

Théorie des Langages

Axel PIGEON

26 mars 2025

Table des matières

1	Rappels sur les ensembles	2
2	Mots et Langages	3
2.1	Alphabets et Mots	3
2.2	Relations d'Ordre	6
2.3	Langage	7
2.4	Langage Décidable	9
3	Automate Fini Déterministe	10
3.1	Définition et représentation	10
3.2	Mot et Langage Automatique	11
4	Automate Fini Non Déterministe	13
4.1	Généralités	13
4.2	Juxtaposition et construction d'un AFD	15
5	Automate Fini A Epsilon Transition	18
5.1	Généralités	18
5.2	Déterminisation	19
6	Opérations entre automates	21
6.1	Langages Elémentaires	21
6.2	Automate Complémentaire	22
6.3	Somme d'automates	22
6.4	Intersection d'Automates	23
6.5	Différence d'Automates	23
6.6	Langages Automatiques	24
7	Langage d'un Automate	25
7.1	Langage d'un automate fini	25
7.2	Lemme d'Arden	26
8	Langages Algébriques et Automates à Piles	28
8.1	Grammaires Formelles	28

Chapitre 1

Rappels sur les ensembles

Rappels sur les ensembles

Chapitre 2

Mots et Langages

Contents

2.1	Alphabets et Mots	3
2.1.1	Premières Définitions	3
2.1.2	Opérations sur les mots	4
2.1.3	Puissance d'un mot	4
2.2	Relations d'Ordre	6
2.2.1	Ordre Préfixe	6
2.2.2	Ordre Lexicographique	6
2.3	Langage	7
2.3.1	Définition	7
2.3.2	Opérations	7
2.3.3	Propriétés	8
2.3.4	Expressions Régulières	8
2.4	Langage Décidable	9

Si on connaît plusieurs langages de programmation, on remarque que chaque langage, ou plutôt chaque paradigme de langage est spécialisé pour la résolution d'une catégorie de problèmes. On pourrait se demander s'il est possible de créer un langage permettant de résoudre tous les problèmes. Pour cela, il nous faudrait d'abord être capable de définir formellement un langage, des mots, etc...

2.1 Alphabets et Mots

2.1.1 Premières Définitions

Commençons tout d'abord par redéfinir correctement la notion d'alphabet et de mot.

Définition (Alphabet) . Un alphabet est simplement un ensemble fini, noté Σ . On nomme "lettre" ou "symbole" les éléments d'un alphabet.

Exemple Quelques exemples d'alphabets :

- $\Sigma = \{a, b\}$
- $\Sigma = \{a, b, \dots, \%, \$\}$

Définition (Mot) . Un mot est une suite finie de lettres d'un alphabet.

Proposition Le mot vide est noté ε . On note l'ensemble des mots d'un alphabet Σ^* .

Définition (Longueur d'un mot) . On note $|w|$ la longueur d'un mot $w \in \Sigma^*$ qui correspond au nombre de lettres, avec répétition du mot w .

Définition (Egalité de mots) . On dit que deux mots $u, v \in \Sigma^*$ sont égaux ssi :

- $|v| = |u|$
- $\forall i \in \llbracket 1, |v| \rrbracket, u[i] = v[i]$ où $u[i]$ est la i ème lettre du mot u .

Deux mots sont égaux ssi ils sont de même longueur et sont composés des mêmes lettres dans le même ordre.

2.1.2 Opérations sur les mots

Sur les mots, on ne définit qu'une seule opération, la **concaténation**.

Définition (Concaténation) . Soient $u, v \in \Sigma^*$ deux mots définis sur un même alphabet. On appelle la concaténation l'application :

$$\begin{cases} \Sigma^* \times \Sigma^* \longrightarrow \Sigma^* \\ (u, v) \longmapsto w = u.v \end{cases}$$

Elle est définie telle que $\forall u, v \in \Sigma^*$ de longueur $n, p \in \mathbb{N}$, on ait :

- $|u.v| = |u| + |v| = n + p$
- $\forall i \in \llbracket 1, n \rrbracket u.v[i] = u[i]$ et $\forall i \in \llbracket 1, p \rrbracket, u.v[n + i] = v[i]$

On parlera identiquement de concaténation ou de produit.

Remarque Cette définition reprend bien la caractérisation de deux mots égaux.

Proposition (Propriétés de la concaténation) La concaténation est une application :

- **Associative** : $\forall u, v, w \in \Sigma^*, w.(u.v) = (w.u).v$
- **Pas commutative** pour un alphabet de plus d'une lettre.
- admet pour **élément neutre** le mot vide ε .

2.1.3 Puissance d'un mot

Une fois la concaténation définie pour un mot, on peut alors parler de puissance de mot. Définissons celle-ci par récurrence.

Définition (Puissance d'un mot) . Soit Σ un alphabet et $u \in \Sigma^*$, on a :

- $u^0 = \varepsilon$
- $u^1 = u$
- $\forall n \in \mathbb{N}, u^{n+1} = u^n.u$

Exemple $(ba)^3 = bababa$

Proposition Soient $u \in \Sigma^*$ on peut appliquer les règles "connues" des puissances d'où :

$$\forall n, p \in \mathbb{N}, u^{n+p} = u^n.u^p = u^p.u^n$$

On remarque que l'on peut effectuer des simplifications sur les égalités de mots.

Propriété (Simplifications) . L'ensemble Σ^* est simplifiable à gauche et à droite.

- $\forall u, v, w \in \Sigma^*, \quad u.w = v.w \implies u = v$
- $\forall u, v, w \in \Sigma^*, \quad w.u = w.v \implies u = v$

Ici, pas besoin d'inverse, la démonstration repose sur la définition de l'égalité entre deux mots.

2.2 Relations d'Ordre

Dans l'alphabet dit "classique" on possède un ordre lexicographique des mots permettant de les classer en fonction de leurs lettres et de la position de leurs lettres. Ici, nous allons définir deux types de relations d'ordre sur les mots.

Commençons par rappeler la définition de relation d'ordre.

Définition (Relation d'Ordre) . Soit \triangleleft une relation sur un ensemble E . On dit que \triangleleft est une **relation d'ordre** ssi pour tout $x, y, z \in E$, \triangleleft est :

- **Réflexive** : $x \triangleleft x$
- **Anti-Symétrique** : $x \triangleleft y$ et $y \triangleleft x \implies x = y$
- **Transitive** : $x \triangleleft z$ et $z \triangleleft y \implies x \triangleleft y$

2.2.1 Ordre Préfixe

Naturellement, on munit Σ^* d'un ordre préfixe permettant de classer les mots en fonction de leur préfixe. Cette relation peut être vue comme une forme d'inclusion de mots.

Définition (Ordre Préfixe) . Soient $u, w \in \Sigma^*$, on définit la relation d'ordre préfixe \sqsubseteq telle que :

$$u \sqsubseteq w \iff \exists v \in \Sigma^*, w = u.v$$

Autrement dit, u est un préfixe de w ssi il existe un mot v tel que w soit composé de la concaténation de u et v .

Remarque L'ordre préfixe ne nécessite pas de relation d'ordre directement sur l'alphabet Σ .

