MATHÉMATIQUES ET SCIENCES HUMAINES

LUC BOASSON
MAURICE NIVAT

Transductions et familles de langages

Mathématiques et sciences humaines, tome 35 (1971), p. 31-37
<http://www.numdam.org/item?id=MSH_1971__35__31_0>

Math. Sci. hum. (9e année, n° 35, 1971, p. 31-37)

TRANSDUCTIONS ET FAMILLES DE LANGAGES

par

Luc BOASSON et Maurice NIVAT [1]

*La théorie des langages formels a des applications évidentes dans le champ des langages de programmation et l'on peut s'interroger sur son utilité pour une meilleure compréhension des mécanismes de production et de reconnaissance d'une langue naturelle. Or, le problème de l'*analyse syntaxique *d'une phrase consiste à décrire les algorithmes de reconnaissance qui permettent de décider, lorsqu'on connaît les règles de formation des phrases d'un langage, si cette phrase donnée appartient ou non au langage et ensuite, en cas de réponse favorable, à donner une suite de règles au moyen desquelles cette phrase peut être formée. C'est pour répondre à ce problème que l'on a introduit les* transductions, *applications d'un monoïde libre dans un autre monoïde, qui sont des généralisations naturelles des homomorphismes de monoïde libre. Les résultats obtenus dans la théorie des langages formels, n'ont certes pas tous une interprétation linguistique immédiate, mais c'est un exemple de théorie mathématique où :*

1° *Les concepts mathématiques sont idéalisés (langage formel, grammaire formelle, syntaxe) à partir d'objets linguistiques non mathématiques ;*

2° *On pousse « les conséquences jusqu'au bout » et étudie par l'algèbre les conséquences qui peuvent devenir des résultats (positif ou plus souvent négatif); voir « l'attitude formalisante en linguistique »*, Math. Sci. hum., *n° 34, qui serviront à tester l'adéquation d'un modèle linguistique.*

J. P. Desclés

L'utilisation systématique de la notion de transductions rationnelles dans l'étude des familles de langages (en particulier de langages algébriques [5]) permet de simplifier un grand nombre de définitions et de résultats. Nous donnons ci-dessous définitions et résultats concernant les AFL ou familles abstraites de langages introduites dans [3]. Et nous décrivons quelques problèmes qui font l'objet de travaux en cours.

I. TRANSDUCTIONS RATIONNELLES

Rappelons qu'un monoïde est un ensemble muni d'une loi de composition associative que nous noterons multiplicativement et d'un élément neutre pour cette loi que nous noterons e.

1. Institut de Programmation, Faculté des Sciences, Paris.

Dans l'ensemble des parties d'un monoïde M, on définit les opérations d'union, produit, étoile et étoile tronquée de la façon suivante :

— l'union est l'union ensembliste,

— le produit de A et B noté AB est égal à :

$$AB = \{ f g \mid f \in A \quad \text{et} \quad g \in B \},$$

— l'étoile tronquée de A noté A^+ est égale à :

$$A^+ = \bigcup_{n \geqslant 1} A^n,$$

— l'étoile de A notée A^* est égale à :

$$A^* = \{ e \} \cup A^+.$$

Très généralement, on appelle *partie rationnelle* de M, tout élément de la plus petite famille de parties de M contenant les parties finies et fermée pour les opérations d'union, produit et étoile.

Les parties rationnelles du monoïde libre X^* engendré par l'alphabet fini sont aussi appelées *langages rationnels*, *langages réguliers* ou *langages de Kleene.*

Nous appelons *transduction* du monoïde libre X^* dans le monoïde libre Y^* toute application de X^* dans l'ensemble des parties de Y^*.

Si τ est une transduction de X^* dans Y^*, nous notons $\hat{\tau}$ le sous-ensemble de $X^* \times Y^*$ défini par :

$$\hat{\tau} = \{ (f, g) \mid f \in X^*, g \in \tau (f) \},$$

τ^{-1} la transduction de Y^* dans X^* définie par :

$$\tau^{-1} (g) = \{ f \in X^* \mid (f, g) \in \hat{\tau} \}$$

pour tout $g \in Y^*$.

