3.1 Langages formels

3.1.1 Monoïde

Déf. 1 (Alphabet)

Un ***alphabet*** X est un ensemble fini de symboles (lettres). La *taille* de l'alphabet est le nombre de symboles. On parle aussi de ***vocabulaire***.

Déf. 2 (Mot)

Un *mot* sur l'alphabet X est une suite finie de lettres de X.
Les symboles qui composent le mot u peuvent donc être caractérisés par leur *position* dans la suite : on parle d'*occurrence*.
Le ***mot vide***, que l'on notera ε (ou $\mathbb{1}_X$) est la suite qui ne comporte aucun élément.
La longueur du mot u est notée $|u|$, c'est le nombre d'occurrences dans u. $|\varepsilon| = 0$.

Formellement, on définit $[p] = (1, 2, 3, 4, ..., p)$ (suite entière ordonnée).
Alors le mot u sur l'alphabet X est une fonction

$$u : [p] \longrightarrow X$$

qui à chaque entier entre 1 et p associe un symbole.

Déf. 3 (Sous-mot, facteur)

w est un *sous-mot* de u si w est une sous-suite de lettres de u.
w est un *facteur* de u si w est un sous-mot de u, dont les lettres sont adjascentes dans u.
NB : L'ordre est conservé.

Déf. 4 (Concaténation)

Soient $[p] \xrightarrow{u} X$, $[q] \xrightarrow{w} X$. On définit la concaténation de u et w, notée uw (quelquefois $u.w$ ou $u\hat{}w$) :

$$uw : \quad [p+q] \longrightarrow X$$
$$uw_i = \begin{cases} u_i & \text{pour} \quad i \in [1, p] \\ w_{i-p} & \text{pour} \quad i \in [p+1, p+q] \end{cases}$$

Définition informelle : uw est le mot formé de tous les symboles de u suivis immédiatement de tous les symboles de w.

On démontre facilement que l'opération de concaténation ainsi définie a les propriétés suivantes :

- La concaténation est non commutative (en général)
- La concaténation est associative : $u(vw) = (uv)w = uvw$
- La concaténation admet un élément neutre, le mot vide, noté $\emptyset$ ou ε ou 1_X ou $\mathbb{1}_X$.

Cette opération munit donc l'ensemble des mots sur un alphabet X ($(X, .)$, noté X^*) d'une structure de *monoïde*. De plus, ce monoïde est *libre* car les éléments de l'alphabet, qu'on appelle la *base*, sont indépendant les uns des autres.

3.1.2 Langage formel

Déf. 5 (Langage)

Un langage sur un alphabet X est une partie de X^* (c'est-à-dire un sous-ensemble).
$L \subseteq X^*$. $L \in \mathcal{P}(X^*)$ $\left(= 2^{X^*}\right)$

Problèmes sur un langage
- Caractérisation
- Reconnaissance
- Classification

Opération ensemblistes courantes sur les langages :

$$\begin{aligned} L_1 \cup L_2 &= \{u \,/\, u \in L_1 \text{ ou } u \in L_2\} \\ L_1 \cap L_2 &= \{u \,/\, u \in L_1 \text{ et } u \in L_2\} \\ L_1 \setminus L_2 &= \{u \,/\, u \in L_1 \text{ et } u \notin L_2\} \end{aligned}$$

Opération induite par l'opération de concaténation, appelée produit :

$$L_1.L_2 = \{uv \,/\, u \in L_1 \text{ et } v \in L_2\}$$

On démontre facilement que l'opération de produit ainsi définie a les propriétés suivantes :

- Associative
- Admet un élément neutre, l'ensemble $\{\epsilon\}$.
- Admet un élément absorbant, l'ensemble vide $\emptyset$.
- Distributive par rapport à $\cup$.
- Non distributive par rapport à $\cap$.

Déf. 6 (Étoile)

Par analogie, on notera A^2 le produit A.A pour un langage A.
En généralisant, on peut proposer la notation

$$\begin{aligned} A^0 &= \{\mathbb{1}_X\} \\ A^1 &= A \\ A^{i+1} &= A.A^i \end{aligned}$$

d'où

$$A^n = \{a_1 ... a_n \,/\, a_i \in A\}$$

et définir ainsi l'*étoile* de A :

$$A^* = \bigcup_{n \geq 0} A^n$$

Remarque : La notation est cohérente : $X^* = \bigcup_{n\geq 0} X^n$ où $X^n = \{\text{mots de longueur } n\}$.

3.2 Grammaires formelles

3.2.1 Définition

Une **grammaire formelle** est un quadruplet $\langle X, V, S, P\rangle$ où
- X est l'alphabet (du langage engendré)
- V est un alphabet disjoint de X dit « *non terminal* »
- $S \in V$ est un élément distingué de V, appelé *axiome*
- P est un ensemble de « *règles (de production)* », c'est-à-dire une partie finie du produit cartésien $(X \cup V)^* V (X \cup V)^* \times (X \cup V)^*$.

3.2.2 Dérivation

Soient $\mathcal{G} = \langle X, V, S, P\rangle$ une grammaire, $(f, g) \in (X \cup V)^*$, r une règle de production de P, de la forme $r : A \longrightarrow u$ $(A \in (X \cup V)^* V (X \cup V)^* \times (X \cup V)^*)$.

- f se **réécrit** (ou **dérive immédiatement**) en g avec la règle r (notation $f \xrightarrow{r} g$) ssi $\exists v, w$ t.q. $f = vAw$ et $g = vuw$
- f se **réécrit** (ou **dérive**) en g dans la grammaire $\mathcal{G}$ (notation $f \xrightarrow{\mathcal{G}} g$) ssi $\exists r \in P$ t.q. $f \xrightarrow{r} g$.
- $f \xrightarrow{\mathcal{G}*} g$ si $f = g$, ou
  $\exists f_1 = f, f_2, ..., f_n = g$ t.q. $f_{i-1} \longrightarrow f_i$

On note $L_{\mathcal{G}}(f)$ l'ensemble des terminaux engendrés par f dans la grammaire $\mathcal{G}$.
$L_{\mathcal{G}}(f) = \{g \in X^* / f \xrightarrow{\mathcal{G}*} g\}$
Par convention, on notera $L_{\mathcal{G}}$ le langage $L_{\mathcal{G}}(S)$.

3.2.3 Arbre de dérivation

3.2.4 Classes de grammaire

La fameuse « hiérarchie de Chomsky » :

type 0 Aucune restriction sur $P \subset (X \cup V)^* V (X \cup V)^* \times (X \cup V)^*$.

type 1 (*grammaires contextuelles*, *context-sensitive*) Tout élément de P est de la forme $(u_1 S u_2, u_1 m u_2)$, où u_1 et $u_2 \in (X \cup V)^*$, $S \in V$ et $m \in (X \cup V)^+$.

type 2 (*grammaires algébriques*, *context-free*)
Tout élément de P est de la forme (S, m), où $S \in V$ et $m \in (X \cup V)^*$.

type 3 (*grammaires régulières*)
Tout élément de P est de la forme (S, m), où $S \in V$ et $m \in X.V \cup X \cup \{\varepsilon\}$.