Propriété (Ordre Préfixe et égalité) . Soient $u, v \in \Sigma^*$ on a :

$$u \sqsubseteq v \text{ et } v \sqsubseteq u \implies u = v$$

Démonstration Soient $u, v \in \Sigma^*$ tels que $\exists x, y \in \Sigma^*$ tels que

$$u = v.x \quad \text{et} \quad v = u.y$$

On a alors que :

$$\begin{cases} v = v.y.x \\ yx = \varepsilon \end{cases} \implies \begin{cases} y = \varepsilon \\ x = \varepsilon \end{cases} \implies u = v$$

Remarque Attention : l'ordre préfixe est une relation d'ordre partielle. Autrement dit, tous les éléments d'un même alphabet ne sont pas comparables.

2.2.2 Ordre Lexicographique

Définition (Ordre Lexicographique) . Soit Σ un alphabet que l'on muni d'une relation d'ordre \leq . L'ordre lexicographique \leq est une relation d'ordre totale sur Σ^* .

Remarque Ici, nous avons bien besoin de définir au préalable un ordre sur notre alphabet Σ .

Propriété (Compatibilité) . L'ordre lexicographique est compatible avec l'ordre préfixe. Plus formellement,

$$\forall u, v \in \Sigma^*, \quad u \sqsubseteq v \implies u \leq v$$

2.3 Langage

Maintenant que nous sommes au clair sur la définition de lettre et de mot, on peut enfin définir l'objet principal de ce chapitre, les langages.

2.3.1 Définition

Définition (Langage) . Soit Σ un alphabet, on appelle langage sur Σ toute partie de Σ^* .

Remarque L'ensemble de tous les langages d'un alphabet Σ est donc $\mathcal{P}(\Sigma^*)$, l'ensemble de toutes les parties de Σ^* .

Définition (Complémentaire) . Soit L un langage sur Σ . On définit le complémentaire de L dans Σ^* le langage :

$$\bar{L} = \{w, w \notin L\}$$

2.3.2 Opérations

De même que pour les alphabets et les mots, on peut définir des opérations sur les langages.

Définition (Opérations sur les langages) . Soit Σ un alphabet et $L_1, L_2 \subseteq \Sigma^*$ deux langages de Σ . On définit 4 principales opérations sur des langages :

- **Somme** : notée $+$, la somme de deux langages d'appartenance à l'union des ensembles.

$$L_1 + L_2 = \{w, w \in L_1 \text{ ou } w \in L_2\}$$

C'est une opération :

- Commutative
- Associative
- dont \emptyset est le neutre.

- **Intersection** : de même que pour les ensembles :

$$L_1 \cap L_2 = \{w, w \in L_1 \text{ et } w \in L_2\}$$

C'est une opération :

- Commutative
- Associative
- dont Σ^* est le neutre

- **Différence** : comme les ensembles, on définit la différence de langages :

$$L_1 / L_2 = \{w, w \in L_1 \text{ et } w \notin L_2\} = L_1 \cap \bar{L}_2$$

- **Produit de concaténation** : de même que pour les mots, on peut généraliser le produit de concaténation aux langages :

$$L_1.L_2 = \{u.v, u \in L_1, v \in L_2\}$$

C'est une opération :

- Associative
- Distributive par rapport à l'union
- D'élément neutre $\{\varepsilon\}$.

Définition (Puissance de langage) . Soit L un langage, on définit **par récurrence** la puissance de L par :

- $L^0 = \{\varepsilon\}$
- $L^1 = L$
- $\forall n \in \mathbb{N}^*, L^n = L^{n-1}.L$

Une fois définies des opérations "simples" sur les langages, on peut en définir des plus complexes, permettant de "générer" un langage infini à partir d'un langage fini ou infini.

Définition (Langage plus et étoile) . Soit L un langage sur un alphabet Σ . On définit le langage plus de L comme le langage :

$$L^+ = L^1 + L^2 + \dots$$

De même le langage étoile de L est défini par :

$$L^* = \{\varepsilon\} + L^1 + L^2 + \dots$$

2.3.3 Propriétés

Voyons quelques propriétés des langages...

Proposition Soient L_1 et L_2 deux langages sur un alphabet Σ , on a les propriétés suivantes :

- $\forall p \in \mathbb{N}$, on a :

$$(L_1)^p.(L_2)^p \subseteq (L_1)^*.(L_2)^*$$

- L'opération étoile est idempotente :

$$(L^*)^* = L^*$$

- $L^* = \{\varepsilon\} + L^+$
- $\varepsilon \in L \iff L^+ = L^*$

2.3.4 Expressions Régulières

Lorsque l'on manipule des langages infinis, il serait appréciable d'avoir une expression pratique pour un langage permettant de directement voir la forme des mots qu'il contient. On définit ainsi les expressions régulières.

Définition (Expression Régulière) . On définit récursivement une expression régulière sur un alphabet Σ :

- ε est une expression régulière.
- $\forall w \in \Sigma$ est une expression régulière.
- Si E est une expression régulière alors (E) l'est aussi.
- Si E_1 et E_2 sont des expressions régulières, alors $E_1 + E_2$ l'est aussi.
- Si E_1 et E_2 sont des expressions régulières, alors $E_1.E_2$ l'est aussi.
- Si E est une expression régulière, alors E^* l'est aussi.

Exemple Voyons quelques exemples d'expressions régulières sur un alphabet $\Sigma = \{a, b\}$:

$$a^*b, \quad (a+b)^*, \quad (a+b)^*ba(a+b)^*$$

Définition (Langage Régulier) . On dit qu'un langage est régulier si il peut s'écrire sous la forme d'une expression régulière.

Il sera donc préférable de travailler avec des langages réguliers.

2.4 Langage Décidable

L'objectif de ce cours est bien entendu de comprendre comment fonctionne un compilateur, pour pouvoir en créer un par nous même. Pour rappel, on doit d'abord bien comprendre les notions de langage et de mot pour pouvoir ensuite déterminer si un ensemble de mots est syntaxiquement corrects lors de la compilation.

Lors de la compilation, il faut d'abord commencer par savoir si un mot traité appartient au langage défini ou pas. Pour des langages finis, l'opération n'est pas compliquée, pour chaque mot il suffit de vérifier si il appartient à un ensemble fini. Pour des langages infinis, l'opération semble plus complexe, il va falloir trouver une manière systématique et efficace de définir si un mot appartient au langage ou pas.

On appelle ce genre de problème un problème de **décision**.

Définition (Langage Décidable) . Un langage L est dit décidable si il existe un algorithme permettant de dire si un mot w appartient ou pas au langage L .

Théorème (Nombre de Langages Décidables) . Il existe un nombre fini de langages décidables.

Autrement dit, il existe un nombre infini de langages non décidables...

Chapitre 3

Automate Fini Déterministe

Contents

3.1	Définition et représentation	10
3.2	Mot et Langage Automatique	11
3.2.1	Lecture d'un mot	11
3.2.2	Automate Complet	12
3.2.3	Complémentaire	12

Comme abordé dans le chapitre précédent, on cherche une méthode pratique et efficace pour déterminer si un mot appartient à un langage ou pas. On veut donc un modèle qui soit d'une part très pratique mathématiquement pour nous permettre de démontrer des choses dessus mais aussi facilement implémentable algorithmiquement.

Alerte Spoiler : de solides connaissances en théorie des graphes seront plus qu'utiles...