La transduction τ de X^* dans Y^* est dite :

— *d'image finie*, si et seulement si, $\tau (f)$ est un sous-ensemble fini de Y^* pour tout $f \in X^*$,

— *fidèle* si et seulement si, τ^{-1} est d'image finie,

— *propre* si et seulement si, $\tau (f) \subset Y\ Y^*$ pour tout f dans X^*,

— *rationnelle* si et seulement si, $\hat{\tau}$ est une partie rationnelle de $X^* \times Y^*$.

Les *transductions rationnelles* ont été introduites par Elgot et Mezei [1] et étudiées par Nivat [5] qui en donne la caractérisation suivante :

Théorème 1

La transduction τ de X^* dans Y^* est rationnelle si et seulement s'il existe un alphabet Z, un langage rationnel K dans Z et deux homomorphismes alphabétiques φ et ψ de Z^* dans X^* et Y^* respectivement tels que :

$$\hat{\tau} = \{ (\varphi (h), \psi (h)) \mid h \in K \}.$$

On rappelle que l'homomorphisme φ de Z^* dans X^* est *alphabétique* (resp. strictement alphabétique, resp. continu) si et seulement si :

$$\varphi (Z) \subset X \cup \{ e \} \quad (\text{resp. } \varphi (Z) \subset X, \quad \text{resp. } \varphi (Z) \subset X\ X^*).$$

On peut alors écrire :

$$\tau (f) = \psi (\varphi^{-1} (f) \cap K)$$

et

$$\tau^{-1} (g) = \varphi (\psi^{-1} (g) \cap K)$$

pour tout $f \in X^*$ et $g \in Y^*$.

Nous donnons ci-dessous trois lemmes nouveaux permettant de préciser cet énoncé :

Lemme 1

On peut toujours supposer, dans l'énoncé du théorème 1, que pour tout $z \in Z$, l'un au moins des deux mots $\varphi(z)$ et $\psi(z)$ n'est pas vide. (Ce lemme nous a été suggéré par J. Berstel.)

Lemme 2

On peut toujours supposer, dans l'énoncé du théorème 1, que Z ne contient pas deux lettres distinctes z_1 et z_2 telles que :

$$\varphi(z_1) = \varphi(z_2) \qquad \text{et} \qquad \psi(z_1) = \psi(z_2).$$

Lemme 3

La transduction rationnelle τ de X^* dans Y^* est *fidèle et propre* si et seulement s'il existe un alphabet Z, un langage rationnel K sur Z, un homomorphisme φ de Z^* dans X^* et un homomorphisme ψ strictement alphabétique de Z^* dans Y^* tels que :

$$\hat{\tau} = \{ (\varphi(h), \psi(h)) \mid h \in K \}.$$

Démonstration du lemme 3

1) *La condition est suffisante.* Nous savons que :

$$\tau^{-1}(g) = \varphi(\psi^{-1}(g) \cap K)$$

pour tout g dans Y^*.

Si ψ est strictement alphabétique, tout mot h de Z^* tel que $\psi(h) = g$ est de même longueur que g. Mais alors, si $M = \max \{ |\varphi(z)| ; z \in Z \}$; on sait que $|\varphi(h)| \geqslant M \,.\, |h|$. On en déduit que tout mot f dans X^* élément de $\tau^{-1}(g)$ est borné en longueur par $M \,.\, |g|$. On en déduit que $\tau^{-1}(g)$ est un sous-ensemble fini de X^* quelque soit g dans Y^* et donc que τ est fidèle.

2) *La condition est nécessaire.* τ étant une transduction rationnelle, on peut trouver un alphabet Z, un rationnel K sur Z et deux homomorphismes alphabétiques φ et Ψ de Z^* dans X^* et Y^* respectivement, tels que :

$$\tau(f) = \psi(\varphi^{-1}(f) \cap K)$$

pour tout $f \in X^*$.

Soit N_0 l'entier associé au rationnel K, tel que :

$$h = h_1 h_2 h_3 \in K \qquad \text{avec} \qquad |h_2| > N_0 \Rightarrow \exists\, u \neq 1$$

tel que $h_2 = h_2' u h_2''$ et $h_1 h_2' u^* h_2'' h_3 \subset K$.

Soit enfin :

$$Z_0 = \{ z \in Z \mid \psi(z) = e \} \qquad \text{et} \qquad Z_1 = Z / Z_0.$$

D'après le lemme 1, on peut supposer que si $z \in Z_0$, $\varphi(z)$ est non vide.

