INTRODUCTION AUX LANGAGES

À la fin de ce chapitre, je sais :

- ☞ définir les concepts d'alphabet, de mot, de mot vide et de langage
- ☞ expliquer les concepts de suffixe, de préfixe, de facteur et de sous-mot
- ☞ expliquer le résultat du lemme de Levi

L'informatique est la construction de l'information par le calcul. Force est de constater que le seul outil conceptuel, universel et pratique pour construire et manipuler l'information est le langage : un langage est un moyen de communiquer, stocker et transformer de l'information. Le calcul de l'information par un ordinateur au moyen d'un ou plusieurs langages peut créer un sens ou pas, tout comme l'interprétation d'un texte par un humain.

C'est pourquoi la théorie des langages est un des principaux fondements de l'informatique. Qui dit langage dit alphabet, mots, préfixes, suffixes mais aussi ensembles de mots, agrégation de mots... Comment définir clairement ces concepts afin de pouvoir les calculer ? C'est la question qui guide ce chapitre.

A Alphabets

■ **Définition 1 — Ensemble.** Un ensemble est une collection de concepts qu'on appelle éléments. L'ensemble vide est noté $\emptyset$.

■ **Définition 2 — Cardinal d'un ensemble fini.** Le cardinal d'un ensemble fini E est son nombre d'éléments. On le note $|E|$.

■ **Définition 3 — Alphabet.** Un alphabet est un ensemble Σ de lettres (ou symboles) non vide.

■ **Définition 4 — Longueur d'un alphabet.** La longueur d'un alphabet est le nombre de lettres de celui-ci, c'est-à-dire $|\Sigma|$.

■ **Exemple 1 — Alphabet latin commun.** L'alphabet latin commun

$$\Sigma = \{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Z\}$$

a une longueur de 26.

■ **Exemple 2 — Héxadécimal.** L'alphabet héxadécimal

$$\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$$

a une longueur de 16.

■ **Exemple 3 — ASCII.** L'alphabet ASCII est constitué des nombres entiers de 0 à 127 et représente les caractères nécessaire à l'écriture de l'américain, y compris les caractères de contrôle nécessaires à la pagination. Il possède une longueur de 128.

B Mots

Un mot d'un alphabet Σ est une séquence de lettres de Σ. Un mot peut être une séquence vide notée ϵ, car $\emptyset \subset \Sigma$ par définition d'un ensemble. On peut donner plusieurs définition d'un mot.

■ **Définition 5 — Mot (comme application).** Un mot de longueur $n \in \mathbb{N}^\star$ est une application de $[\![1, n]\!] \longrightarrow \Sigma$. À chaque position dans le mot correspond une lettre de l'alphabet.

■ **Définition 6 — Mot vide ϵ.** Le mot vide ϵ est l'application de l'ensemble vide dans Σ.

R Le mot vide ϵ est un mot ne comportant aucun symbole. Dans le contexte des mots, le mot vide est l'élément neutre de la concaténation de mots. On peut comparer son rôle au 1 pour la multiplication des entiers naturels $\mathbb{N}$.

■ **Définition 7 — Longueur d'un mot.** La longueur d'un mot w est le nombre de lettres qui composent sa séquence. On note souvent cette longueur $|\, w \,|$.

■ **Définition 8 — Ensemble de mots possibles.** On note Σ^* l'ensemble des mots possibles crées à partir d'un alphabet Σ.

■ **Définition 9 — Ensemble de mots de longueur n.** On note Σ^n l'ensemble de tous les mots de longueur n crées à partir d'un alphabet Σ.

■ **Définition 10 — Concaténation de mots.** Soit $v, w \in \Sigma^*$. On appelle concaténation de v

et w l'opération $\circ$ notée $vw = v \circ w$ qui est obtenue par agrégation du mot w à la suite du mot v.

R Dans les notations suivantes, on omettra d'écrire le symbole $\circ$ entre les éléments, la concaténation étant juste une agrégation de symboles.

Théorème 1 — La concaténation est une loi de composition interne sur un ensemble Σ^* et ϵ en est l'élément neutre. .