3.1 Définition et représentation

Définition (Automate fini déterministe) . Un Automate Fini Déterministe est un quintuplet :

$$\mathcal{A} = (\Sigma, Q, T, q_0, A)$$

où :

- Σ est un alphabet
- Q est un ensemble fini d'états (souvent une partie finie de \mathbb{N})
- $T : Q \times \Sigma \longrightarrow Q$ est une application qui, à un état et une lettre associe un autre état.
- q_0 un état initial
- $A \subseteq Q$ les états acceptants

On représentera ainsi un automate fini déterministe de plusieurs façons en fonction de son utilisation :

- **Mathématique** : $\mathcal{A} = (\Sigma, Q, T, q_0, A)$
- **Table de Transition** : Elle va permettre de trouver rapidement les différents types d'états.
- **Sagittale** : Sous forme de graphe
- **En Python** : Nous représenterons les automates finis déterministes sous la forme de quintuplet aussi.

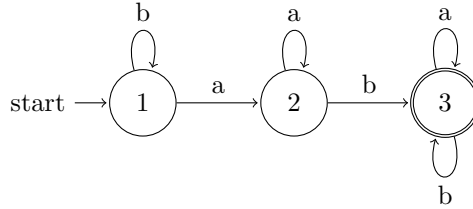
Regardons en détail ces différentes représentations au travers d'un exemple.

Exemple Soit $\mathcal{A} = (\Sigma, Q, T, q_0, A)$ un automate fini déterministe.

Mathématiquement nous avons :

- $Q = \{1, 2, 3\}$
- $\Sigma = \{a, b\}$
- $q_0 = 1$
- $A = \{3\}$

La représentation **sagittale** de notre automate sera :



On représente de façon doublement cerclée les états acceptants. Les états sont les sommets du graphe. Les arcs valués sont les antécédants/images de la fonction T .

Enfin, la **table de transition** de l'automate est représentée par le tableau suivant :

Q/ Σ	a	b
1	2	3
2	2	3
3	3	3

La première ligne présente les lettres de l'alphabet et la première colonne les différents états. Pour chaque état, le tableau donne l'état obtenu en fonction de la lettre suivante lue.

3.2 Mot et Langage Automatique

3.2.1 Lecture d'un mot

Définition (Lecture d'un mot) . Soit Σ un dictionnaire, $l \in \Sigma$ et \mathcal{A} un automate. On lit la lettre l en dérivant d'un état $q \in Q$ vers un état $q' \in Q$ et si $T(q, l) = q'$. On notera la lecture d'un mot de longueur p par la lecture successive de ses lettres :

$$q_1 \xrightarrow{l_1} q_2 \dots q_{p-1} \xrightarrow{l_p} q_p$$

Concrètement, pour la lecture d'un mot, on va partir de l'état initial et en fonction des valeurs de l'état courant et de la lettre lue, on va "bouger" d'un état (sommets) à un autre.

Définition (Mot refusé) . Un mot $w \in \Sigma^*$ est dit refusé par un automate \mathcal{A} si sa lecture à partir de l'état initial se termine sur un état refusant ou ne se termine pas. Dans le cas contraire, w est dit accepté.

Définition (Langage d'un Automate) . Le langage d'un automate \mathcal{A} est l'ensemble des mots acceptés par l'automate. On le note $L(\mathcal{A})$.

On parle de **langage automatique** si il est reconnaissable par un automate. Deux automates sont dits **équivalents** si ils reconnaissent le même langage.

3.2.2 Automate Complet

Définition (Automate Complet) . Un automate \mathcal{A} est dit complet si

$$\forall i \in Q, \forall l \in \Sigma, \quad T(i, l) \text{ est défini}$$

Autrement dit, un automate est dit complet si pour toute lettre et pour tout état fixés, il est possible de changer d'état dans l'automate.

Définition (Puit) . Un état $q \in Q$ est un puit ssi

$$\forall l \in \Sigma, \quad T(q, l) = q$$

Un puit est un état duquel on ne peut sortir.

On définit aussi la notion de piège comme un puit refusant (i.e un puit dont l'état est refusant).

3.2.3 Complémentaire

Puisque la notion de complémentaire existe pour les langages et que les automates semblent très étroitement liés aux langages, on peut se demander si un automate peut admettre un complémentaire...

Soit \mathcal{A} un automate associé à un langage L . On cherche $\mathcal{A}' = \overline{\mathcal{A}}$.

$$w \in \overline{\mathcal{A}} \iff w \notin L \iff w \notin L(\mathcal{A})$$

Il semble falloir que \mathcal{A} soit complet. Si c'est le cas, on pourrait inverser \mathcal{A} en inversant les sommets acceptants/refusants.

Proposition Tout automate fini peut être complété par des puits refusants en un automate complet.

Théorème (Complémentarité) . L'ensemble des automates est stable par complémentarité.

Chapitre 4

Automate Fini Non Déterministe

Contents

4.1	Généralités	13
4.1.1	Définitions	13
4.1.2	Lecture d'un mot	14
4.2	Juxtaposition et construction d'un AFD	15
4.2.1	Juxtaposition	15
4.2.2	Déterminisation	15

Dans le chapitre précédent nous avons vu un modèle très efficace pour vérifier l'appartenance d'un mot à un langage. En plus d'être facilement représentable en mémoire (i.e Python), il est facile à utiliser à la main et hérite de toute la théorie des graphes vue précédemment ce qui en fait un très beau modèle mathématiquement parlant.

Malgré tout cela, nous ne savons pas comment, à partir de plusieurs langages simples, constituer un automate reconnaissant la somme de ces langages. Nous n'avons pas défini de somme/union d'automate et celles-ci semblent assez difficiles vu la rigidité de notre modèle.

Nous allons donc construire un modèle d'automate, appelé non déterministe, nous permettant de faire ces opérations d'union (que nous appellerons juxtaposition). Elles nous permettront de construire des automates complexes à partir de somme de langages.

4.1 Généralités

4.1.1 Définitions

Définition (AFN) . Un automate fini non déterministe \mathcal{A} est un quintuplet :

$$\mathcal{A} := (Q, \Sigma, T, I, A)$$

tel que :

- Q est l'ensemble des états de l'automate
- Σ est un alphabet
- $T : Q \times \Sigma \longrightarrow \mathcal{P}(Q)$ est une application
- $I \subseteq Q$ est l'ensemble des états initiaux
- $A \subseteq Q$ est l'ensemble des états acceptants.

Remarque Plusieurs remarques concernant ce nouveau modèle. Premièrement, on remarque que l'on peut maintenant définir des transitions multiples entre les états de l'automate. Deuxièmement, il existe plusieurs états initiaux.

4.1.2 Lecture d'un mot

Définition (Arbre de lecture) . Soient $w \in \Sigma^*$, L un langage sur Σ et \mathcal{A} un automate reconnaissant L . L'arbre de lecture de w par \mathcal{A} est l'arbre résultat du parcours de \mathcal{A} en fonction des lettres de w .

