{-1} = (m_o,n_o)\eta^{-1}$, ce qui démontre la proposition.

Preuve du théorème 2.1

On notera $M(L)$ le monoïde syntactique de L. Soit $\mathfrak{U}$ ($\mathfrak{V}$) la variété correspondant à $\underline{V*A}$ ($\underline{V}$) et soit $\mathfrak{W}$ la plus petite variété de langages satisfaisant les conditions (1) et (2) de l'énoncé du théorème.

Si $L \in A^*\mathfrak{U}$ et $a \in A$, on a d'après 2.2,

$$M(aL) < (M(L) \times U_1) \circ U_1$$

Comme $M(L) \in \underline{V}*\underline{A}$ et que $U_1 \in \underline{A}$, on a aussi $M(aL) \in \underline{V*A}$ et donc $\mathfrak{U}$ satisfait la condition (2). De même, si $\alpha : A^* \to B^*$ est un codage préfixe pur, on a d'après 2.3,

$$M(L\alpha) < M(L) \circ N \text{ pour un certain } N \in \underline{A}$$

et donc $M(L\alpha) \in \underline{V*A}$, ce qui montre que $\mathfrak{U}$ satisfait la condition (1). On en déduit finalement $\mathfrak{W} \subset \mathfrak{U}$.

L'inclusion opposée est plus difficile à établir. La première étape est constituée par le lemme suivant :

Lemme 2.5 $\mathfrak{W}$ contient la variété des langages apériodiques.

Preuve Soit P un code préfixe fini pur sur un alphabet A et soit $\alpha : B^* \to A^*$ un morphisme de codage tel que $B\alpha = P$. Alors $B^* \in B^*\mathfrak{W}$ par définition d'une variété et donc $B^*\alpha = P^* \in A^*\mathfrak{W}$ d'après la condition (1) du théorème 1. Or d'après un résultat de [11] la variété des langages apériodiques est la plus petite variété contenant les langages P^* avec P préfixe fini pur. (En fait l'énoncé original est donné pour les +variétés, mais le résultat dont nous avons besoin s'en déduit aisément).

Soit maintenant $L \in A^*\mathfrak{U}$. Il existe alors $M \in \underline{V}$ et $N \in \underline{A}$ tels que L soit reconnu par $\eta : A^* \to M*N$. D'après le principe du produit semidirect, L peut s'écrire comme union de langages de la forme $X \cap Y\sigma^{-1}$ avec $X \subset A^*$ reconnu par N, $Y \subset B^*$ reconnu par M, $B = N\times A$ et où σ est la fonction séquentielle définie plus haut. Puisque $N \in \underline{A}$, on a $X \in A^*\mathfrak{W}$ d'après le lemme 2.5. D'autre part, puisque $M \in \underline{V}$, on a $Y \in B^*\mathfrak{V}$ et il suffit d'établir que $Y\sigma^{-1} \in A^*\mathfrak{W}$.

L'étape suivante consiste à décomposer la transduction σ^{-1}.
Posons $N = \{z_0, z_1, \ldots, z_n\}$ où $z_0 = 1$, neutre de N. On définit une action de A sur $\{0,1,\ldots,n\}$ en posant, pour $0 \leqslant i \leqslant n, z_{ia} = z_i(a\varphi)$. Soit τ une fonction de $\mathbb{N}$ dans $\mathbb{N}$ satisfaisant les conditions :

(1) $0\tau = 0$

(2) $a\tau + b\tau = c\tau + d\tau$ entraine $\{a,b\} = \{c,d\}$

Une telle fonction existe, par exemple $n\tau = 2^n - 1$. Finalement, posons $C = A \cup \{c\}$ où c est une nouvelle lettre. Le résultat (non trivial) suivant est démontré en [9]

Proposition 2.5 Si N est apériodique, le morphisme $\alpha : B^* \to C^*$ défini par $(z_k, a)\alpha = c^{k\tau}ac^{n\tau - k\alpha\tau}$ est un codage préfixe pur

Il reste à définir un morphisme $\gamma : A^* \to C^*$ par $a\gamma = c^{n\tau}a$ et on obtient la décomposition souhaitée :

