

Theorie des Langages

Matthew Coyle

January 21, 2018

1 Introduction

Les langages formels ont été étudiés par:

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

Exemples de langages:

- Les entiers naturels (suites de chiffres parmi 0..9)
- Les entiers naturels impairs (même représentation)
- Les mots français (du dictionnaire)
- Les phrases en français
- Les programmes (syntaxiquement corrects) écrits en C++

Points communs:

- Chaque langage est un ensemble d'éléments appelés mots ou "chaînes".
- Chaque chaîne est une suite de symboles pris parmi un ensemble fini de symboles.
- Chaque chaîne est de longueur finie.

On étudie des modèles pour représenter de manière finie des langages:

automate fini

expressions régulières

grammaire formelles

Applications pratiques:

Recherche de "motifs" dans les fichiers

traitement de texte

modélisation de circuits

modélisation de machines à états

Compilation de langages de programmation

un alphabet est un ensemble fini, non vide, de symboles. On le note généralement Σ .

Exemple d'alphabets:

$$\Sigma_{entiers} = \{0,1,2,3,4,5,6,7,8,9\}$$

$$\Sigma_{mots} = \{a,b,c,\dots,z, ', -\}$$

$$\Sigma_{ident} = \{a,\dots,z,A,\dots,Z,0,\dots,9,-\}$$

$$\Sigma_{prog} = \{int, float, bool, while, i, toto, a, \dots\}$$

Un mot ou une chaîne w formée(e) sur un alphabet est une suite finie $s_1s_2\dots s_n$ de symboles de cet alphabet

La concatenation de deux chaînes u et v , notée $u.v$ ou uv est la chaîne obtenue en écrivant les symboles de u suivis de ceux de v .

$$\text{si } u = a_1a_2\dots a_n \text{ et } v = b_1b_2b_p$$

$$\text{alors } uv = a_1a_2\dots a_nb_1b_2\dots b_p$$

Un préfixe d'une chaîne w est une suite de symboles débutant w .

Un suffixe de w est une suite de symboles terminant w .

$\forall x, y \mid w = x.y$, x est un préfixe de w , y un suffixe.

Une sous-chaîne d'une chaîne w est une suite de symboles apparaissant consécutivement dans w .

Un langage est un ensemble de chaînes.

Exemple de langages:

$$\{toto, titi, tata\}$$

$$\{1, 11, 101, 1001\}$$

$$\{1^n \mid n \geq 0\} = \{e, 1, 11, 111, 1111, 11111, \dots\}$$

$$\text{Nombres binaires impaires: } \{1, 11, 101, 111, 1001, 1011, \dots\}$$

$$\text{Nombres binaires premiers: } \{1, 10, 11, 101, 111, 1011, \dots\}$$

le Langage vide, noté \emptyset , ne contient aucune chaîne (ensemble vide).

Remarque: $\emptyset \neq \{\varepsilon\}$

Le langage plein, noté Σ^*

l'Union de deux langages A et B est le langage, noté $A \cup B$, composé de toutes les chaînes qui apparaissent dans l'un au moins de langages A ou B .

La concatenation de deux 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 .

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

Propriétés:

Associativité: $(A.B).C = A.(B.C)$

$\{\varepsilon\}$ est un élément neutre: $A.\{\varepsilon\} = \{\varepsilon\}.A = A$

est élément absorbant

Distributivité de la concaténation sur l'union:

Puissance d'un langage A A^k est le langage formé par la concaténation de k occurrences de A .

$$A^0 = \{\varepsilon\}$$

$$A^1 = A$$

Etoile de Kleene (fermeture ou clôture par $*$).

- la fermeture de Kleene d'un langage A est le langage, noté A^*

$$A^* = \bigcup_{i=0}^{\infty} A^i$$

2 Modèles et Langages

- contextuels
- langages récursivement énumérables