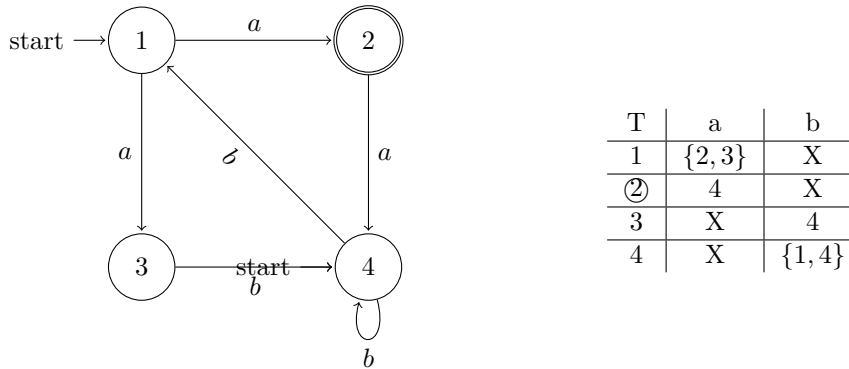
Autrement dit dans l'arbre de lecture G de w , un noeud a est le fils d'un noeud b si il existe une lettre l de w telle que $T(b, l) = a$. Une feuille de cet arbre est un état acceptant ou refusant de l'automate.

Définition (Lecture acceptante) . Soient $w \in \Sigma^*$, L un langage sur Σ et \mathcal{A} un automate reconnaissant L . Une lecture de w par \mathcal{A} est dite acceptante si il existe un chemin d'un état initial vers un état acceptant dans l'arbre de lecture de w par \mathcal{A} .

Proposition On peut dire plusieurs choses de la lecture d'un mot $w \in \Sigma^*$ par un automate \mathcal{A} :

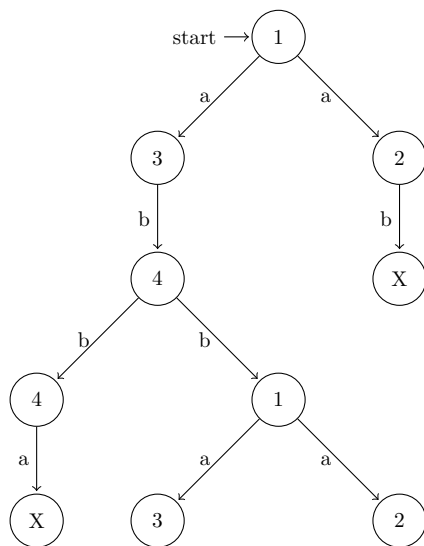
- Si l'automate possède plusieurs états initiaux, la lecture produit un arbre de lecture pour chaque état initial, nous auront donc une forêt d'arbres de lecture.
- Une lecture sera donc acceptante ssi il existe un arbre de la forêt dont au moins une des feuilles est un état acceptant.

Exemple Voyons tout cela sur un exemple. Soit $\mathcal{A} := (Q, \Sigma, T, I, A)$ un automate fini non déterministe \mathcal{A} sur l'alphabet $\Sigma : \{a, b\}$. Représentons notre automate sous forme de graphe orienté valué et sa table de transition :



C'est un automate non déterministe puisqu'il contient deux transitions multiples et deux états initiaux.

Posons $w := abba$, déterminons si ce mot appartient au langage $L(\mathcal{A})$. Nous allons construire un seul arbre permettant d'avoir une condition suffisante de validation du mot.



L'arbre de lecture du mot *abba* contient un état acceptant comme feuille.

Autrement dit, il existe un chemin menant d'un état initial à un état acceptant dans la forêt de lecture de *abba*. Donc $abba \in L(\mathcal{A})$.

4.2 Juxtaposition et construction d'un AFD

4.2.1 Juxtaposition

Rappelons la problématique principale du chapitre. On cherche un modèle dérivant des AFD nous permettant de définir des opérations dessus et qui puisse être converti algorithmiquement vers un AFD pour construire des automates d'un langage complexe à partir de langages plus simple.

Autrement dit, on veut pouvoir appliquer l'opération de somme de langages sur les automates fini déterministes.

Définition (Juxtaposition d'AFN) . Soient L_1 et L_2 deux langages reconnus par deux automates \mathcal{A}_1 et \mathcal{A}_2 . Le langage $L_1 + L_2$ est reconnu par la **juxtaposition disjointe** de \mathcal{A}_1 et \mathcal{A}_2 .

Théorème (Langages automatiques et stabilité) . L'ensemble des langages automatiques est stable par somme.

On peut maintenant, à partir de deux langages automatiques, définir le nouveau langage résultant de la somme des deux qui sera lui aussi automatique. Il suffit de faire la juxtaposition disjointe des deux automates de départ.

4.2.2 Déterminisation

Théorème (Existence et équivalence) . Pour tout automate fini non déterministe, il existe un automate déterministe équivalent.

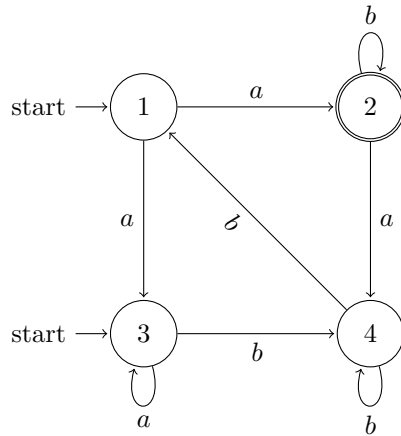
Ce théorème est peut être un peu obscur mais permet de dire qu'il est toujours possible de passer d'un automate fini non déterministe (obtenu par exemple par juxtaposition) à un automate fini déterministe qui reconnaisse le même langage. En tout cas, il nous dit qu'il en existe un...

L'intérêt de vouloir repasser chez les automates fini déterministes vient du fait que la lecture d'un mot par un AFN est de complexité exponentielle alors que la lecture d'un mot par un AFD est polynômiale... Lors de la vérification syntaxique de très long mots pour des langages très complexes, cela fait une différence cruciale pour la compilation.

Ce processus est appelé **déterminisation** d'un AFN.

Proposition Soit $\mathcal{A} = (Q, \Sigma, T, I, A)$ un AFN. On cherche à construire un AFD \mathcal{A}' équivalent à \mathcal{A} . L'idée est de raisonner sur l'application $T : Q \times \Sigma \longrightarrow \mathbb{Q}$. Dans un AFN, cette application n'est pas injective, on va donc poser une nouvelle application dans l'espace quotient de $Q \times \Sigma$ par le noyau de T . Nous obtiendrons donc une application injective et donc un AFD.

Exemple (Déterminisation) Soit \mathcal{A} l'automate défini sur l'alphabet $\Sigma = \{a, b\}$ non déterministe et sa table de transition suivants :



T	a	b
1	{2, 3}	X
②	4	2
3	3	4
4	X	{1, 4}

Déterminisons cet automate. Pour cela, nous allons renommer tous les états de l'automate en prenant en compte les ensembles. L'algorithme consiste donc à construire la table de transition du nouvel automate. Pour chaque itération (i.e ajout d'une ligne dans la table), on effectue un parcours en largeur du nouvel automate pour "découvrir" de nouveau état. On crée ainsi un "automate des parties".

	a	b
$I = \{1, 3\}$	II	III
$II = \{2, 3\}$	IV	V
$III = \{4\}$	-	VI
$IV = \{3, 4\}$	VII	VI
$V = \{2, 4\}$	III	$VIII$
$VI = \{1, 4\}$	II	VI
$VII = \{3\}$	VII	III
$VIII = \{1, 2, 4\}$	IX	$VIII$
$IX = \{2, 3, 4\}$	IV	$VIII$