Proposition 2.6 Pour tout langage Y de B^*, on a l'égalité :

$$Y\sigma^{-1} = (c^{n\tau}((Y\alpha)(c^*)^{-1}))\gamma^{-1}$$

Preuve Soit $u = (z_{i_1}, a_1)\ldots(z_{i_n}, a_r)$. On a

$$c^{n\tau}((u\alpha)(c^*)^{-1}) = \{c^{n\tau + i_1\tau}a_1c^{n\tau - i_1a_1\tau}c^{i_2\tau}a_2c^{n\tau - i_2a_2\tau}\ldots c^{i_r\tau}a_rc^j \mid 0 \leqslant j \leqslant n\tau - i_ra_r\tau\}$$

Par conséquent $[c^{n\tau}(u\alpha)(c^*)^{-1}]\gamma^{-1}$ est vide sauf dans le cas où

$$i_1\tau = 0 \qquad i_1a_1\tau = i_2\tau \quad \dots \quad i_{r-1}a_{r-1}\tau = i_r\tau$$

i.e. $i_1 = 0 \qquad i_2 = i_1a_1 \quad \dots \quad i_r = i_{r-1}a_{r-1}$

Dans ce dernier cas, on a
$(c^{n\tau}(u\alpha)(c^*)^{-1})\gamma^{-1} = a_1 \dots a_r$

Par ailleurs, $u\sigma^{-1}$ est vide sauf dans le cas où

$$u = (1,a_1)(a_1\beta,a_2)\dots((a_1\dots a_{r-1})\beta,a_r)$$

c'est-à-dire dans le cas où

$$z_{i_1} = 1 \qquad z_{i_2} = z_{i_1}(a_1\beta) \quad \dots \quad z_{i_r} = z_{i_{r-1}}(a_{r-1}\beta)$$

soit encore dans le cas où

$$i_1 = 0 \qquad i_2 = i_1a_1 \quad \dots \quad i_r = i_{r-1}a_{r-1}$$

Dans ce dernier cas, on a $u\sigma^{-1} = a_1 \dots a_r$, ce qui établit la proposition 2.6

On voit que les opérations utilisées dans la décomposition de σ^{-1} sont soit des opérations de variétés ($L \to L(c^*)^{-1}$ et $L \to L\gamma^{-1}$) soit l'une des opérations de codage préfixe pur ($L \to L\alpha$) ou de produit à gauche par une lettre ($L \to cL$). On en déduit, puisque $Y \in B^*\mathbb{V} \subset B^*\mathbb{W}$, que $Y\sigma^{-1} \in A^*\mathbb{W}$ ce qui conclut la preuve du théorème 2.1.

Voici une version un peu plus faible du théorème 2.1 :

Corollaire 2.7 La variété de langages correspondant à $\underline{V*A}$ est la plus petite variété contenant $\mathbb{V}$ qui soit fermée par codage préfixe pur et par l'opération $L \to aL$ où a est une lettre.

Preuve La seule chose à vérifier, par rapport au théorème 2.1, est que la variété correspondant à $\underline{V*A}$ soit fermée par codage préfixe pur, ce qui résulte de la proposition 2.3.

3. Lien avec la complexité des semigroupes

Définissons une suite croissante de variétés $\underline{V}_n$ par les formules :

$$\underline{V}_0 = \underline{A} \quad \text{et} \quad \underline{V}_{n+1} = \underline{V}_n * \underline{G} * \underline{A}$$

On peut énoncer le théorème classique de Krohn et Rhodes sous la forme suivante : tout monoïde (fini) M appartient à l'une au moins des variétés $\underline{V}_n$. Le plus petit entier k tel que $M \in \underline{V}_k$ est appelé la complexité de M. Il découle également des résultats généraux de la théorie de la complexité (cf. Tilson [19]) l'importante égalité :

$$\text{pour tout } n \geqslant 0, \ \underline{A}^{-1}\underline{V}_n = \underline{V}_n.$$

Il en résulte en particulier, d'après le résultat de Straubing rappelé plus haut, que les variétés de langages correspondant aux variétés $\underline{V}_n$ sont fermées par produit. Compte tenu de la proposition 2.3, on peut donc énoncer :

Proposition 3.1 Pour tout $n \geqslant 0$, la variété de langages correspondant à $\underline{V}_n$ est fermée par produit et par codage préfixe pur.