Démonstration. Soit $v, w \in \Sigma^*$. On observe que $vw \in \Sigma^*$. Donc c'est une application de $\Sigma^* \times \Sigma^* \longrightarrow \Sigma^*$. De plus, on peut observer que $v\epsilon = \epsilon v = v$. Donc ϵ est l'élément neutre de cette loi. ■

■ **Définition 11 — Monoïde.** Un ensemble E muni d'une loi de composition interne associative et d'un élément neutre e est nommée monoïde $(E, \star)$.

R On peut facilement montrer que la concaténation de mots est une loi de composition interne associative. C'est pourquoi, $(\Sigma^*, \circ)$ est un **monoïde**.

Il faut bien remarquer cependant qu'il n'existe pas a priori de concept d'inverse dans un monoïde, c'est-à-dire il **n'existe pas** de mots v et w tels que $vw = \epsilon$.

Toutefois, on peut simplifier par la gauche ou la droite :

$$vw = vx \Longrightarrow w = x \tag{1}$$
$$wv = xv \Longrightarrow w = x \tag{2}$$

R La longueur d'un mot est un morphisme de monoïde car $|vw| = |v| + |w|$.

■ **Définition 12 — Ensemble de mots non vides.** On note Σ^+ l'ensemble des mots non vides crées à partir d'un alphabet Σ. C'est le plus petit ensemble tel que :

$$\forall a \in \Sigma, a \in \Sigma^+ \tag{3}$$
$$\forall w \in \Sigma^+, \forall a \in \Sigma, wa \in \Sigma^+ \tag{4}$$

On a $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$.

C Mots définis inductivement

■ **Définition 13 — Mot (inductivement par la droite).** Soit Σ un alphabet. Alors on définit un mot de manière inductive par la droite ainsi :

Base ϵ est un mot sur Σ,

Règle de construction si w est un mot sur Σ et a une lettre de Σ, alors $w.a$ est un mot sur Σ.

où l'opération . est l'ajout d'une lettre à droite à un mot.

■ **Définition 14 — Mot (inductivement par la gauche)).** Soit Σ un alphabet. Alors on définit un mot de manière inductive par la droite ainsi :

Base ϵ est un mot sur Σ,

Règle de construction si w est un mot sur Σ et a une lettre de Σ, alors $a.w$ est un mot sur Σ.

où l'opération . est l'ajout d'une lettre à gauche à un mot.

■ **Définition 15 — Concaténation de mots (définie inductivement sur la première opérande).** On définit l'opérateur concaténation de mots $\circ$ par

$$\forall w \in \Sigma^*, \epsilon \circ w = w \text{ (Base)} \tag{5}$$

$$\forall v, w \in \Sigma^*, \forall a \in \Sigma, (a.v) \circ w = a.(v \circ w) \text{ (Règle de construction)} \tag{6}$$

où l'opération . est l'ajout d'une lettre à gauche à un mot.

■ **Définition 16 — Concaténation de mots (définie inductivement sur la seconde opérande).** On définit l'opérateur concaténation de mots $\circ$ par

$$\forall w \in \Sigma^*, w \circ \epsilon = w \tag{7}$$

$$\forall v, w \in \Sigma^*, \forall a \in \Sigma, v \circ (w.a) = (v \circ w).a \tag{8}$$

où l'opération . est l'ajout d'une lettre à droite à un mot.

Théorème 2 — ϵ est l'élément neutre de la concaténation.

Démonstration. on procède par induction sur la première opérande.

- Cas de base : pour $w = \epsilon$, on a $\epsilon \circ \epsilon = \epsilon$.
- Pas d'induction : soit $w \in \Sigma^*$. On suppose que ϵ est l'élément neutre pour ce mot : $w \circ \epsilon = \epsilon \circ w = w$. Considérons maintenant un élément a de l'alphabet Σ pour créer un mot plus long à partir de w. Par construction on a :

  $$(a.w) \circ \epsilon = a.(w \circ \epsilon)$$

  En utilisant l'hypothèse d'induction, on en déduit que $(a.w) \circ \epsilon = a.w$.
  ϵ est donc toujours l'élément neutre. On procède de même avec la définition sur la deuxième opérande.