FIGURE 4.1 – Table de l'AFD

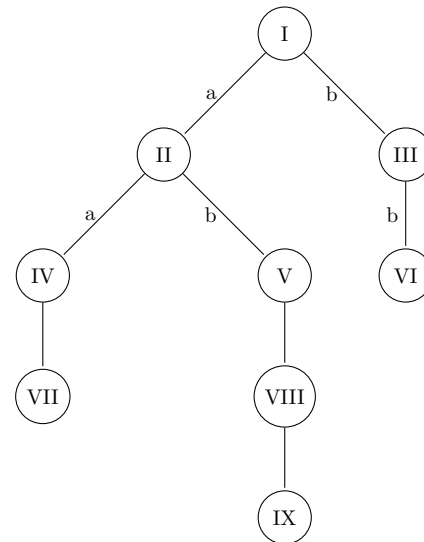


FIGURE 4.2 – Automate des parties

Chapitre 5

Automate Fini A Epsilon Transition

Contents

5.1	Généralités	18
5.2	Déterminisation	19

Définissons un nouveau type d'automates non déterministes. Les automates non déterministes à ε -transition. Il diffèrent des premiers puisque l'on va permettre le changement d'état sans lecture de lettres lors de la lecture d'un mot. Pour cela, nous allons définir une transition ε . Cela peut se voir comme une transition via le mot vide.

5.1 Généralités

Définition (Automate Fini à ε -transitions (AFN ε)). Un automate fini à ε -transitions est un quintuplet :

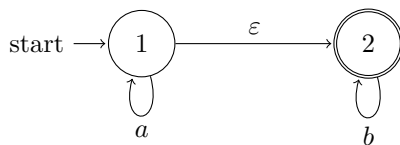
$$\mathcal{A} = (Q, \Sigma, T, I, A)$$

définit de la même façon que les automates précédents mais où :

$$T : Q \times \Sigma \cup \{\varepsilon\} \longrightarrow \mathcal{P}(Q)$$

Ici, le changement spontané d'état sans lecture de lettre sera donc caractérisé par une nouvelle entrée dans la table de transition ε .

Exemple Soit le langage $L := a^*b^*$. Un automate reconnaissant ce langage peut être écrit avec une ε -transition. Ecrivons aussi sa table de transition :



T	a	b	ε
1	1	-	2
2	-	2	-

Lors de la lecture d'un mot, les transitions peuvent être très aléatoires en fonction du nombre d' ε -transitions possibles de l'état courant. Un tel automate est donc hautement non déterministe.

Définition (Lecture d'un mot) . Soit $w = l_1 l_2 \dots l_n$ un mot sur Σ . Soit $w' = a_1 a_2 \dots a_p$ le mot w ε -complété (rembourré par des ε) tel que :

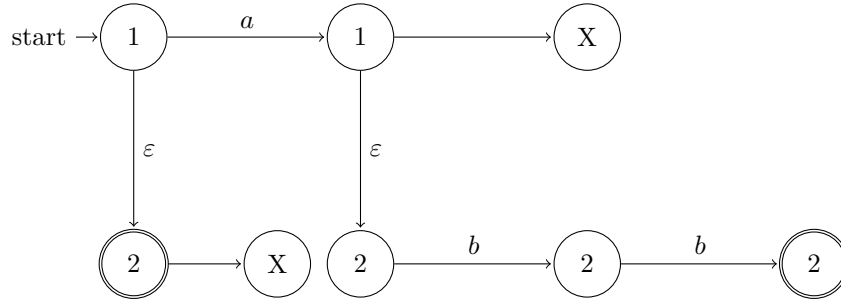
- $p \geq n$
- $\forall i \in \llbracket 1, n \rrbracket, a_i \in \{l_1, \dots, l_n\} \cup \{\varepsilon\}$
- $l_1 l_2 \dots l_n = a_1 a_2 \dots a_p$ (du point de vue du produit de concaténation)

Une lecture du mot w par \mathcal{A} est une lecture par \mathcal{A} de n'importe quel w' , un ε -complété de w .

Remarque Tout comme pour les automates précédents, un mot appartient au langage d'un automate ssi la lecture de ce mot par celui-ci se finit sur au moins un état acceptant de l'automate.

La lecture d'un mot par un $\text{AFN}\varepsilon$ conduira donc à la construction d'une forêt de lecture de ce mot par l'automate.

Exemple (Lecture d'un mot) Soit l'automate fini non déterministe à ε -transitions précédent. Soit le mot abb . Construisons la forêt de lecture de abb par \mathcal{A} .



Il existe un chemin d'un racine vers une feuille acceptante dans la forêt de lecture :

$$1 \xrightarrow{a} 1 \xrightarrow{\varepsilon} 2 \xrightarrow{b} 2 \xrightarrow{b} 2$$

Donc par définition, $abb \in L(\mathcal{A})$.

5.2 Déterminisation

Dans le chapitre précédent, un théorème nous permet de dire que pour tout automate fini non déterministe, il existe un automate fini déterministe équivalent. Ainsi, à chaque fois que l'on considère un $\text{AFN}\varepsilon$, on est sûr qu'il existe un AFD équivalent.

Pour la déterminisation d'un $\text{AFN}\varepsilon$, nous allons avoir besoin d'une définition supplémentaire...

Définition (Clôture) . Soit $\mathcal{A} = (Q, \Sigma, T, I, A)$ un $\text{AFN}\varepsilon$. Pour tout $q \in Q$, on appelle clôture de q l'ensemble des états accessibles à partir de q sans lecture de lettre lors de la lecture d'un mot dans l'automate.

Autrement dit, la clôture de q est l'ensemble des états accessibles depuis q dans le sous-graphe de \mathcal{A} restreint aux ε -transitions.

On note $cl(q)$ la clôture de q .

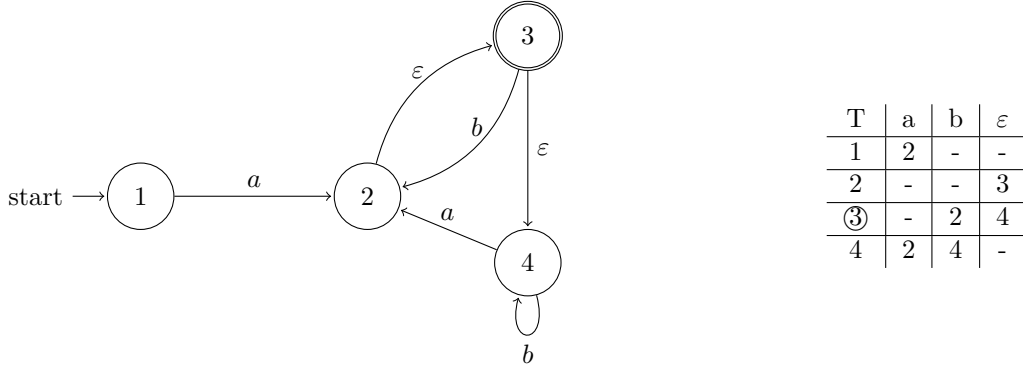
L'idée de la déterminisation d'un $\text{AFN}\varepsilon$ est, non plus de regrouper des états, mais d'étendre les transitions de l'automate à tous les états accessibles par ε -transitions.

De plus, si un état acceptant est accessible uniquement par ε -transition depuis un état, alors celui-ci hérite du caractère acceptant de l'état atteint.

