REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE ROUGE

G. THIERRIN

Décomposition des langages réguliers

Revue française d'informatique et de recherche opérationnelle. Série rouge, tome 3, n° R3 (1969), p. 45-50

<http://www.numdam.org/item?id=M2AN_1969__3_3_45_0>

R.I.R.O.
(3e année, N° R-3, 1969, p. 45-50)

DECOMPOSITION DES LANGAGES REGULIERS

G. THIERRIN

Département d'Informatique, Université de Montréal.

Résumé. — *Dans cet article, on établit quelques propriétés des langages réguliers simples à droite et d'une classe particulière de ces langages, celle des langages réguliers purs à droite. On montre que tout langage régulier simple à droite est d'une manière unique produit d'un langage régulier pur à droite et d'un langage régulier étoilé obtenu à partir d'un langage régulier pur à droite. De là on déduit une décomposition des langages réguliers à partir des langages réguliers purs à droite.*

1. DEFINITIONS ET RESULTATS PRELIMINAIRES

Soit $I = \{ a_1, a_2, ..., a_n \}$ un alphabet fini non vide et soit I^* le demi-groupe libre avec élément unité Λ engendré par I. Tout sous-ensemble U de I^* est dit un *langage* (sur I). Soit $a \in I^*$. Le *quotient à droite* $U .^{\cdot} a$ et le *quotient à gauche* $U^{\cdot} . a$ de U par a sont définis respectivement par

$$U .^{\cdot} a = \{ x \mid x \in I^*, ax \in U \} \quad , \quad U^{\cdot} . a = \{ x \mid x \in I^*, xa \in U \}$$

A tout langage U est associée une *congruence à droite* notée R_U et une *congruence à gauche* notée ${}_UR$ définies respectivement par

$$a \equiv b(R_U) \Leftrightarrow U .^{\cdot} a = U .^{\cdot} b, \quad a \equiv b({}_UR) \Leftrightarrow U^{\cdot} . a = U^{\cdot} . b$$

Un langage U est dit *régulier* si et seulement si U est réunion de classes d'une congruence d'index fini de I^*. Il est bien connu [3] que les propriétés suivantes sont équivalentes :

1) U est un langage régulier.
2) U est réunion de classes d'une congruence à droite d'index fini de I^*.
3) U est réunion de classes d'une congruence à gauche d'index fini de I^*.
4) La congruence à droite R_U est d'index fini.
5) La congruence à gauche ${}_UR$ est d'index fini.
6) U est l'ensemble des mots acceptés par un automate fini.

Un langage régulier est dit *simple à droite* si et seulement si U est une classe de R_U. On a alors nécessairement $U \neq \varnothing$.

Étant donné un langage régulier $U \neq \varnothing$, les propriétés suivantes sont équivalentes :

1) U est simple à droite.

2) U est classe d'une congruence à droite de I^*.

3) La relation $Ua \cap U \neq \varnothing$ entraîne $Ua \subseteq U$.

Un langage U est dit *étoilé* si et seulement si U est un sous-demi-groupe de I^* contenant l'élément unité Λ. Si V est un langage régulier, alors

$$V^* = \bigcup_{i=0}^{\infty} V^i$$

est un langage régulier étoilé.

Si U est un langage étoilé, tout langage V tel que $U = V^*$ est dit une *racine* de U. D'après [1], tout langage régulier étoilé U possède une *racine minimale unique* $V = (U - \Lambda) - (U - \Lambda)^2$ et cette racine est un langage *régulier*.

Un langage U est dit *pur à droite* si et seulement si $U \neq \varnothing$ et si $Ua \cap U \neq \varnothing$ entraîne $a = \Lambda$. Tout langage régulier pur à droite est évidemment simple à droite. Si $\Lambda \in U$, alors $U = \{ \Lambda \}$.

Soit U un langage régulier non vide. Si $\Lambda \in U$, un élément $u \in U$ est dit *pur à droite* si et seulement si $u = \Lambda$. Si $\Lambda \notin U$, un élément $u \in U$ est dit *pur à droite* si et seulement si $u = u_1 a$ avec $u_1 \in U$ entraîne $a = \Lambda$. Il est évident que U est pur à droite si et seulement si tous ses éléments sont purs à droite.