La variété $\underline{V}_1$ peut être étudiée de façon plus complète. Rappelons qu'un langage à groupe est un langage dont le monoïde syntactique est un groupe. On déduit alors du théorème 2.1 le

Théorème 3.2 La variété de langages correspondant à $\underline{V}_1$ est la plus petite variété de langages contenant les langages à groupe qui soit fermée par produit et par codage préfixe pur.

On peut améliorer le théorème 3.2 en remplaçant les langages à groupe par une famille de langages explicitement donnée. Pour celà, nous aurons besoin d'un résultat de [8], dont une version moins précise est donnée en [9]

Proposition 3.3 Soit n un entier positif et soit M un monoïde de $\underline{V}_n$.

Il existe alors un code préfixe fini P effectivement constructible tel que

(1) M divise le monoïde syntactique N de P^*

(2) $N \in \underline{V}_n$

En appliquant ce résultat à $\underline{V} = \underline{V}_1$ et à $M = \mathfrak{S}_n$, le groupe symétrique sur n lettres, on obtient le code P_n, construit sur l'alphabet $A = \{a,b\}$ et défini par les formules :

$$P_{n+1} = \{a^{2^i} b a^{2^n - 2^{i+1}} \mid 0 \leqslant i \leqslant n-1\} \cup \{a^{2^n} b a^{2^n - 1}, \ b^2 a^{2^n - 2}, \ ab^2 a^{2^n - 1}\}$$
$$\cup \{a^{2^{i-1}} b^2 a^{2^n - 2^i} \mid 2 \leqslant i \leqslant n\}$$

On en déduit finalement :

Théorème 3.4 La variété de langages correspondant à $\underline{V}_1$ est la plus petite variété de langages contenant les langages $P_n^*(n \geqslant 1)$ et qui soit fermée par produit et par codage préfixe pur.

BIBLIOGRAPHIE

[1] J. BERSTEL Transductions and Context-Free Languages. Teubner (1979)

[2] J. BRZOZOWSKI Hierarchies of aperiodic languages, RAIRO Informatique Théorique, Vol 10 (1976), 33-49.

[3] J. BRZOZOWSKI Open problems about regular languages, Proc. Formal Language Theory Symposium, Santa Barbara, December 1979.

[4] S. EILENBERG Automata, Languages and Machines, Vol B, Academic Press, New-York (1976).

[5] G. LALLEMENT Semigroups and Combinatorial applications, Wiley, New-York, (1979).

[6] S.W. MARGOLIS à paraître

[7] S.W. MARGOLIS et J.E. PIN Minimal non-commutative variéties of finite monoïds. Soumis pour publication.

[8] S.W. MARGOLIS et J.E. PIN On variéties of rational languages and variable-length codes II. En préparation.

[9] J.E. PIN On varieties of rational languages and variable-length codes. A paraître dans Journal of Pure and Applied Algebra.

[10] J.E. PIN Variétés de langages et monoïde des parties, Semigroup Forum vol 20 (1980) 11-47.

[11] J.E. PIN Une caractérisation de trois variétés de langages bien connues 4th GI Conference, Lecture Notes in Computer Science N° 67 Springer (1979) 233-243.

[12] J.E. PIN Propriétés syntactiques du produit non ambigu. 7th ICALP. Lecture Notes in Computer Science N° 85 Springer (1980) 483-499.

[13] J.E. PIN Variétés de langages et variétés de semigroupes. Thèse Paris (1981).

[14] J.E. PIN et J. SAKAROVITCH Une application de la représentation matricielle des transductions. En préparation.

[15] Chr. REUTENAUER Sur les variétés de langages et de monoïdes. 4th GI conference Lecture Notes in Computer Science 67, Springer (1979), 260-265.

[16] H. STRAUBING Varieties of recognizable sets whose syntactic monoids contain solvable groups. Ph. D. Thesis, University of California, Berkeley (1978)

[17] H. STRAUBING Aperiodic homomorphisms and the concatenation product of recognizable sets. J. Pure and Applied Algebra, Vol. 15, 1979, 319-327.

[18] H. STRAUBING Recognizable sets and power séts of finite semigroups. Semigroup Forum Vol 18 (1979) 331-340.

[19] B. TILSON Chapitres 11 et 12 de la référence (4).