Proposition (Algorithme de Déterminisation) Soit $\mathcal{A} = (Q, \Sigma, T, I, A)$ un AFN ε . On construit l'automate fini déterministe \mathcal{A}' équivalent à \mathcal{A} en :

1. Calculer les clôtures de \mathcal{A}
2. Héritage : Tous les états dont la clôture contient un état acceptant sont acceptants.
3. Calculer les transitions étendues
4. Déterminisation de \mathcal{A}' par l'automate des parties (voir chap précédent)

Exemple (Déterminisation) Soit \mathcal{A} l'AFN ε suivant :



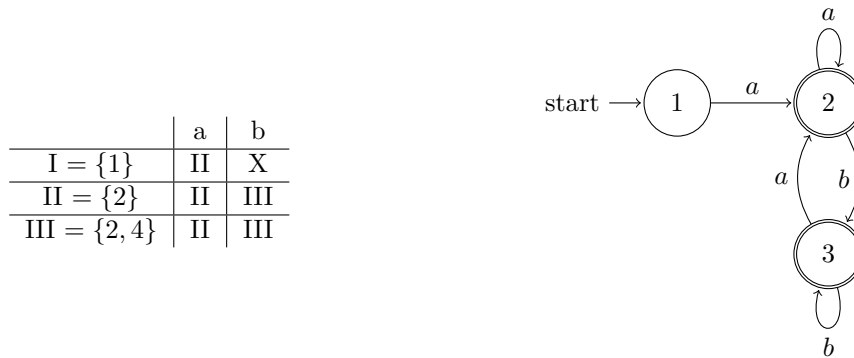
1. Calcul des clôtures et héritage

T	a	b	ε	$cl(q)$
1	2	-	-	{1}
②	-	-	3	{2, 3, 4}
③	-	2	4	{3, 4}
4	2	4	-	{4}

2. Calcul des transitions étendues

\tilde{T}	a	b	$cl(q)$
1	2	-	
②	2	{2, 4}	{2, 3, 4}
③	2	{2, 4}	{3, 4}
4	2	4	

3. Déterminisation par l'automate des parties



La déterminisation d'un AFN ε nous permet donc de passer de la lecture d'un mot de complexité exponentielle (voire infinie) à un automate permettant de lire tous les mots avec une complexité linéaire.

Chapitre 6

Opérations entre automates

Contents

6.1	Langages Elémentaires	21
6.2	Automate Complémentaire	22
6.3	Somme d'automates	22
6.4	Intersection d'Automates	23
6.5	Différence d'Automates	23
6.6	Langages Automatiques	24
6.6.1	Théorème de pompage	24

Les chapitres précédents nous ont montré que les ε -transitions et les transitions multiples nous permettent de modéliser plus facilement des langages complexes. Nous allons donc définir des opérations sur les automates, analogues à celles sur les langages. Elles nous permettront, à partir de la construction d'un langage par opérations, de construire son automate par opérations aussi.

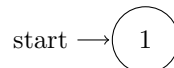
Problématique : Soient L_1 et L_2 deux langages reconnus respectivement par \mathcal{A}_1 et \mathcal{A}_2 . Soit T une opération entre langages. Comment construire un automate reconnaissant $L_1 T L_2$?

6.1 Langages Elémentaires

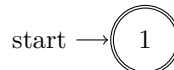
Pour définir des opérations sur les automates, et ainsi, construire un automate par opérations pour un langage lui-même construit par opérations, nous allons avoir besoin de définir des langages élémentaires. Ces langages seront reconnus par des automates fixés, que nous connaissons d'avance. Nous en choisissons un nombre fini pour pouvoir les stocker en mémoire. Ils vont représenter les briques de base nous permettant de construire des automates plus complexes par la suite.

Proposition (Langages Elémentaires) Soit $\Sigma = \{a, b\}$ un alphabet. On définit les langages élémentaires de Σ comme :

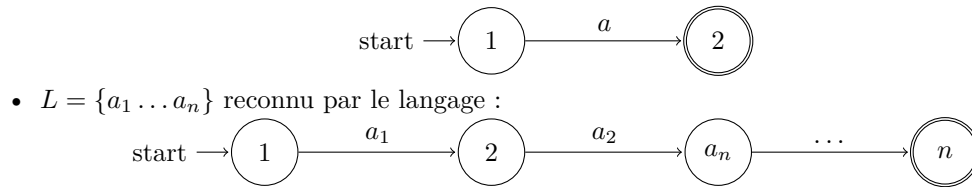
- $L = \emptyset$ reconnu par l'automate :



- $L = \{\varepsilon\}$ reconnu par l'automate :



- $L = \{a\}$ reconnu par le langage :



6.2 Automate Complémentaire

A partir d'un langage, on peut définir son langage complémentaire. De même, on peut définir l'automate complémentantaire reconnaissant ce langage.

Définition (Automate Complémentaire) . Soit L un langage reconnu un automate $\mathcal{A} = (Q, \Sigma, T, q_0, A)$ **complet**. L'automate reconnaissant le langage complémentaire de L est $\bar{\mathcal{A}} = (Q, \Sigma, T, q_0, Q \setminus A)$.

Remarque Attention, pouvoir "passer au complémentaire" pour un automate, il faut que celui-ci soit complet.

6.3 Somme d'automates

Définition (Somme d'automates) . Soient L_1 et L_2 deux langages respectivement reconnus par \mathcal{A}_1 et \mathcal{A}_2 . L'automate reconnaissant $L_1 + L_2$ est l'automate fini non déterministe construit par l'**union disjointe** de \mathcal{A}_1 et \mathcal{A}_2 . On l'appelle **automate somme** des automates \mathcal{A}_1 et \mathcal{A}_2 .

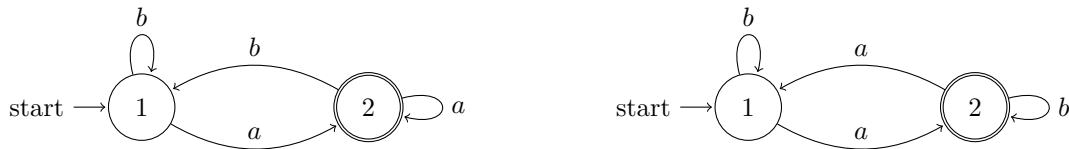
Cet automate n'est pas déterministe puisqu'il contient deux états initiaux mais que l'on peut déterminer.

Exemple Soient les langages suivants sur $\Sigma = \{a, b\}$:

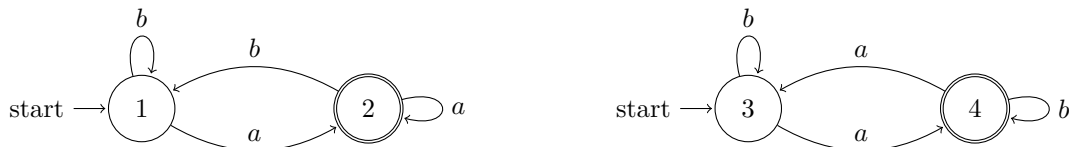
$$L_1 = \{\text{mots de } \Sigma \text{ terminant par } a\}$$

$$L_2 = \{\text{mots de } \Sigma \text{ contenant un nombre pair de } a\}$$

Ces langages sont reconnus par les automates suivants :



On construit l'automate \mathcal{A} comme la juxtaposition disjointe des deux automates :



Que l'on doit ensuite déterminer...