α) K ne contient pas de mot admettant plus de N_0 lettres de Z_0 consécutives : en effet, supposons qu'il existe un mot h dans K tel que :

$$h = h_1 h_2 h_3, \quad h_2 \in Z_0^* \qquad \text{et} \qquad |h_2| > N_0.$$

Alors, on sait que :

$$h_2 = h_2' u h_2'' \qquad u \neq 1 \qquad \text{et} \qquad h_1 h_2' u^* h_2'' h_3 \subset K.$$

Or :

$$\psi(h_1 h_2' u^* h_2'' h_3) = \psi(h_1 h_3) = g,$$

$$\varphi(h_1 h_2' u^* h_2'' h_3) = \varphi(h_1 h_2') (\varphi u)^* \varphi(h_2'' h_3) = f_1 \alpha^* f_2$$

et $\alpha = \varphi(u) \neq e$.

Donc : $f_1 \alpha^* f_2 \subset \tau^{-1}(g)$ et τ n'est pas fidèle, ce qui contredit l'hypothèse.

β) Soit alors un alphabet $T = T_1 \cup T_2$ défini par :

$$T_1 = \{ t_\alpha \mid t \in Z_1 \quad \text{et} \quad \alpha \in \bigcup_{n=0}^{N_0} Z_0^n \}$$

$$T_2 = \{ {}_\alpha t_\beta \mid t \in Z_1 \quad \text{et} \quad \alpha, \beta \in \bigcup_{n=0}^{N_0} Z_0^n \}.$$

On définit les deux homomorphismes φ' et ψ' dans X^* et Y^* respectivement par :

$$\psi'(t_\alpha) = \psi(t) \in Y, \qquad \varphi'(t_\alpha) = \varphi(t_\alpha),$$
$$\psi'({}_\alpha t_\beta) = \psi(t) \in Y, \qquad \varphi'({}_\alpha t_\beta) = \varphi({}_\alpha t_\beta).$$

Notons qu'alors ψ' est strictement alphabétique.

Soit θ, l'homomorphisme T^* dans Z^* défini par :

$$\theta(t_\alpha) = t_\alpha$$
$$\theta({}_\alpha t_\beta) = {}_\alpha t_\beta.$$

Soit enfin $K' = \theta^{-1} K \cap T_2 T_1$. On vérifie que la traduction rationnelle τ de X^* dans Y^* définie par $\tau'(f) = \psi'(\varphi'^{-1}(f) \cap K')$, est égale à τ. En effet :

$$g \in \tau(f) \Leftrightarrow \exists\, h \in K \text{ tel que } \varphi(h) = f \text{ et } \psi(h) = g$$
$$g \neq e \text{ car } \tau \text{ est propre. Donc :}$$
$$h = h_1 \alpha_1 h_2 \alpha_2 \ldots h_p \alpha_p h_{p+1}$$
$$h_i \in Z_0^* \quad |h_i| < N_0$$
$$\alpha_i \in Z_1$$
$$\Leftrightarrow \exists\, h' \in K'$$
$$h' = {}_{h_1}\alpha_{1h_2}\alpha_{2h_3}\alpha_{3h_4} \ldots \alpha_{ph_{p+1}}$$
$$\text{tel que } \varphi'(h') = f$$
$$\psi'(h') = g$$
$$\Leftrightarrow g \in \tau'(f)$$

C.Q.F.D.

II. FAMILLES ABSTRAITES DE LANGAGES

Nous traduisons par *famille abstraite de langages*, l'expression « abstract family of languages » en abrégé AFL, introduite par Ginsburg et Greibach [3]. Nous dirons nous aussi AFL pour abréger. Les semi-AFL ont été introduits par Greibach [4].

Définition 1

Une famille $\mathscr{F}$ de langages est dite constituer un *semi-AFL* (resp. *semi-AFL plein*) si et seulement si, elle est fermée par homomorphisme continu (resp. homomorphisme), homomorphisme inverse, intersection avec un langage rationnel et union.

Nous noterons $\mathscr{T}(\mathscr{F})$ (resp. $\mathscr{T}^+(\mathscr{F})$) la famille des langages de la forme $\tau(L)$ où $L \in \mathscr{F}$ et τ est une transduction rationnelle (resp. une transduction rationnelle fidèle). On vérifie immédiatement.