■

R Une conséquence de ces définitions est qu'on peut confondre les opérateurs . et ∘ dans les notations. C'est ce qui est fait dans la suite de ce cours. On omettra également souvent l'opérateur lorsqu'il n'y a pas d'ambiguïtés.

■ **Définition 17 — Puissances d'un mot.** Les puissances d'un mot sont définies inductivement :

$$w^0 = \epsilon \tag{9}$$

$$w^n = w w^{n-1} \text{ pour } n \in \mathbb{N}^\star \tag{10}$$

D Langages

■ **Définition 18 — Langage.** Un langage sur un alphabet Σ est un ensemble de mots sur Σ.

R Un langage peut être vide, on le note alors $\mathcal{L} = \emptyset$, son cardinal est nul. C'est l'élément neutre de l'union des langages et l'élément absorbant de la concaténation de langages [a]. Il ne faut pas confondre ce langage vide avec le langage qui ne contient que le mot vide $\mathcal{L} = \{\epsilon\}$ dont le cardinal vaut un et qui est l'élément neutre de la concaténation des langages.

a. comme le zéro pour l'addition et la multiplication des entiers

R Σ^*, l'ensemble de tous les mots sur Σ, est également appelé langage universel.

■ **Exemple 4 — Langages courants et concrets.** Voici quelques exemples de langages concrets utilisés couramment :

- le langage des dates : une expression est-elle une date ? Par exemple, les dates 21/11/1943 et 11/21/43 sont-elles admissibles ?
- le langage des emails : utilisé pour détecter la conformité ou les erreurs dans les adresses emails,
- les protocoles réseaux : par exemple le protocole DHCP.

■ **Exemple 5 — Langage des mots de longueur paire.** Soit l'ensemble E de mots sur l'alphabet Σ de longueur paire. On peut définir ce langage en compréhension comme suit :

$$E = \{w \in \Sigma^*, |w| = 0 \bmod 2\}$$

■ **Exemple 6 — Langage des puissances n d'un alphabet.** Soit l'ensemble E de mots sur l'alphabet $\Sigma = \{a, b\}$ qui comportent autant de a que de b. On peut définir ce langage en

compréhension comme suit :

$$E = \{w \in \Sigma^*, \exists n \in \mathbb{N}, w \text{ est une permutation de } (a^n b^n)\}$$

Un langage est un ensemble. On peut donc définir les opérations ensemblistes sur les langages.

Soit deux langages $\mathcal{L}_1$ sur Σ_1 et $\mathcal{L}_2$ sur Σ_2.

■ **Définition 19 — Union de deux langages.** L'union de $\mathcal{L}_1$ et $\mathcal{L}_2$ est le langage défini sur $\Sigma_1 \cup \Sigma_2$ contenant tous les mots de $\mathcal{L}_1$ et de $\mathcal{L}_2$.

$$\mathcal{L}_1 \cup \mathcal{L}_2 = \{w, w \in \mathcal{L}_1 \text{ ou } w \in \mathcal{L}_2\} \tag{11}$$

■ **Définition 20 — Intersection de deux langages.** L'intersection de $\mathcal{L}_1$ et $\mathcal{L}_2$ est le langage défini sur $\Sigma_1 \cap \Sigma_2$ contenant tous les mots à la fois présents dans $\mathcal{L}_1$ et dans $\mathcal{L}_2$.

$$\mathcal{L}_1 \cap \mathcal{L}_2 = \{w, w \in \mathcal{L}_1 \text{ et } w \in \mathcal{L}_2\} \tag{12}$$

■ **Définition 21 — Complémentaire d'un langage.** Le complémentaire d'un langage $\mathcal{L}$ est le langage défini sur Σ qui contient tous les mots non qui ne sont pas dans $\mathcal{L}$.

$$C(\mathcal{L}) = \overline{\mathcal{L}} = \{w, w \in \Sigma^* \text{ et } w \notin \mathcal{L}\} \tag{13}$$

■ **Définition 22 — Différence de deux langages .** La différence de $\mathcal{L}_1$ et $\mathcal{L}_2$ est le langage défini sur Σ_1 contenant tous les mots présents dans $\mathcal{L}_1$ qui ne sont pas dans $\mathcal{L}_2$.

