Formaliser la syntaxe

max.silberztein@univ-fcomte.fr

Les grammaires formelles

- Comment décrire mathématiquement les ensembles infinis de séquences de mots

- Comment savoir si une séquence appartient à un langage infini ?

- Linguistes (ex. Chomsky) : décrire les langues
- Mathématiciens (ex. Schützenberger) : décrire les ensembles infinis
- Informaticiens (ex. Gödel) : décrire ce qu'on peut calculer

Langages, grammaires et machines

- Alphabet et lettres :
- Un alphabet est un ensemble fini d'éléments. Ses éléments sont les lettres de l'alphabet

- Ex. A = { a, e, i, o, u } a 5 lettres

- Ex. A_F = { a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z } a 26 lettres

Mots et langages

- A partir d'un alphabet donné, on peut construire des suites constituées de lettres : les mots

- Un langage est un ensemble de mots

- Ex. { aime, aimes, aimons, aimez, aiment }
- Ex. { mais, ou, et, donc, or, ni, car }
- Ex. { dès, ès, grès, près, très }

Vocabulaire et phrases

- Niveau orthographique : Lettres → Alphabet → Mots → Langages
- Niveau syntaxique : ALU → Vocabulaire → Phrases → Langages
- Ex. V = { chat, chien, le, voit }
- Ex. L = { "le chat le voit", "le chat voit", "le chat voit le chien", "le chat voit le chat", "le chien le voit", "le chien voit", "le chien voit le chat", "le chien voit le chien" }

Mot vide, langage vide, singleton du mot vide

- Le mot vide est constitué de 0 lettre (orthographe) ; La phrase vide est constituée de 0 ALU (syntaxe)
  - Les linguistes et les mathématiciens le notent ε
  - Les informaticiens le notent ""
  - Dans NooJ, on le note <E>

- Le langage vide ne contient aucun mot, ex. { } ou Ø
- Le singleton { <E> } contient un élément : le mot vide

Les grammaires formelles

- Une **grammaire formelle** est un ensemble de règles qui permet de représenter les phrases d'un langage

- Il existe de nombreux types de grammaires formelles, plus ou moins puissantes : les grammaires XFST, les grammaires GPSG, les grammaires LFG, les grammaires HPSG, les grammaires NooJ...

- Toutes ces grammaires sont des variantes, équivalentes à des **grammaires génératives**

Grammaires génératives

- Introduites par Schützenberger et Chomsky (1957)
- Une **grammaire générative** est un ensemble de **règles de réécriture** ; chaque règle de réécriture est constituée de deux membres : le membre gauche se réécrivant en membre droit. Une grammaire générative contient un **symbole de départ**, ex. :

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

Grammaires génératives

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

- Chaque membre gauche et droit d'une règle de réécriture peut contenir des **symboles auxiliaires** et des **symboles terminaux**

Dérivation (et machine)

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

Dérivation (et machine)

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE => GN voit GN

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

Dérivation (et machine)

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE => GN voit GN => <u>le NOM</u> voit GN

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

Dérivation (et machine)

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE => GN voit GN => le NOM voit GN => le NOM voit un NOM

PHRASE → **GN** voit **GN**

GN → le **NOM**

GN → un **NOM**

NOM → chat

NOM → chien

Dérivation (et machine)

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE => GN voit GN => le NOM voit GN => le NOM voit un NOM => le NOM voit un chat

Dérivation (et machine)

PHRASE → **GN** voit **GN**
GN → le **NOM**
GN → un **NOM**
NOM → chat
NOM → chien

- En partant du texte PHRASE, on effectue n'importe quelle réécriture, c'est-à-dire qu'on remplace n'importe quel membre gauche d'une règle par son membre droit, jusqu'à épuisement des possibilités de réécriture. Par exemple :

PHRASE
=> GN voit GN
=> le NOM voit GN
=> le NOM voit un NOM
=> le NOM voit un chat
=> le chien voit un chat

Il n’y a plus de possibilité de remplacement, donc la phrase *le chien voit un chat* fait partie du langage décrit par la grammaire.

Exercice

- Trouver d'autres dérivations pour la grammaire :

PHRASE → **GN** voit **GN**
GN → le **NOM**
GN → un **NOM**
NOM → chat
NOM → chien

- Combien y a-t-il de phrases dans le langage représenté par cette grammaire ?

Dérivation et langage

- Si on calcule toutes les dérivations possibles pour la grammaire précédente, on obtient le langage correspondant, *i.e.*, l'ensemble des 16 phrases suivantes :

{ "un chat voit un chat", "un chat voit un chien", "un chat voit le chat", "un chat voit le chien", "le chat voit un chat", "le chat voit un chien", "le chat voit le chat", "le chat voit le chien", "un chien voit un chat", "un chien voit un chien", "un chien voit le chat", "un chien voit le chien", "le chien voit un chat", "le chien voit un chien", "le chien voit le chat", "le chien voit le chien" }

Hiérarchie de Chomsky-Schützenberger (1957)

- Quatre types de grammaires de réécriture :

- Grammaires régulières, grammaires rationnelles (Type 3)
- Grammaires hors contexte, grammaires algébriques, grammaires de Chomsky (Type 2)
- Grammaires contextuelles (Type 1)
- Grammaires non restreintes (Type 0)

Grammaires régulières (type 3)

