Introduction à la théorie des langages

Claire.Lefevre@univ-angers.fr

1

Les langages formels ont été étudiés par :

- les informaticiens
  ⇒ langages de programmation
  (définir syntaxe, vérifier la syntaxe d'un programme, le traduire en langage machine)
- les linguistes
  ⇒ langues naturelles
  (les décrire et essayer de les traiter automatiquement)

2

- Exemples de langages
  – Les entiers naturels (suite de chiffres parmi 0..9)
  – Les entiers naturels impairs (même représentation)
  – Les mots français (du dictionnaire)
  – Les identificateurs en C++
  – Les phrases en français
  – Les programmes (syntaxiquement corrects) écrits en C++
- Points communs
  – Chaque langage est un ensemble d'éléments (« chaînes »)
  – Chaque chaîne est une suite de « symboles » pris parmi un ensemble fini de symboles
  – Chaque chaîne est de longueur finie (même s'il n'y a pas de limite à cette longueur)

3

- On étudie des modèles pour représenter de manière finie des langages :
  – Automates finis
  – Expressions régulières
  – Grammaires formelles
  – …
- Applications pratiques
  – Recherche de « motifs » dans des fichiers
  – Traitement de texte
  – Modélisation de circuits
  – de machines à états
  – Compilation de langages de programmation
  – …

4

Langages
Concepts de base

5

Alphabets

- Un alphabet est un ensemble **fini**, non vide, de symboles
- On le note généralement Σ (sigma)

Exemples

$\Sigma_{entiers}$ = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
Σ_{mots} = {a, b, …, z, é, è, ç, à, ù,…, ', -}
Σ_{idents} = {a, ..., z, A, …, Z, 0, …, 9, _}
Σ_{prog} = {int, float, bool, while, do, for, …,
<, <=, >, >=, =, !=, +, -, /, *, ;, …,
0, 1, …, 25, 26, 27,…, 12.56, …,
a, b, toto, compteur, Tab, …}

6

Chaînes

- Un mot ou une chaîne ω formé(e) sur un alphabet est une suite finie $s_1 s_2 \ldots s_n$ de symboles de cet alphabet
- La chaîne vide, notée ε (epsilon), est une chaîne ne contenant aucun symbole
- La longueur d'une chaîne ω, notée |ω|, est le nombre de symboles composant la chaîne ω

Opérations sur les chaînes

- La concaténation de 2 chaînes u et v, notée u.v ou uv, est la chaîne obtenue en écrivant les symboles de u suivis de ceux de v

  si $u = a_1 a_2 \ldots a_n$ *et* $v = b_1 b_2 \ldots b_p$

  alors $uv = a_1 a_2 \ldots a_n b_1 b_2 \ldots b_p$

Propriétés :

- $|u.v| = |u| + |v|$
- Associativité : (u.v).w = u.(v.w)
- ε est élément neutre : u.ε = ε.u = u

- Puissances d'une chaîne ω
  - ω^k est la chaîne formée par la concaténation de k occurrences de ω

    $$\omega^k = \underbrace{\omega\,\omega\,\omega\ \ldots\ \omega\,\omega}_{k \text{ fois}}$$

  - $\omega^0 = \varepsilon$

- Un préfixe d'une chaîne ω est une suite, éventuellement vide, de symboles débutant ω
- Un suffixe de ω est une suite de symboles terminant ω

  ∀x,y t.q. ω = x.y x est un préfixe de ω et y est un suffixe de ω

- Une sous-chaîne d'une chaîne ω est une suite de symboles apparaissant consécutivement dans ω
- Notation : $|\omega|_x$ est le nombre d'occurrences de la chaîne x dans la chaîne ω

Langages

- Un langage est un ensemble de chaînes

Exemples

- {toto, titi, tata}
- {1, 11, 101, 1001}
- $\{1^n \mid n >= 0\} = \{\varepsilon, 1, 11, 111, 1111, 11111, \ldots\}$
  c'est un langage infini (nombre infini de chaînes) dont chaque chaîne est de longueur finie
- Nombres binaires impairs
  {1, 11, 101, 111, 1001, 1011, …}
- Nombres binaires qui sont premiers
  {1, 10, 11, 101, 111, 1011, …}