$$\mathcal{L}_1 \setminus \mathcal{L}_2 = \{w, w \in \mathcal{L}_1 \text{ et } w \notin \mathcal{L}_2\} \tag{14}$$

■ **Définition 23 — Produit de deux langages ou concaténation .** Le produit de $\mathcal{L}_1$ et $\mathcal{L}_2$ est le langage défini sur $\Sigma_1 \cup \Sigma_2$ contenant tous les mots formés par une mot de $\mathcal{L}_1$ suivi d'un mot de $\mathcal{L}_2$.

$$\mathcal{L}_1.\mathcal{L}_2 = \{vw, v \in \mathcal{L}_1 \text{ et } w \in \mathcal{L}_2\} \tag{15}$$

■ **Définition 24 — Puissances d'un langage.** Les puissances d'un langage $\mathcal{L}$ sont définies par induction :

$$\mathcal{L}^0 = \{\epsilon\} \tag{16}$$

$$\mathcal{L}^n = \mathcal{L}.\mathcal{L}^{n-1} \text{ pour } n \in \mathbb{N}^\star \tag{17}$$

■ **Définition 25 — Fermeture de Kleene d'un langage.** La fermeture de Kleene d'un langage $\mathcal{L}$ ou étoile de Kleene notée $\mathcal{L}^*$ est l'ensemble des mots formés par un nombre fini de

concaténation de mots de $\mathcal{L}$. Formellement :

$$\mathcal{L}^* = \bigcup_{n \geqslant 0} \mathcal{L}^n \tag{18}$$

La fermeture d'un langage peut également être définie inductivement par :

$$\epsilon \in \mathcal{L}^* \tag{19}$$
$$v \in \mathcal{L}, w \in \mathcal{L}^* \Longrightarrow vw \in \mathcal{L}^* \tag{20}$$
$$v \in \mathcal{L}^*, w \in \mathcal{L} \Longrightarrow vw \in \mathcal{L}^* \tag{21}$$

R Il existe un nombre dénombrable de mots sur un alphabet Σ, c'est à dire qu'on peut les mettre en bijection avec $\mathbb{N}$, il y en a une infinité mais on peut les compter. Néanmoins, le nombre de langages sur Σ n'est pas dénombrable puisqu'il s'agit des parties d'un ensemble dénombrable.

E Préfixes, suffixes, facteurs et sous-mots

■ **Définition 26 — Préfixe.** Soit v et w deux mots sur Σ. v est un préfixe de w et on le note $v \preceq w$ si et seulement s'il existe un mot u sur Σ tel que :

$$vu = w \tag{22}$$

■ **Définition 27 — Suffixe.** Soit v et w deux mots sur Σ. w est un suffixe de v si et seulement s'il existe un mot u sur Σ tel que :

$$uw = v \tag{23}$$

■ **Définition 28 — Facteur.** Soit v et w deux mots sur Σ. v est un facteur de w si et seulement s'il existe deux mots t et u sur Σ tel que :

$$tvu = w \tag{24}$$

Vocabulary 1 — Subword ⇜⇝ Facteur. Attention l'imbroglio n'est pas loin...

■ **Définition 29 — Sous-mot.** Soit $w = a_1 a_2 \dots a_n$ un mot sur $\Sigma = \{a_1, a_2, \dots, a_n\}$ de longueur n. Alors $v = a_{\psi(1)} a_{\psi(2)} \dots a_{\psi(p)}$ est un sous-mot de w de longueur p si et seulement s'il ψ : $[\![1, p]\!] \longrightarrow [\![1, n]\!]$ est une application strictement croissante.

R Cette définition implique que l'ordre d'apparition des lettres dans un sous-mot est préservé par rapport à l'ordre de lettres du mot.

Vocabulary 2 — Scattered Subword ⇜⇝ Sous-mot...

■ **Exemple 7 — Illustrations des concepts précédents.** Prenons par exemple le mot le plus long de la langue française, *anticonstitutionnellement*. Alors

- *anti* est un préfixe, tout comme antico mais uniquement pour les informaticiens, pas les linguistes...
- *ment* est un suffixe,
- *constitution* est un facteur,
- *colle* est un sous-mot.