2. LANGAGES REGULIERS SIMPLES A DROITE

Pour tout langage U, posons

$$U.^{\cdot} U = \{ x \mid x \in I^*, Ux \subseteq U \}$$

Proposition 1. — Soit U un langage régulier non vide. Alors $U.^{\cdot} U$ est un langage régulier étoilé. Si de plus U est simple à droite, alors $U.^{\cdot} U$ est aussi simple à droite.

Preuve. — U étant régulier, la congruence à gauche ${}_UR$ est d'index fini. Montrons que $S = U.^{\cdot} U$ est réunion de classes de ${}_UR$. Soit $s \equiv a({}_UR)$ avec $s \in S$. De $Us \subseteq U$ suit $U \subseteq U^{\cdot}.s$. Mais $U^{\cdot}.s = U^{\cdot}.a$. Donc $U \subseteq U^{\cdot}.a$ et $Ua \subseteq U$. D'où $a \in S$. Il est immédiat que S est un sous-demi-groupe de I^* contenant Λ.

Supposons maintenant U simple à droite et montrons que S est classe de R_S. Soit $sa \in S$ avec $s \in S$. Alors $Usa \subseteq U$ et $Ua \cap U \neq \varnothing$. Comme U est simple,

on a $Ua \subseteq U$ et $a \in S$. Il suit de là que pour tout $s \in S$, on a $S \cdot^* s = S$ et donc S est contenu dans une classe de R_S. Soit $s \equiv a(R_S)$. Alors

$$S \cdot^* s = S \cdot^* a = S.$$

Donc $aS \subseteq S$. Mais $\Lambda \in S$ et donc $a \in S$, ce qui montre que S est une classe de R_S.

Proposition 2. — Soit U un langage régulier simple à droite, soit U_p l'ensemble des éléments purs à droite de U et soit $U_S = U \cdot^* U$. Alors :

1) $U_S = \{ x \mid x \in I^*, U_P x \subseteq U \}$ et $U = U_P U_S$.

2) Si $p, q \in U_P$ avec $p \neq q$, alors $pU_S \cap qU_S = \varnothing$.

3) Pour tout $p \in U_P$, pU_S est un langage régulier simple à droite.

4) U_P est un langage régulier pur à droite et si m et n sont respectivement le nombre de classes de R_{U_P} et R_U, on a $m \leqslant n + 2$.

Preuve. (1) L'ensemble U_P n'est pas vide et, comme U est simple à droite, la première partie de (1) est immédiate. De plus $U_P U_S \subseteq U$. Montrons que $U \subseteq U_P U_S$. Soit $u \in U$. Si u est pur à droite, alors $u = u\Lambda$ avec $u \in U_P$ et $\Lambda \in U_S$. Si u n'est pas pur à droite, alors il existe $u_1 \in U_P$ tel que $u = u_1 a$. D'où $Ua \cap U \neq \varnothing$ et $Ua \subseteq U$. Donc $a \in U_S$ et $u \in U_P U_S$.

2) Supposons $pU_S \cap qU_S \neq \varnothing$. Il existe alors $s, t \in U_S$ tels que $ps = qt$. Désignons par $lg(a)$ la longueur d'un élément $a \in I^*$. De $p \neq q$ et $ps = qt$ suit $lg(p) \neq lg(q)$.

Si $lg(p) < lg(q)$, alors $q = pa$ avec $a \neq \Lambda$ et q n'est pas pur à droite. Si $lg(q) < lg(p)$, on obtient un résultat analogue.

3) Le langage pU_S est régulier, car U_S est régulier (proposition 1). Montrons qu'il est simple à droite. Soit $pU_S a \cap pU_S \neq \varnothing$. Alors $U_S a \cap U_S \neq \varnothing$. Comme U_S est simple à droite, on a $U_S a \subseteq U_S$. D'où $pU_S a \subseteq pU_S$ et pU_S est simple à droite.

