Théorie des langages formels (corrigé rapide)

Un monoïde est un ensemble M muni d'une loi de composition . (une application de M^2 dans M) satisfaisant les propriétés suivantes:

- $\forall x,y,z \in M \quad x.(y.z) = (x.y).z$ (. associative)
- $\exists \epsilon \in M \quad \forall x \in M \quad x.\epsilon = \epsilon.x = x$ (M possède un élément neutre ϵ)

On remarque qu'un monoïde ne peut avoir qu'un élément neutre. S'il y en avait deux, disons 1 et ϵ on aurait $1\epsilon = 1$ (car ϵ est neutre) et $1\epsilon = \epsilon$ (car 1 est neutre), et par suite on aurait $1 = \epsilon$.

Autre remarque: en ajoutant l'existence, pour tout élément, d'un inverse, on obtient la définition d'un groupe.

Attention: les monoïdes ne sont en général pas commutatifs.

Un exemple de monoïde (qui n'est pas un groupe et qui n'est pas commutatif): les matrices $n \times n$ munies du produit matriciel, l'identité étant l'élément neutre.

Un sous-monoïde est simplement un sous-ensemble d'un monoïde qui est lui même un monoïde et qui a le même élément neutre[1]

Pour vérifier qu'une partie $M' \subset M$ est un sous monoïde d'un monoïde $(M,.,\epsilon)$ il suffit (exercice: pourquoi?) de vérifier que M' est clos pour . et contient ϵ:

- $\forall x,y \in M' \quad x.y \in M'$
- $\epsilon \in M$

En informatique le monoïde considéré sera le plus souvent le monoïde libre V^* sur un alphabet V, les éléments de celui-ci étant les suites finies d'éléments de V, l'opération la concaténation et l'élément neutre la suite vide. Dans ce contexte on dira plutôt mot ou chaîne que suite finie.[2]

1 Sous-monoïdes de $\{a_1, \ldots ,a_s\}^*$

A Soient $w,w' \in A$; il existe $p,n \in \mathbb{N}$ tels que $|w| = 2p$ et $|w'| = 2n$; par suite $|ww'| = 2p + 2n = 2(p + n)$, et donc $ww' \in A$.
$|\epsilon| = 0 = 2 \times 0$ donc $\epsilon \in A$.
A est donc un sous monoïde de V^*.

B On a $a_1 \in B$ et $a_2 \in B$, et $a_1a_2 \notin B$, B n'est donc pas un (sous)monoïde.
Autre argument: $|\epsilon| = 0$ donc $\epsilon \notin B$, B n'est pas un sous monoïde de V^*.

C Soient $w,w' \in C$; il existe $p,n \in \mathbb{N}$ tels que $w = (a_1a_2)^p$ et $w = (a_1a_2)^n$; donc $ww' = (a_1a_2)^p(a_1a_2)^n = (a_1a_2)^{n+p} \in C$
$\epsilon = (a_1a_2)^0 \in C$
C est donc un sous monoïde de V^*

D $a_1a_2 \in D$ or $a_1a_2a_1a_2 \notin D$, D n'est pas un sous monoïde de V^*.

1. Par exemple $\{0\}$ est un sous ensemble de $(N, \times ,1)$ qui est stable par $\times$, mais qui n'est pas un sous monoïde de $(N, \times ,1)$ car son élément neutre est 0 et non 1.

2. Ce monoïde est dit libre parce que quelque soit une application ϕ de V dans un monoïde $(M, \circ ,1)$ il existe un unique morphisme $\tilde{\phi}$ de monoïdes ϕ de $(V^*,.,\epsilon)$ dans $(M, \circ ,1)$ telle que $\phi(x) = \tilde{\phi}(x)$ — le deuxième x étant la chaîne constituée d'un élément x qui appartient à V^*.

E Soient w,w' deux chaînes contenant a_1 et a_2 en nombre identique. w contient n fois a_1 et n fois a_2, et w' contient p fois a_1 et p fois a_2. Par conséquent ww' contient $n+p$ fois a_1 et $n+p$ fois a_2 et donc $ww' \in E$.
De plus ϵ contient 0 fois a_1 et 0 fois a_2.
Donc E est un sous monoïde de V^*.

2 Génération

Etant donné un ensemble de non terminaux N et un ensemble de terminaux V avec $V \cap N = \emptyset$ une grammaire est la donnée d'un nombre fini de règles de production, qui a un mot (non vide) w de $(V \cup N)^+$ (on peut se ramener à: contenant au moins un non terminal) associe un mot w' de $(V \cup N)^*$, ce qui est noté $w \to w'$

Si la taille d'un membre droit est toujours supérieure à celle du membre gauche correspondant la grammaire est dite contextuelle ou de type 1. La règle $S \to \epsilon$ peut être admise si S ne figure jamais dans un membre droit.

Si les membres gauches sont toujours réduit à un non terminal, la grammaire est dite non-contextuelle, hors-contexte ou algébrique ou de type 2. La règle $S \to \epsilon$ peut être admise si S ne figure jamais dans un membre droit.

Si la grammaire est de type 2 et que de plus chaque membre droit contient au plus un non terminal toujours situé à la fin du mot, la grammaire est dite régulière ou rationnelle ou de type 3. [Il en est de même si le non terminal est toujours situé en tête de chaque membre droit.] La règle $S \to \epsilon$ peut être admise si S ne figure jamais dans un membre droit.

