1 Langages non reconnaissables : lemmes de pompage (encore appelés lemmes de l'étoile)

Lemme 1.1 *Soit L un langage reconnaissable par automates. Alors il existe un entier $N > 0$ tel que pour tout mot $w \in L$ de longueur supérieure ou égale à N, il existe trois mots $u_1, u_2, u_3 \in A^*$ tels que $w = u_1u_2u_3$, $u_2 \neq \varepsilon$ et $u_1u_2^*u_3 \subset L$.*

Démonstration : Soit $\mathcal{A}$ un automate déterministe reconnaissant L, soit N le nombre de ses états, q_0 son état initial, δ sa fonction de transition. Soit $w = c_1 \dots c_p \in L$ avec $p \geq N$. Définissons l'état q_i par $q_i = \delta(q_{i-1}, c_i)$ si bien que q_p est un état final de l'automate. Comme $p \geq N$ les états $q_0, \dots, q_p$ ne sont pas distincts, donc il existe $i < j$ tels que $q_i = q_j$. Posons $u_1 = c_1 \dots c_i$, $u_2 = c_{i+1} \dots c_j \neq \varepsilon$ et $u_3 = c_{j+1} \dots c_p$. On a $q_i = q_j = \delta^*(q_i, u_2)$, si bien que l'on a encore, pour tout $n \in \mathbf{N}$,
$q_i = q_j = \delta^*(q_i, u_2^n)$, d'où $q_p = \delta^*(q_0, u_1u_2^nu_3)$, ce qui montre que le mot $u_1u_2^nu_3$ est encore reconnu par $\mathcal{A}$ donc appartient à L.

Application 1 *Soit $L = \{a^n b^n \mid n \in \mathbf{N}\} \subset \{a, b\}^*$. Alors L n'est pas reconnaissable par automate.*

Démonstration Supposons que L est reconnu par un automate et soit N comme dans le lemme. Soit $w = a^N b^N \in L$. On peut écrire $w = u_1 u_2 u_3$ avec $u_2 \neq \varepsilon$ et $\forall n \in \mathbf{N},\ u_1 u_2^n u_3 \in L$ et en particulier $w_2 = u_1 u_2^2 u_3 \in L$. Mais alors trois cas ont possibles :

- si u_2 ne contient que des a, w_2 contient plus de a que de b
- si u_2 ne contient que des b, w_2 contient plus de b que de a
- si u_2 contient à la fois des a et des b, alors w_2 contient des b (du premier u_2) qui précèdent des a (du deuxième u_2).

C'est absurde.

Application 2 *Le langage des parenthèses bien formées n'est pas reconnaissable par automates. Le langage des expressions algébriques n'est pas reconnaissable par automates.*

Lemme 1.2 *Soit L un langage reconnaissable par automates. Alors il existe un entier $N > 0$ tel que pour tout mot $w = w_1 w_2 w_3 \in L$ avec $|w_2| \geq N$, il existe trois mots $u_1, u_2, u_3 \in A^*$ tels que $w =_2 u_1 u_2 u_3$, $u_2 \neq \varepsilon$ et $w_1 u_1 u_2^* u_3 w_3 \subset L$.*

Démonstration : Soit $\mathcal{A}$ un automate déterministe reconnaissant L, soit N le nombre de ses états, q_{-1} son état initial, δ sa fonction de transition, δ^* son extension à A^*. Soit $q_0 = \delta^*(q_{-1}, w_1)$, $w_2 = c_1 \ldots c_p \in L$ avec $p \geq N$. Définissons l'état q_i par $q_i = \delta(q_{i-1}, c_i)$ si bien que $\delta^*(q_p, w_3)$ est un état final de l'automate. Comme $p \geq N$ les états $q_0, \ldots, q_p$ ne sont pas distincts, donc il existe $i < j$ tels que $q_i = q_j$. Posons $u_1 = c_1 \ldots c_i$, $u_2 = c_{i+1} \ldots c_j \neq \varepsilon$ et $u_3 = c_{j+1} \ldots c_p$. On a $q_i = q_j = \delta^*(q_i, u_2)$, si bien que l'on a encore, pour tout $n \in \mathbf{N}$, $q_i = q_j = \delta^*(q_i, u_2^n)$, d'où $q_p = \delta^*(q_0, u_1 u_2^n u_3)$, et donc

$$\delta^*(q_{-1}, w_1 u_1 u_2^n u_3 w_3) = \delta^*(q_{-1}, w_1 u_1 u_2 u_3 w_3) = \delta^*(q_{-1}, w)$$

est un état final, ce qui montre que le mot $w_1 u_1 u_2^n u_3 w_3$ est encore reconnu par $\mathcal{A}$ donc appartient à L.

Application 3 *Soit* $L = \{w \in \{a,b\}^* \mid |w|_a = |w|_b\}$ *(langage des mots contenant autant de a que de b). Alors L n'est pas reconnaissable par automate.*

Démonstration Supposons que L est reconnu par un automate et soit N comme dans le lemme. Soit $w = a^N b^N \in L$. Posons $w_1 = \varepsilon$, $w_2 = a^N$ et $w_3 = b^N$. On peut alors écrire $a^N = w_2 = u_1 u_2 u_3$ avec $u_2 \neq \varepsilon$ et $\forall n \in \mathbf{N},\ w_1 u_1 u_2^n u_3 w_3 \in L$. Mais ceci est absurde, car si $u_2 = a^p, p \geq 1$, on a $w_1 u_1 u_2^2 u_3 w_3 = a^{N+p} b^N$ qui n'appartient pas à L.

Application 4 *Soit L le langage des palindromes (mots égaux à leur miroir) sur l'alphabet* $\{a,b\}$. *Alors L n'est pas reconnaissable par automate.*

Démonstration Supposons que L est reconnu par un automate et soit N comme dans le lemme. Soit $w = a^N b a^N \in L$. Posons $w_1 = \varepsilon$, $w_2 = a^N$ et $w_3 = b a^N$. On peut alors écrire $a^N = w_2 = u_1 u_2 u_3$ avec $u_2 \neq \varepsilon$ et $\forall n \in \mathbf{N},\ w_1 u_1 u_2^n u_3 w_3 \in L$. Mais ceci est absurde, car si $u_2 = a^p, p \geq 1$, on a $w_1 u_1 u_2^2 u_3 w_3 = a^{N+p} b a^N$ qui n'appartient pas à L.

Autres exemples de langages non reconnaissables : $L = \{w \mid |w|_a < |w_b|\}$ (prendre $w = a^N b^{N+1}$), $L = \{uu \mid u \in A^*\}$ (prendre $w = a^N b a^N b$), $L = \{a^n \mid n \text{ est premier}\}$ (prendre $w = a^n$ avec n premier supérieur à N et considérer $u_1 u_2^{n+1} u_3$).