4) Pour montrer que U_P est régulier, il suffit de montrer que R_{U_P} est d'index fini. L'ensemble U_P est une classe de R_{U_P}. Soit $W = \{ w \mid w \in I^*, U_P \cdot^* w = \varnothing \}$. On a $W \neq \varnothing$. L'ensemble W est une classe de R_{U_P} et $U_P \cap W = \varnothing$. Soit K le complément de $U_P \cap W$. Si $a \equiv b(R_U)$ avec $a, b \in K$, on a $a \equiv b(R_{U_P})$, c'est-à-dire $U_P \cdot^* a = U_P \cdot^* b$. En effet, soit $ax \in U_P$. Puisque $U_P \subseteq U$ et $a \equiv b(R_U)$, on a $bx \in U = U_P U_S$. Supposons que $bx \notin U_P$. Alors $bx = ps$ avec $p \in U_P$, $s \in U_S$, $s \neq \Lambda$.

Si $lg(b) = lg(p)$, alors $b = p$ et $b \in U_P$, ce qui est contradictoire puisque $b \in K$.

Si $lg(b) < lg(p)$, alors $p = br \in U_P \subseteq U$. De $a \equiv b(R_U)$ suit $ar \equiv br(R_U)$ et donc $ar \in U$. De $bx = ps = brs$ suit $x = rs$, $ax = ars$. Comme $q = ax$ est

un élément pur à droite de U, $q \neq \Lambda$, alors de $q = ars$ avec $ar \in U$ suit $s = \Lambda$, ce qui est contradictoire.

Si $\lg(p) < \lg(b)$, alors $b = pr$ avec $r \neq \Lambda$. Comme $b \notin W$, il existe x tel que $bx = q \in U_P$. D'où $q = prx$ et q n'est pas un élément pur de U, ce qui est contradictoire.

On a donc montré que $U_P \cdot^{\cdot} a \subseteq U_P \cdot^{\cdot} b$. Par symétrie, on a l'inclusion inverse. D'où $U_P \cdot^{\cdot} a = U_P \cdot^{\cdot} b$.

De ce qui précède, il résulte que R_{U_P} peut avoir au plus deux classes de plus que R_U. Donc R_{U_P} est d'index fini.

Il est immédiat que U_P est pur à droite.

Proposition 3. — Un langage régulier non vide $U \neq \{ \Lambda \}$ est pur à droite si et seulement si U est la racine minimale d'un langage régulier étoilé simple à droite $\neq \{ \Lambda \}$.

Preuve. — Supposons U pur à droite et soit $U^* = \bigcup_{i=0}^{\infty} U^i$ le langage régulier étoilé engendré par U.

Montrons que U^* est simple à droite. Soit $U^*a \cap U^* \neq \varnothing$. Il existe $r, s \in U^*$ tels que $ra = s$. Si $r = \Lambda$, alors $a = s \in U^*$ et $U^*a \subseteq U^*$. Soit $r \neq \Lambda$. Alors $r = r_1 r_2 \ldots r_k$ avec $r_1, r_2, \ldots, r_k \in U$, $r_i \neq \Lambda$, et $s = s_1 s_2 \ldots s_n$ avec $s_1, s_2, \ldots, s_n \in U$, $s_j \neq \Lambda$. Comme U est pur à droite, l'égalité

$$r_1 r_2 \ldots r_k a = s_1 s_2 \ldots s_n$$

entraîne $r_1 = s_1, r_2 = s_2, \ldots, r_k = s_k$. D'où $a = s_{k+1} \ldots s_n$ ou $k = n$ et $a = \Lambda$. Par conséquent $U^*a \subseteq U^*$ et U^* est simple à droite. Il est immédiat que U est la racine minimale de U^*.

Soit W un langage régulier étoilé simple à droite $\neq \{ \Lambda \}$ et soit $V = W - \{ \Lambda \}$. Alors $V \neq \varnothing$. Soit U la racine minimale de W. D'après [1], on a $U = V - V^2$ et U est un langage régulier non vide. Montrons que U est pur à droite. Soit $Ua \cap U \neq \varnothing$. Il existe $u, v \in U$ tels que $ua = v$. De plus $Wa \cap W \neq \varnothing$, ce qui entraîne, puisque W est simple, $Wa \subseteq W$ et $a \in W$. Si $a \neq \Lambda$, alors $a \in V$ et $v = ua \in V^2$, ce qui est en contradiction avec $v \in U$. Donc $a = \Lambda$ et U est pur à droite.

3. UNE DECOMPOSITION DES LANGAGES REGULIERS

Proposition 4. — Tout langage régulier simple à droite non vide U possède une décomposition unique de la forme :

$$U = PQ^*$$

où P et Q sont des langages réguliers purs à droite.