- Un et un seul symbole auxiliaire dans les membres à gauche des règles de réécriture.
- Les membres droits contiennent :

Soit <E>, ex.	**GN** → <E>
Soit une seule ALU, ex.	**GN** → Luc
Soit une ALU + un symbole auxiliaire, ex. :	**GN** → le **NOM**

Grammaires régulières (type 3)

- La grammaire suivante n'est pas une grammaire régulière :

PHRASE → **GN** voit **GN**
GN → le **NOM**
GN → un **NOM**
NOM → chat
NOM → chien

Grammaires régulières (type 3)

- La grammaire suivante est une grammaire régulière :

PHRASE → le **SUITE**

PHRASE → un **SUITE**

SUITE → chat **GVERBE**

SUITE → chien **GVERBE**

GVERBE → voit **OBJET**

OBJET → le **NOM2**

OBJET → un **NOM2**

NOM2 → chat

NOM2 → chien

Exercice : construire une grammaire régulière

- sur l'alphabet { a, b, c } qui représente tous les mots qui commencent par un a

Exercice : construire une grammaire régulière

- sur l'alphabet { a, b, c } qui représente tous les mots qui commencent par un a

MOT → a **SUITE**
SUITE → a **SUITE**
SUITE → b **SUITE**
SUITE → c **SUITE**
SUITE → <E>

Exercice : construire une grammaire régulière

- sur l’alphabet { a, b, c } qui représente tous les mots qui contiennent un et un seul b

Exercice : construire une grammaire régulière

- sur l'alphabet { a, b, c } qui représente tous les mots qui contiennent un et un seul b

MOT → a **MOT**
MOT → c **MOT**
MOT → b **OK**
OK → a **OK**
OK → c **OK**
OK → <E>

Exercice : construire une grammaire régulière

- sur le vocabulaire { Luc, voit, un, chat } qui représente toutes les phrases correctes

Exercice : construire une grammaire régulière

- sur le vocabulaire { Luc, voit, un, chat } qui représente toutes les phrases correctes

PHRASE → Luc **GVERBE**
PHRASE → un **NOMGVERBE**
NOMGVERBE → chat **GVERBE**
GVERBE → voit **COMPLEMENT**
COMPLEMENT → Luc
COMPLEMENT → un **NOM**
NOM → chat
COMPLEMENT → <E>

Grammaires hors contexte (type 2)

- Un et un seul symbole auxiliaire dans les membres gauches des règles de réécriture

- Les membres droits contiennent n'importe quelle séquence d'ALU, de symboles et/ou de mots vides, ex. :

GN → DET NOM de **NOM**

Grammaires hors contexte (Type 2)

- La grammaire suivante est une grammaire hors contexte :

PHRASE → **GN** voit **GN**
GN → le **NOM**
GN → un **NOM**
NOM → chat
NOM → chien

Exercice : construire une grammaire hors contexte

- sur le vocabulaire { Luc, donne, un, chat, chien, à Marie }
- Construire une grammaire hors-contexte qui représente tous les phrases correctes

Exercice : construire une grammaire hors contexte

- sur le vocabulaire { Luc, donne, un, chat, chien, à Marie }
- Construire une grammaire hors-contexte qui représente tous les phrases correctes

PHRASE → GN VERBE GN à GN

PHRASE → GN VERBE GN

GN → DET NOM

GN → NOMPROPRE

DET → un

NOM → chat

NOM → chien

VERBE → donne

NOMPROPRE → Luc

NOMPROPRE → Marie

Grammaires contextuelles (Type 1)

- Les grammaires contextuelles sont comme les grammaires hors contexte, mais les règles de réécriture peuvent contenir en plus un symbole auxiliaire identique dans les deux membres, ex.

PHRASE → PLURIEL PHRASE

PLURIEL PHRASE → PLURIEL GN voient **PLURIEL GN**

PLURIEL GN → PLURIEL les **NOMPLURIEL**

PLURIEL → <E>

NOMPLURIEL → chats

NOMPLURIEL → chiens

- Exercice : que reconnaît cette grammaire ?
- Ajouter à cette grammaire des règles pour qu'elle reconnaisse :

un chat voit un chien, un chien voit un chat

Grammaires non restreintes (type 0)

- Aucune restriction : les membres gauches et les membres droits des règles de réécriture peuvent contenir n'importe quelle séquence d'ALU, de symboles auxiliaires et de mots vides.

- Les grammaires non restreintes permettent de décrire n'importe quel langage, i.e. n'importe quel ensemble de séquences d'ALU

Grammaires non restreintes (Type 0)

- Dans les grammaires non restreintes, il n'y a aucune contrainte sur les membres gauches et droits, ex.

PHRASE → **GN1 VERBE GN2**

GN1 VERBE GN2 → **GN1** ne **VERBE** pas **GN2**

GN1 VERBE GN2 → C'est **GN1** qui **VERBE GN2**

GN1 VERBE GN2 → C'est **GN2** que **VERBE GN1**

GN1 → Luc

GN2 → Paul

VERBE → voit

- Exercice : que reconnaît cette grammaire ?
- Ajouter à cette grammaire des règles pour qu'elle reconnaisse

Luc ne va pas voir Paul, Paul va voir Luc,
C'est Luc qui peut voir Paul, C'est Paul que Luc ne peut pas voir