Chapter 2

Mots et langages

2.1 Introduction

Dans une machine actuelle, on n'a notre disposition que des 0 et des 1 pour stocker l'information. Il s'agit donc d'arriver à représenter nos données uniquement avec des suites de 0 et de 1.

Par exemple, une représentation très utilisée pour les caractères est le code ASCII : à chaque caractère on associe une séquence de 8 bits, 1 octet, qui le représente de façon unique. Une chaine de caractère, en C, est donc une suite de caractères, qui eux-mêmes sont des groupes de 8 bits. Un caractères spécial est utilisé pour désigner la fin de la chaîne.

Si on prend comme unité le bit, un caractère est une suite de bits; si on prend comme unité le caractère, une chaîne est une suite de caractères. Il semble donc important de développer des connaissances sur les "suites de lettres", puisqu'on aura toujours à s'en servir.

La notion de code permet des représentations un peu plus compliquées que le code ASCII, mais peut-être plus efficaces (plus compactes par exemple). Il s'agit de représenter nos objets, toujours par des suites de bits, mais pas nécessairement de longueur fixe. On veut cependant que le découpage soit sans ambiguité afin qu'une même suite de bits ne représente pas 2 objets différents.

2.2 Mots

2.2.1 Définitions

Définition 4 *Un alphabet est un ensemble fini dont les éléments sont appelés des lettres.*

Exemple : $B = \{0, 1\}$ l'alphabet binaire, $A = \{a, b, c\}$, $C = \{a \cdots, z\}$. On peut considérer n'importe quel ensemble fini, par exemple $A = \{hello, word\}$ est un alphabet à deux lettres.

Définition 5 *Un mot sur l'alphabet A est une suite finie d'élément de A. On note A^* l'ensemble des mots sur A. La suite peut ne contenir aucun élément, il s'agit dans ce cas du mot vide, noté ϵ, qui ne contient aucune lettre.*

Exemple : 0011010, 0110 sont des mots sur l'alphabet $\{0, 1\}$. *abbab* est un mot sur $\{a, b\}$. *hello word word hello* est un mot sur $A = \{hello, word\}$.

Mathématiquement, on a

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

avec la convention $A^0 = \{\varepsilon\}$. On représente les n-uplets sans les virgules et les parenthèses.

Deux mots sont égaux s'ils ont les mêmes lettres dans le même ordre.

2.2.2 Concaténation

Définition 6 *Soit $u = u_1u_2\cdots u_n$ un mot de taille n sur l'alphabet A, les u_i étant les lettres le composant, et soit $v = v_1v_2\cdots v_m$ un autre mot, de taille m sur l'alphabet A. On note $u\cdot v$ la concaténation de u et de v, qui est le mot :*

$$u \cdot v = u_1 \cdots u_n v_1 \cdots v_m$$

Proposition 7 *La concaténation sur A^* est associative et admet pour élément neutre le mot vide ϵ. Elle n'est pas commutative.*

Définition 8 *Un monoïde est ensemble muni d'une loi interne associative et possédant un élément neutre. A^* muni de la concaténation est un monoïde.*

Définition 9 *Soient $(X,+)$ et $(Y,*)$ deux monoïdes de neutres respectifs ε_X et ε_Y. Une application f de X dans Y est un* morphisme de monoïde *(on dit juste* morphisme *quand il n'y a pas d'ambiguïté), quand*

- $f(\varepsilon_X) = \varepsilon_Y$
- *pour tout $x_1, x_2 \in X$, $f(x_1 + x_2) = f(x_1) * f(x_2)$*

2.2.3 Longueur

Définition 10 *La taille, ou encore la longueur, d'un mot $u \in A^*$ est son nombre de lettres. On la note $|u|$. dans le mot u. Le mot vide est de taille* 0.

Notation 11 *Soit $a \in A$, le nombre de a présents dans un mot $u \in A^*$ est noté $|u|_a$.*

Proposition 12 *Pour tout $u, v \in A^*$ on a :*

- $|uv| = |u| + |v|$
- *pour tout $a \in A$, $|uv|_a = |u|_a + |v|_a$*

2.2.4 Découpage des mots

Définition 13 *On dit qu'un mot v est un* préfixe *d'un mot u sur A, s'il existe un mot $w \in A^*$ tel que $u = vw$. On dit qu'un mot v est un* suffixe *d'un mot u sur A, s'il existe un mot $w \in A^*$ tel que $u = wv$. On dit qu'un mot v est un* facteur *d'un mot u sur A, s'il existe deux mots $w, w' \in A^*$ tel que $u = wvw'$.*

Définition 14 *Un préfixe (resp. suffixe, facteur) d'un mot u est un* prefixe strict *(resp.* suffixe strict, facteur strict*) s'il est différent de u.*

Définition 15 *Soit $u = u_1\cdots u_n$ un mot de A^*. Un sous-mot v de u est un mot v de longueur m tel qu'il existe une injection strictement croissante ϕ de $\{1,\cdots,m\}$ dans $\{1,\cdots,n\}$ tel que :*

$$v = u_{\phi(1)}u_{\phi(2)}\cdots u_{\phi(m)}$$

2.2.5 Opérations sur les mots

Définition 16 *Le miroir d'un mot $u = u_1 \cdots u_n$ sur A est le mot $\overline{u} = u_n \cdots u_1$ obtenu lisant les lettres dans l'autre sens.*

Proposition 17 *Pour tout $u, v \in A^*$, $\overline{u \cdot v} = \overline{v} \cdot \overline{u}$*

2.2.6 Ordre sur les mots

On suppose que l'alphabet A est totalement ordonné.

Définition 18 *Soient u et v deux mots de A^* on dit que u est plus petit que v pour l'*ordre lexicographique, *noté $u \leq v$ ou encore $u \leq_{lex} v$ quand :*

- *ou bien u est préfixe de v,*
- *ou bien il existe $a < b$ dans A, il existe des mots w, u', v' dans A^*, tels que $u = wau'$ et $v = wbv'$.*

Donner des exemples.

Définition 19 *Soient u et v deux mots de A^* on dit que u est plus petit que v pour l'*ordre militaire, *noté $u \leq_{mil} v$, quand*

- *ou bien $|u| < |v|$,*
- *ou bien $|u| = |v|$ et $u \leq_{lex} v$.*