Preuve. — Soit P l'ensemble des éléments premiers de U et soit $U_S = U \cdot^* U$. D'après la proposition 2, P est un langage régulier pur à droite. D'après la proposition 1, U_S est un langage régulier étoilé simple à droite. Si $U_S = \{ \Lambda \}$, alors $U = PQ^*$, où $Q = U_S$ est pur à droite. Soit $U_S \neq \{ \Lambda \}$ et soit Q la racine minimale de U_S. D'après la proposition 3, Q est un langage régulier pur à droite. Par conséquent $U = PQ^*$.

Soit $U = RV^*$ une autre décomposition de U, où R et V sont réguliers et purs à droite. On a $P \subseteq RV^*$ et comme P est l'ensemble des éléments purs à droite de U, il s'ensuit que $P \subseteq R$. Soit $r \in R, r = ps$ avec $p \in P, s \in Q^*$. Comme $p \in R$, on a $Rs \cap R \neq \varnothing$ et donc $s = \Lambda$. Par conséquent $P = R$.

D'autre part $V^* \subseteq U \cdot^* U = U_S = Q^*$. Supposons $V^* \neq Q^*$ et soit $t \in Q^*$, $t \notin V^*$. Si $p \in P$, alors $pt \in U$ et donc $pt = p_1 v$ avec $p_1 \in P$, $v \in V^*$. De là suit $p = p_1$ et $t = v \in V^*$, ce qui est contradictoire. Par conséquent $Q^* = V^* = U_S$. Si $U_S \neq \Lambda$, d'après la preuve de la proposition 3, V est la racine minimale de $V^* = U_S$. Donc $V = Q$. Si $U_S = \{ \Lambda \}$, alors $V = Q = \{ \Lambda \}$.

EXEMPLES. — Soit $I = \{ a, b, c \}$.

(1) $$U = \{ b^n a x \mid n \geqslant 0, x \in I^* \}$$

On a $U = PQ^*$ avec $P = \{ b^n a \mid n \geqslant 0 \}$, $Q = I$.

(2) $U = \{ ab^n a \mid n \geqslant 0 \}$

On a $U = PQ^*$ avec $P = U, Q = \{ \Lambda \}$.

(3) Soit k un entier fixé $\geqslant 0$ et soit $U = \{ a^r b^s a^n \mid n \geqslant 0, r + s = k \}$.

On a $U = PQ^*$ avec $P = \{ a^r b^s \mid r + s = k \}$, $Q = \{ a \}$.

D'après [4], un langage U est dit *premier* si $U = U_1 U_2$, où U_1 et U_2 sont des langages, entraîne $U_1 = \{ \Lambda \}$ ou $U_2 = \{ \Lambda \}$. D'après [2], tout langage régulier est le produit d'un nombre fini de langages étoilés et de langages premiers. En général, cette décomposition n'est pas unique. On voit facilement que tout langage régulier simple à droite et premier est pur à droite. L'inverse n'est pas vrai. Ainsi le langage $U = \{ ab^n a \mid n \geqslant 0 \}$ de l'exemple (2) ci-dessus est régulier et pur à droite, sans être premier puisque $U = U_1 U_2 U_3$ avec $U_1 = U_3 = a$, $U_2 = b^*$.

Proposition 5. — Un langage non vide U est régulier si et seulement s'il existe un nombre fini de langages réguliers purs à droite $P_1, \ldots, P_k, Q_1, \ldots, Q_k$ tels que

$$U = \bigcup_{i=1}^{k} P_i Q_i^*$$

Preuve. — Cela découle du théorème précédent et du fait qu'un langage est régulier si et seulement s'il est réunion de classes d'une congruence à droite d'index fini.

REFERENCES

[1] J. A. BRZOZOWSKI, « Roots of star events », *J. ACM*, vol. 14 (1967), p. 466-477.

[2] J. A. BRZOZOWSKI and R. COHEN, « On decompositions of regular events », *J. ACM*, vol. 16 (1969), p. 132-144.

[3] S. GINSBURG, *An introduction to mathematical machine theory*. Addison-Wesley, Reading, Mass., 1962.

[4] A. PAZ and B. PELEG, On concatenative decompositions of regular events. *IEEE Trans.*, vol. C-17 (1968), p. 229-237.