Les définitions de langages élémentaires nous permettent de savoir que tout langage réduit à un mot est automatique. On en déduit le théorème suivant :

Théorème (Langage Fini) . Tout langage fini est automatique.

6.4 Intersection d'Automates

Proposition Soient L_1 et L_2 reconnus par les automates fini déterministes suivants \mathcal{A}_1 et \mathcal{A}_2 . On a alors :

$$\overline{L_1 \cap L_2} = \overline{L_1} + \overline{L_2}$$

On peut donc en conclure que :

$$L_1 \cap L_2 = \overline{\overline{L_1} + \overline{L_2}}$$

Ainsi, l'intersection de deux automates peut être construite par somme et complémentarité.

En pratique nous utiliseront plutôt l'automate des couples :

Définition (Automate des Couples) . Soient $\mathcal{A}_1 = \{Q_1, \Sigma, T_1, Q_0, A_1\}$ reconnaissant le langage L_1 et $\mathcal{A}_2 = \{Q_2, \Sigma, T_2, q'_0, A_2\}$ reconnaissant le langage L_2 . On définit l'automate des couples :

$$A = \{Q_1 \times Q_2, \Sigma, T, (q_0, q'_0), A_1 \times A_2\}$$

reconnaissant le langage $L_1 \cap L_2$ où

$$\forall i \in \Sigma, \forall q_1, q_2 \in Q_1, \forall q'_1, q'_2 \in Q_2 \text{ tels que } q_2 = T_1(q_1, i) \text{ et } q'_2 = T_2(q'_1, i)$$

$$\text{alors } T'((q_1, q'_1), i) = (q_2, q'_2)$$

Proposition Si \mathcal{A}_1 possède $n \in \mathbb{N}$ états et que \mathcal{A}_2 possède $p \in \mathbb{N}$ états, alors $\mathcal{A}_1 \cap \mathcal{A}_2$ possède $n \times p$ états. Pour de gros automates, cette méthode peut donc engendrer des très gros. Même si la façon de les construire est assez simple et ressemble beaucoup à la détermination. L'automate obtenu est, de plus, déterministe.

6.5 Différence d'Automates

Proposition Soient \mathcal{A}_1 et \mathcal{A}_2 deux automates reconnaissant respectivement les langages L_1 et L_2 . Pour reconnaître le langage $L_1 \setminus L_2$, on peut simplement construire l'automate reconnaissant :

$$L_1 \setminus L_2 = L_1 \cap \overline{L_2}$$

6.6 Langages Automatiques

Essayons maintenant de déduire des conditions nécessaires pour qu'un langage soit automatique. D'après ce que l'on a vu grâce aux opérations, on peut déjà énoncer la proposition suivante :

Proposition Les langages réguliers sont tous automatiques.

Proposition que nous élargirons plus tard grâce au théorème de Kleene.

6.6.1 Théorème de pompage

Théorème (Pompage (faible)) . Soit L un langage. Supposons L automatique. Soit $w \in L$ alors pour toute décomposition de w de la forme :

$$\exists w_1, w_2, w_3 \in L, w = w_1 w_2 w_3$$

alors cette décomposition est gonflable.

$$\text{i.e } \forall k \in \mathbb{N}, w_1 w_2^k w_3 \in L$$

Ce théorème n'est pas idéal pour montrer qu'un langage est automatique. En revanche, sa contraposé de la forme :

Non Pompable \implies Non Automatique

Est en pratique très utilisée pour montrer qu'un langage n'est pas automatique.

Chapitre 7

Langage d'un Automate

Contents

7.1	Langage d'un automate fini	25
7.2	Lemme d'Arden	26

Maintenant que nous savons bien manipuler les automates fini déterministes et non déterministes, il serait utile, à partir d'un automate de pouvoir déterminer le langage qu'il reconnaît. Pour cela nous aurons besoin de définir les systèmes d'équations aux langages, d'introduire le Lemme d'Arden. Nous finirons par énoncer le théorème de Kleene caractérisant les langages automatiques.

7.1 Langage d'un automate fini

Définition (Langage d'arrivée) . Soit $\mathcal{A} = \{Q, \Sigma, T, q_0, A\}$ un automate fini. On définit le langage d'arrivée à l'état $q \in Q$, noté L_q l'ensemble des mots de Σ^* dont la lecture par \mathcal{A} débute par q_0 et finit en q .

On peut alors définir la langage d'un automate comme :

Proposition Soit $\mathcal{A} = \{Q, \Sigma, T, q_0, A\}$ un automate fini. Soit $\{L_q \mid q \in A\}$ l'ensemble des langages d'arrivés aux états acceptants de l'automate \mathcal{A} . On a alors l'égalité suivante :

$$L(\mathcal{A}) = \sum_{q \in A} L_q$$

Autrement dit, le langage de \mathcal{A} est la somme de tous ses langages d'arrivé aux états acceptants.

Définition (Système d'équations aux langages) . Soient un ensemble X_1, \dots, X_n de langages sur un même alphabet Σ . Un système d'équations aux langages est un ensemble d'équations de la forme :

$$X_i = \sum_{j=1}^n a_{ij} X_j + b_i \quad \forall i \in \llbracket 1, n \rrbracket$$

où $\forall i, j \in \llbracket 1, n \rrbracket, a_{ij} \in \Sigma, b_i \in \Sigma^*$

Proposition Soit $\mathcal{A} = \{Q, \Sigma, T, q_0, A\}$ un automate fini. A partir des définitions, on peut donc représenter le langage d'un automate par un système d'équations aux langages. Elles associent à chaque L_q une équation de langages dérivant de l'automate.

Ainsi si un état q a des transitions vers d'autres états selon les lettres $a \in \Sigma$ et si q est un état acceptant alors l'équation associée à L_q est de la forme

$$\begin{cases} L_q = \sum_{(q,a,q') \in T} aL_{q'} + \{\varepsilon\} & \text{si } q \text{ est un état initial} \\ L_q = \sum_{(q,a,q') \in T} aL_{q'} & \text{sinon} \end{cases}$$

Il nous faut maintenant être capable de résoudre ces équations pour déterminer les L_q et ainsi le langage de l'automate.

7.2 Lemme d'Arden

Le Lemme d'Arden permet de résoudre de telles équations. Il fut démontré en 1961 par Dean N. Arden. Voici son énoncé :

Lemme (Arden) Soient A et B deux langages. Le langage

$$L = A^*B$$

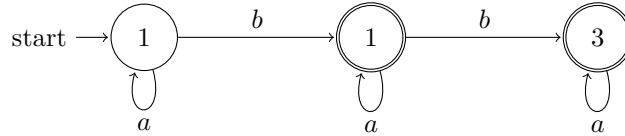
est le plus petit langage (pour l'inclusion ensembliste) qui est solution de l'équation

$$(E) : X = (AX) \cup B$$

De plus, si A ne contient pas le mot vide ε , A^*B est l'unique solution de cette équation.

Ainsi, puisque chacune des équations aux langages vues précédemment sont de la forme $L = AL + B$, on peut donc résoudre de tels systèmes et calculer explicitement le langage d'un automate.