Propriété 1

La famille $\mathscr{F}$ de langages, fermée par union est un semi-AFL (resp. un semi-AFL plein) si et seulement si, $\mathscr{F} = \mathscr{T}^+(\mathscr{F})$ (resp. $\mathscr{F} = \mathscr{T}(\mathscr{F})$).

Définition 2

Une famille $\mathscr{F}$ de langages est dite constituer un AFL (resp. AFL plein) si et seulement si, c'est un semi-AFL (resp. semi-AFL plein) fermé par produit et étoile tronquée (resp. étoile).

On vérifie immédiatement la

Propriété 2

La famille $\mathscr{F}$ de langages, rationnellement fermée, est un AFL (resp. un AFL plein) si et seulement si, $\mathscr{F} = \mathscr{T}^+(\mathscr{F})$ (resp. $\mathscr{F} = \mathscr{T}(\mathscr{F})$). La famille $\mathscr{F}$ est dite rationnellement fermée si et seulement si, $\mathscr{F}$ est fermée par union, produit et étoile tronquée.

La théorie des langages fournit un nombre considérable d'AFL et d'AFL pleins ; en général, il n'est pas difficile d'établir qu'une famille de langages est un AFL ou un AFL plein. Par contre, une question généralement difficile est de déterminer pour un AFL (resp. AFL plein) donné $\mathscr{A}$ les systèmes minimaux de générateurs, c'est-à-dire les sous-familles minimales $\mathscr{F}$ de $\mathscr{A}$ telles que $\mathscr{A}$ soit le plus petit AFL (resp. AFL plein) contenant $\mathscr{F}$. Pour toute famille $\mathscr{F}$, nous notons $\mathscr{A}(\mathscr{F})$ le plus petit AFL plein contenant $\mathscr{F}$. Dans cette étude, un rôle important est dévolu aux *AFL principaux.*

Définition 3

Un AFL plein $\mathscr{A}$ est dit principal si et seulement s'il existe un langage L tel que $\mathscr{A} = \mathscr{A}(L)$. ∎

Remarque

Il est bien connu que l'ensemble de tous les langages *context-free* (que, conformément à la terminologie d'Eilenberg, nous dénommons langages algébriques) constitue un AFL plein. C'est même un AFL plein principal comme il résulte directement du théorème de Chomsky-Schützenberger ; un langage générateur est par exemple l'ensemble de Dyck sur quatre lettres, c'est-à-dire, sur l'alphabet $\{ a, b, \bar{a}, \bar{b} \}$ la classe d'équivalence du mot vide pour la congruence engendrée par :

$$a\bar{a} = \bar{a}a = b\bar{b} = \bar{b}b = e.$$

Un exemple d'AFL plein non principal nous est fourni par l'ensemble des *langages quasi-rationnels* que l'on définit simplement de la façon suivante :

— la grammaire *algébrique* G est dite *inexpansive* si et seulement s'il n'existe pas de symbole non terminal v qui se dérive en un mot de la forme $f_1 v f_2 v f_3$ avec $f_1 f_2 f_3 \neq e$;

— le langage L est dit quasi-rationnel si et seulement s'il est engendré par une grammaire algébrique inexpansive. Le fait que ces langages constituent un AFL plein se vérifie sans peine. Que cet AFL soit non principal, résulte du théorème 3.2. de [4].

Le théorème suivant est dû à Greibach et Ginsburg [2].

Théorème 2

Le plus petit AFL plein $\mathscr{A}$ contenant le langage L sur X peut se mettre sous la forme $\mathscr{A} = \mathscr{T}((L\alpha)^*)$ où α est un symbole quelconque n'appartenant pas à X.

Démonstration

1) On sait que tout rationnel est élément de $\mathscr{A}$ car $\mathscr{A}$ est un AFL plein. En conséquence, la fermeture par produit et étoile assure que $(L\alpha)^* \in \mathscr{A}$. Mais alors, comme un AFL plein est formé par transduction rationnelle, on a :

$$\mathscr{T}((L\alpha)^*) \subset \mathscr{A}.$$

2) Pour démontrer l'inclusion inverse, nous allons montrer que $\mathscr{T}((L\alpha)^*)$ est rationnellement fermé. Soient L_1 et L_2 deux langages éléments de $\mathscr{T}((L\alpha)^*)$. On peut donc trouver un alphabet Z_i, un rationnel K_i sur Z_i et deux homomorphismes alphabétiques φ_i et ψ_i de Z^* dans X^* et Y_i^* tels que :

$$L_i = \psi_i(\varphi_i^{-1}((L\alpha)^*) \cap K_i)$$

pour $i = 1, 2 \qquad (Z_1 \cap Z_2 = \emptyset)$.