Théorème 3 — Relations d'ordre partielles. Les relations «être préfixe de», «être suffixe de» et «être facteur de » sont des relations d'ordre partiel.

Démonstration. Il suffit de montrer que ces relations sont réflexives, transitives et antisymétriques. C'est un exercice à faire. ■

■ **Définition 30 — Ordre lexicographique.** Soit v et w deux mots sur un alphabet Σ sur lequel on dispose d'un ordre total $\leqslant_\Sigma$. Alors on peut définir [a] l'ordre lexicographique $\leqslant_l$ entre deux mots $v \leqslant_l w$ par :

- v est un préfixe de w
- ou bien $\exists t, v', w' \in \Sigma^+, v = tv', w = tw'$ et la première lettre de v' précède celle de w' dans l'alphabet, c'est-à-dire au sens de $\leqslant_\Sigma$.

a. Il était temps après 18 ans d'école!

■ **Définition 31 — Distance entre deux mots.** Supposons que l'on dispose d'une fonction λ capable de calculer le plus long préfixe, le plus long suffixe ou le plus long facteur commun entre deux mots. Alors on peut définir une distance entre deux mots v et w par :

$$d(v,w) = |vw| - 2\lambda(v,w) \tag{25}$$

■ **Définition 32 — Fermeture d'un langage par préfixe.** La fermeture par préfixe d'un langage $\mathcal{L}$ notée $\mathrm{Pref}(\mathcal{L})$ est le langage formé par l'ensemble des préfixes des mots de $\mathcal{L}$.

$$\mathrm{Pref}(\mathcal{L}) = \left\{w \in \Sigma^*, \exists v \in \Sigma^*, wv \in \mathcal{L}\right\} \tag{26}$$

■ **Définition 33 — Fermeture d'un langage par suffixe.** La fermeture par suffixe d'un langage $\mathcal{L}$ notée $\mathrm{Suff}(\mathcal{L})$ est le langage formé par l'ensemble des suffixes des mots de $\mathcal{L}$.

$$\mathrm{Suff}(\mathcal{L}) = \left\{w \in \Sigma^*, \exists v \in \Sigma^*, vw \in \mathcal{L}\right\} \tag{27}$$

■ **Définition 34 — Fermeture d'un langage par facteur.** La fermeture par facteur d'un langage $\mathcal{L}$ notée Fact($\mathcal{L}$) est le langage formé par l'ensemble des facteurs des mots de $\mathcal{L}$.

$$\text{Fact}(\mathcal{L}) = \left\{w \in \Sigma^*, \exists u, v \in \Sigma^*, uwv \in \mathcal{L}\right\} \tag{28}$$

F Propriétés fondamentales

On peut dors et déjà observer plusieurs faits :

- Soit v et w deux mots de Σ^*. A priori $vw \neq wv$, la concaténation n'est pas commutative.
- La décomposition d'un mot de Σ^* est unique en élément de Σ. Ceci est dû au fait qu'il n'y a pas d'inverse dans un monoïde. On peut le démontrer en raisonnant par l'absurde, en supposant qu'il existe deux décompositions différentes d'un même mot. Comme les lettres ne peuvent pas disparaître par inversion et qu'elles sont atomiques, c'est à dire non décomposables, on aboutit à une contradiction.

Ces deux observations engendrent de nombreux développements dans la théorie des langages.

Théorème 4 — Lemme de Levi. Soient t,u,v et w quatre mots de Σ^*. Si $tu = vw$ alors il existe un unique mot $z \in \Sigma^*$ tel que :

- soit $t = vz$ et $zu = w$,
- soit $v = tz$ et $zw = u$.

Démonstration. Supposons que $|t| \geqslant |v|$. Alors v est un préfixe de t et il existe un mot z tel que $t = vz$. Or, on a $tu = vw = vzu$. Par simplification à gauche, on obtient $w = zu$. On procède de même pour la seconde égalité. ■

Le lemme de Levi est illustré sur la figure 1.

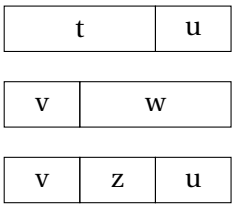

FIGURE 1 – Illustration du lemme de Levi