Exemple Déterminons le langage de l'automate suivant :



Les langages d'arrivées vérifient donc le système d'équations aux langages suivant :

$$\begin{cases} L_1 = L_1a + \varepsilon \\ L_2 = L_1b + L_2a \\ L_3 = L_2b + L_3b \end{cases}$$

Que l'on résout progressivement grâce au Lemme d'Arden :

$$\begin{cases} L_1 = \varepsilon a^* = a^* \\ L_2 = L_2a + a^*b \\ L_3 = L_3a + L_2b \end{cases} \iff \begin{cases} L_1 = a^* \\ L_2 = a^*ba^* \\ L_3 = L_3a + a^*ba^* \end{cases} \iff \begin{cases} L_1 = a^* \\ L_2 = a^*ba^* \\ L_3 = a^*ba^*ba^* \end{cases}$$

D'où :

$$\begin{aligned} L(\mathcal{A}) &= a^*ba^* + a^*ba^*ba^* \\ &= a^*ba^*(\varepsilon + ba^*) \end{aligned}$$

Donc \mathcal{A} reconnaît les mots qui contiennent 1 ou 2 b.

Cet algorithme de résolution est très utile en terme pratique mais il apporte aussi une précision supplémentaire sur les langages automatiques. En effet, à partir d'un automate (i.e langage automatique), on peut écrire le langage reconnu comme une expression régulière. D'où le résultat suivant :

Propriété (Langages Automatiques) . Tout langage automatique peut être décrit par une expression régulière.

On en déduit donc le théorème de Kleene établissant définitivement le lien entre les langages réguliers et automatique :

Théorème (Kleene) . Les langages réguliers sont les mêmes que les langages automatiques.

Chapitre 8

Langages Algébriques et Automates à Piles

Contents

8.1	Grammaires Formelles	28
8.1.1	Contexte et définition	28
8.1.2	Réécriture d'un mot et langages algébriques	29
8.1.3	Arbre de dérivation d'un mot	30

Précédemment, nous avons vu que les langages automatiques sont de très bonnes propriétés. Ils sont stables pour la plupart des opérations définies sur les langages. De plus, le théorème de Kleene nous a permis d'établir le lien direct entre langages automatiques et langages réguliers. La simplicité de la représentation sagittale des automates permet de les implémenter facilement algorithmiquement. Ils sont, de plus, faciles à manipuler à la main et permettent de rapidement "voir" les langages reconnus.

Cependant, cette simplicité a un certain coût, celui de ne pas pouvoir reconnaître des langages "compliqués", notamment ceux où il faut "compter" les lettres. Un automate fini déterministe ne peut donc pas reconnaître le langage composé d'autant de a que de b . On va donc chercher à introduire une nouvelle théorie, celle des **grammaires formelles** qui nous permettra de reconnaître de tels langages.

8.1 Grammaires Formelles

8.1.1 Contexte et définition

Les grammaires formelles ont initialement été développées par des linguistes, notamment Noam Chomsky en 1955. L'objectif était de développer une méthode systématique de traduction entre différentes langues. Ils se sont alors heurtés au problème des mêmes mots qui admettent plusieurs traductions en fonction du contexte de la phrase et n'ont pas pu aboutir leur oeuvre.

Or en informatique, pour l'étude de la syntaxe de langages de programmation, le problème du contexte ne se pose pas. Leur théorie a donc été récupérée pour la vérification syntaxique.

L'idée est donc de représenter un langage **récurivement** par un ensemble de règles de production composées d'un axiome de départ et de différentes règles de productions ou de réécriture. Nous utilisons souvent cette approche pour la gestion de types en Caml en définissant des types récurivement.

Définition (Grammaire Formelle) . Une grammaire formelle est un quadruplet

$$G = (\Sigma, V, S, P)$$

où

- Σ est un **alphabet terminal** dont chaque élément ne peut se réécrire plus simplement.
- V est l'**alphabet auxiliaire** (disjoint de Σ) composé de variables, qui ne peuvent pas non plus se réécrire.
- S est la variable de départ, appelé axiome.
- P est un ensemble de règles dites **de production** ou de réécriture du type

$$X \longrightarrow w \quad X \in V \text{ et } w \in (V \cup \Sigma)^*$$

Par convention, on notera toujours les variables en majuscule et les éléments terminaux en minuscules. En pratique, on regroupera plusieurs réécritures d'une même variable sur la même ligne en les séparant par des barres verticales de la forme :

$$X \longrightarrow w_1 | w_2 | \dots | w_p \iff \begin{cases} X \longrightarrow w_1 \\ X \longrightarrow w_2 \\ \vdots \\ W \longrightarrow w_p \end{cases}$$

8.1.2 Réécriture d'un mot et langages algébriques

L'objectif d'une grammaire formelle, vous l'aurez compris, est de réécrire un mot récursivement jusqu'à arriver à des éléments terminaux.

Définition (Réécriture d'un mot) . Soit $G = (\Sigma, V, S, P)$ une grammaire formelle. Soient $u, v \in (V \cup \Sigma)^*$ deux mots. On dit que u **peut se réécrire en v en une étape** et on note :

$$u \vdash v$$

si il existe des décompositions de u et v en

$$u = u_1 X u_2 \text{ et } v = u_1 w u_2$$

et que G contient la règle de production :

$$X \longrightarrow w$$

Plus généralement, on peut définir la réécriture en plusieurs étapes de la forme :

Définition (Réécriture (2)) . Soit $G = (\Sigma, V, S, P)$ une grammaire formelle. Soient $u, v \in (V \cup \Sigma)^*$ deux mots. On dit que u **peut se réécrire en v** ou que v **dérive en u** en un nombre quelconque de fois si il existe $u_1, \dots, u_p \in (V \cup \Sigma)^*$ tels que

$$u \vdash u_1 \vdash u_2 \vdash \dots \vdash u_p \vdash v$$

On note alors

$$u \vdash^* v$$

On peut maintenant définir les langages engendrés par des grammaires formelles et les langages algébriques, le coeur de ce chapitre.

Définition (Langage Engendré) . La **langage engendré** par une grammaire formelle $G = (\Sigma, V, S, P)$ est l'ensemble des mots de Σ^* qui dérivent de l'axiome S en un nombre quelconque d'étapes. On le note, comme pour les automates, en $L(G)$.

Définition (Langage Algébrique) . Un langage engendré par une grammaire est appelé **langage algébrique**.

8.1.3 Arbre de dérivation d'un mot

Définition (Arbre de dérivation d'un mot) . Soit $G = (\Sigma, V, S, P)$ une grammaire formelle. On appelle l'arbre de dérivation de $w \in \Sigma^*$ l'arbre dont :

- La racine est S
- Tous les sommets intérieurs appartiennent à V
- Toutes les feuilles appartiennent à $\Sigma \cup \{\varepsilon\}$
- Si un sommet intérieur X a pour fils X_1, \dots, X_p alors la règle

$$X \longrightarrow X_1 | \dots | X_p \in P$$

- Le mot obtenu en visitant les feuilles de l'arbre par un parcours profondeur préfixe de l'arbre est un mot de $L(G)$

Définition (Grammaire Ambiguë) . Soit G une grammaire. On dit que G est ambiguë s'il existe un mot $w \in L(G)$ possédant deux arbres de dérivation différents.

En pratique une telle grammaire est pas très utilisée. En effet, en informatique, il ne serait pas très pratique de pouvoir compiler un code en deux expressions différentes d'un autre langage. On ne saurait pas laquelle choisir. Il faut que la dérivation puisse se faire de façon unique.