Alors, si φ et ψ désignent les homomorphismes de $(Z_1 \cup Z_2)^*$ dans X^* et $(Y_1 \cup Y_2)^*$ naturellement définies à partir de φ_1, ψ_1, φ_2 et ψ_2, on a :

$$L_1 \cup L_2 = \psi(\varphi^{-1}((L\alpha)^*) \cap (K_1 \cup K_2)).$$

On peut supposer que $K_1 = K_1' \beta$ avec $\varphi_1(\beta) = \alpha$ et β est un symbole ne figurant pas dans les mots de K_1'. En effet, soit $Z_\alpha = \{ z \in Z_1 \mid \varphi_1(z_1) = \alpha \}$. On peut évidemment supposer que K_1 est inclus dans $Z_1^* Z_\alpha$. Soit alors $\bar{Z}_\alpha = \{ \bar{z} \mid z \in Z_\alpha \}$ un ensemble disjoint de $Z_1 \cup \{ \beta \}$. On définit $\bar{K}_1$ par :

$$\bar{K}_1 = \{ h \bar{z} \beta \mid h z \in K \}.$$

On pose :

$$\varphi_1(\bar{z}) = 1, \varphi_1(\beta) = \varphi_1(z) = \alpha,$$
$$\psi_1(\bar{z}) = \psi(z), \psi_1(\beta) = 1.$$

On a alors :

$$L_1 = \psi_1(\varphi_1^{-1}((L\alpha)^*) \cap \bar{K}_1)$$

avec $\bar{K}_1 = \bar{K}_1' \beta$.

Alors :

$$L_1 L_2 = \psi(\varphi^{-1}((L\alpha)^*) \cap \bar{K}_1 K_2)$$

et

$$L_1^* = \psi_1(\varphi_1^{-1}((L\alpha)^*) \cap \bar{K}_1^*).$$

C.Q.F.D.

III. REMARQUES EN GUISE DE CONCLUSION

Il est bien évident que s'il existe une transduction rationnelle τ_1 de X^* dans Y^* telle que $\tau_1(L_1) = L_2$ et une transduction τ_2 de Y^* dans X^* telle que $\tau_2(L_2) = L_1$ les deux AFL pleins principaux engendrés par L_1 et L_2 sont identiques. Nous disons dans ce cas que L_1 est *rationnellement équivalent* à L_2.

Le problème de montrer que deux langages donnés sont ou ne sont pas rationnellement équivalents est un problème difficile. C'est actuellement le sujet de travaux de L. Boasson qui a démontré que l'ensemble de Dyck sur deux lettres n'est pas rationnellement équivalent à l'ensemble de semi-Dyck également sur deux lettres (l'*ensemble de Dyck* (resp. semi-Dyck) sur $\{ a, b \}$ est la classe d'équivalence de e pour la congruence engendrée par $ab = ba = e$ (resp. $ab = e$)).

C'est aussi le sujet de travaux de J. Berstel qui décrit les classes de cette équivalence rationnelle pour les langages bornés. Signalons enfin les travaux récents de S. Greibach [4].

BIBLIOGRAPHIE

[1] Elgot, C. et Mezei, J., "On relations defined by generalized finite automata", *IBM journal of research and development*, 9, 1965, pp. 47-68.

[2] Ginsburg, S. et Greibach, S., "Principal A F L", *Journal of computer system sciences*, 4, 1970, pp. 308-338.

[3] GINSBURG, S., GREIBACH, S. et HOPCROFT, J., *Studies in abstract family of languages*, memoirs of the American math. Soc. 1969, p. 87.

[4] GREIBACH, S., "Chains of full A FL", *Mathematical system theory*, 4 (3), 1970, pp. 231-242.

[5] NIVAT, M., "Transduction des langages de Chomsky", *Ann. Inst. Fourier*, Grenoble, 18 (1), 1968, pp. 339-456.