Etant donnée une grammaire contenant S dans un membre droit, on obtient une grammaire produisant le même langage plus ϵ en prenant un nouveau symbole de départ $\mathcal{S}$ et en ajoutant les règles: $\mathcal{S} \to S$ et $\mathcal{S} \to \epsilon$.

Pour toute les grammaires de type 2 et 3 les dérivations peuvent être représentées par des arbres de dérivation. On peut représenter un arbre de dérivation par une liste de listes, en indiquant en indice de la parenthèse ouvrante le non terminal étiquetant la racine de ce sous arbre.

Le langage engendré est constitué des mots de *terminaux* obtenus par réécriture successives à partir de S en utilisant les règles de production.

G_1 $[_S\ a\ [_S\ c]\ b\ [_S\ a\ [_S\ c]\ b\ [_S\ c]\ a]\ a]$ qui produit la chaîne $a\ c\ b\ a\ c\ b\ c\ a\ a$ de longueur 9.
$[_S\ a\ [_S\ a\ [_S\ c]\ b\ [_S\ c]\ a]\ b\ [_S\ a\ [_S\ c]\ b\ [_S\ c]\ a]\ a]$ qui produit la chaîne $a\ a\ c\ b\ c\ a\ b\ a\ c\ b\ c\ a\ a$ de longueur 13.
Il s'agit d'une grammaire de type 2: chaque règle produit une séquence de terminaux/non-terminaux à partir d'**un** non terminal.
Par contre ce n'est pas une grammaire de type 3: il y a plusieurs non terminaux à droite (2 fois S).

G_2

$$\begin{array}{lcl}
S & \rightarrow & B\ C\ a\ C\ b\ b\ \underline{A} \\
 & \rightarrow & B\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ \underline{A} \\
 & \rightarrow & B\ C\ a\ \underline{C\ b\ b}\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ A \\
 & \rightarrow & B\ C\ a\ d\ a\ \underline{C\ a}\ C\ b\ b\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ A \\
 & \rightarrow & B\ C\ a\ d\ a\ b\ a\ \underline{C\ b\ b}\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ A \\
 & \rightarrow & \underline{B}\ C\ a\ d\ a\ b\ a\ d\ a\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ A \\
 & \rightarrow & ch\ a\ \underline{C\ a}\ d\ a\ b\ a\ d\ a\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\ A \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ C\ a\ C\ b\ b\ C\ a\ C\ b\ b\underline{A} \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ C\ a\ C\ b\ b\ C\ a\ \underline{C\ b\ b} \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ C\ a\ \underline{C\ b\ b}\ C\ a\ d\ a \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ \underline{C\ a}\ d\ a\ C\ a\ d\ a \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ b\ a\ d\ a\ \underline{C\ a}\ d\ a \\
 & \rightarrow & ch\ a\ b\ a\ d\ a\ b\ a\ d\ a\ b\ a\ d\ a\ b\ a\ d\ a
\end{array}$$

Longueur 18 caractères (ch est un caractère).

$$\begin{array}{lcl}
S & \rightarrow & B\ C\ a\ C\ b\ b\ \underline{A} \\
 & \rightarrow & B\ C\ a\ C\ b\ b \\
 & \rightarrow & B\ C\ a\ \underline{C\ b\ b} \\
 & \rightarrow & B\ \underline{C\ a}\ d\ a \\
 & \rightarrow & \underline{B}\ b\ a\ d\ a \\
 & \rightarrow & ch\ a\ b\ a\ d\ a
\end{array}$$

Longueur 6 caractères.

C'est une grammaire de type 0: en effet on a une règle qui réécrit tout un mot et dont la taille diminue.

G_3 $[_S\ cha\ [_A\ bada\ [_A\ bada]]]$ chaîne $cha\ bada\ bada$ longueur 3 (cha ou $bada$ est un seul caractère).

$[_S\ cha\ [_A\ bada\ [_A\ bada\ [_A\ bada]]]]$ chaîne $cha\ bada\ bada\ bada$ de longueur 4.

C'est une grammaire régulière (linéaire droite): on a un seul non terminal dans chaque membre droit, et qui apparaît toujours le plus à droite.

G_4 Cette grammaire ne produit pas de chaîne de terminaux. En effet, au départ S est un non terminal, et chaque fois que l'on réécrit un non terminal, il produit un non terminal: toute chaîne produite contient donc au moins un non terminal. Le langage engendré est vide.

C'est une grammaire de type 2: les membres gauche sont toujours réduits à un non terminal. Ce n'est pas une grammaire de type 3 car bien qu'il n'y ait qu'un seul non terminal en partie droite, il apparaît tantôt à droite tantôt à gauche d'un membre droit.

G_5

$$\begin{array}{lcl}
S & \rightarrow & \underline{A}\ B \\
 & \rightarrow & a\ A\ \underline{B} \\
 & \rightarrow & a\ \underline{A}\ B\ b \\
 & \rightarrow & a\ a\ \underline{A\ B}\ b \\
 & \rightarrow & a\ a\ c\ c\ b
\end{array}$$

Chaîne de longueur 5.

$$\begin{array}{rcl} S & \rightarrow & \underline{A}\ B \\ & \rightarrow & a\ A\ \underline{B} \\ & \rightarrow & a\ \underline{A\ B}\ b \\ & \rightarrow & a\ c\ c\ b \end{array}$$

Chaîne de longueur 4.

C'est une grammaire de type 1. Ce n'est pas une grammaire de type 2 car des mots se réécrivent. Elle est de type 1 car le membre droit à toujours une taille supérieure ou égale à celle du membre gauche correspondant.