- Le langage vide, noté ∅, ne contient aucune chaîne
- Attention : ∅ ≠ {ε}
- Le langage « plein », noté Σ^*, contient toutes les chaînes que l'on peut former sur l'alphabet Σ
- Σ^+ contient toutes les chaînes *non vides* sur Σ

Rem : $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$

Opérations sur les langages

- L'union de 2 langages A et B est le langage, noté A∪B, composé de toutes les chaînes qui apparaissent dans l'un au moins des langages A ou B :

  $$A \cup B = \{\omega \mid \omega \in A \text{ ou } \omega \in B\}$$

Propriétés :

- Commutativité : A ∪ B = B ∪ A
- Associativité : (A ∪ B) ∪ C = A ∪ (B ∪ C)
- ∅ est élément neutre : A ∪ ∅ = ∅ ∪ A = A
- Idempotence : A ∪ A = A

- La concaténation de 2 langages A et B est le langage, noté A.B ou AB, composé de toutes les chaînes formées par une chaîne de A concaténée à une chaîne de B :

$$A.B = \{u.v \mid u \in A, v \in B\}$$

Propriétés :
- Associativité : (A.B).C = A.(B.C)
- $\{\varepsilon\}$ est élément neutre : $A.\{\varepsilon\} = \{\varepsilon\}.A = A$
- $\varnothing$ est élément absorbant : $A.\varnothing = \varnothing.A = \varnothing$

- Distributivité de la concaténation sur l'union :
  - À gauche : $A . (B \cup C) = A.B \cup A.C$
  - À droite : $(B \cup C) . A = B.A \cup C.A$

- Puissances d'un langage A :

  A^k est le langage formé par la concaténation de k occurrences de A
  - $A^0 = \{\varepsilon\}$
  - $A^1 = A$
  - $A^n = \underbrace{A\ A\ A\ \ldots\ A\ A}_{\text{n fois}}$

  A^k : *« mots formés par la concaténation de k mots de A »*

- Étoile de Kleene (fermeture ou clôture par .)
  - La fermeture de Kleene d'un langage A est le langage, noté A^*, défini par : $A^* = A^0 \cup A^1 \cup A^2 \cup A^3 \cup \ldots$

  « mots formés par la concaténation d'un nbre qcq de mots de A »
  - La fermeture positive de A est le langage, noté A^+, défini par : $A^+ = A^1 \cup A^2 \cup A^3 \cup \ldots$

  « mots formés par la concaténation de 1 ou plusieurs mots de A »

Rem : $A^* = \{\varepsilon\} \cup A^+$

Exercice : calculer A^+ et A^* pour les langages suivants
1. A = {a, ab}
2. A = {ab}
3. A = $\{\varepsilon, ab\}$
4. A = $\{a^n \mid n \in N\}$

Propriété : $A^+ = A.A^* = A^*.A$

Rem : il est possible que $A^+ = A^*$

- L'intersection de 2 langages A et B est le langage, noté $A \cap B$, composé des chaînes apparaissant à la fois dans A et dans B :

$$A \cap B = \{\omega \mid \omega \in A \text{ et } \omega \in B\}$$

- La différence de 2 langages A et B est le langage, noté A\B ou A-B, composé des chaînes de A n'apparaissant pas dans B :

$$A \setminus B = \{\omega \mid \omega \in A \text{ et } \omega \notin B\}$$

- Le complémentaire d'un langage A sur un alphabet Σ est le langage noté $\overline{A}$ composé de toutes les chaînes de Σ^* n'apparaissant pas dans A :

$$\overline{A} = \Sigma^* \setminus A$$

Modèles et langages

- On va étudier 3 modèles pour représenter des langages
  - Automates finis : « machines » qui permettent de déterminer si une chaîne donnée appartient oui ou non à un langage
    => reconnaissance d'un langage
  - Expressions régulières : notation qui permet de définir exactement quelles chaînes constituent un langage
    => spécification d'un langage
  - Grammaires formelles : notation récursive pour spécifier un lang.
- On classifie les langages selon le type de modèles permettant de les représenter :
  - Les langages réguliers (modélisés par les automates finis et les expressions régulières)
  - Les langages non contextuels (modélisés par les grammaires non contextuelles)
  - …