# Automates et logiques pour les mots sur un alphabet infini

Alexis Bès

LACL, Université Paris 12

# Plan

- Mots sur un alphabet infini
- Etat de l'art
- $\mathfrak{M}$-automates
- Propriétés des langages $\mathfrak{M}$-reconnaissables
- Caractérisations logiques
- Lien avec la notion de produit de structures
- Extension du modèle

# Motivations

- Modélisation/vérification
  - Système temporisé :

$$(s_0, t_0)(s_1, t_1)\ldots(s_n, t_n)\ldots$$

    $s_i \in \Sigma$ (ensemble fini des états)
    $t_i \in \mathbb{R}^+$ : durée
    Evolution du système : mot sur l'alphabet infini $(\Sigma \times \mathbb{R}^+)$
  - Gestion de processus

$$(a_0, p_0)(a_1, p_1)\ldots(a_n, p_n)\ldots$$

    $a_i \in \Sigma$ (ensemble fini des actions)
    $p_i \in \mathbb{N}$ (numéro du processus)
  - listes d'entiers, suite des valeurs d'une variable entière...
- Représentation de données semi-structurées (XML) : arbres étiquetés par des mots sur un alphabet fini

# Quels modèles de calcul ?

- Candidat raisonnable : automate fini
- Ambitions :
  - expressivité
  - propriétés de clôture
  - effectivité (et complexité raisonnable)
  - caractérisations : expressions régulières, algèbre, logique

# Quelques résultats

- 1969 : Eilenberg-Elgot-Shepherdson
- 1980 : Autebert-Beauquier-Boasson
- 1994 : Kaminski-Francez : automates à registre
- 2003 : Bouyer-Petit-Thérien
- 2004 : Neven-Schwentick-Vianu : pebble automata
- 2005 : Bojanczyk-David-Muscholl-Schwentick-Segoufin : logique $FO^2$ et *data automata*
- 2005 : Choffrut-Grigorieff
- 2005 : Demri-Lazic-Nowak : variante de LTL

# Automates à registres (Kaminski-Francez)

Automates finis non-déterministes munis d'un nombre fini (fixé) de registres qui permettent de stocker des symboles et d'effectuer des comparaisons avec le symbole courant.

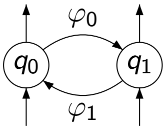
- Clos par $\cup, \cap, \cdot, \ ^*$
- Le problème du vide est décidable
- Pas clos par complémentation
  - $L = \{s_1 \dots s_n : \exists i \neq j \ s_i = s_j\}$ est reconnaissable
  - $\overline{L} = \{s_1 \dots s_n : \forall i \neq j \ s_i \neq s_j\}$ ne l'est pas
- Pas clos par opération miroir :
  - $L = \{s_1 \dots s_n : \forall i > 1 \ s_i \neq s_1\}$ est reconnaissable
  - $\tilde{L} = \{s_1 \dots s_n : \forall i < n \ s_i \neq s_n\}$ ne l'est pas
- le problème de l'universalité est indécidable (Neven & al)
- Les modèles déterministe et non-déterministe ne sont pas équivalents

# $\mathfrak{M}-$automates

- $\Sigma$ un alphabet (fini ou non)
- $\mathfrak{M} = (\Sigma; R_1, \ldots, R_n, =)$ une structure relationnelle
- $\mathfrak{M}-$automate : automate fini non-déterministe dont les transitions sont des formules du 1er ordre écrites dans le langage de $\mathfrak{M}$
- L'automate suit la transition étiquetée par $\varphi(x)$ si le symbole courant $s$ vérifie $\mathfrak{M} \models \varphi(s)$
- Langages $\mathfrak{M}-$*reconnaissables* : langages $L \subseteq \Sigma^*$ reconnaissables par un $\mathfrak{M}$-automate

# Exemple de langage $\mathfrak{M}$-reconnaissable

- $\mathfrak{M} = (\mathbb{N}; +, =)$
- $L$ : ensemble des mots sur $\mathbb{N}$ dont les symboles sont alternativement pairs et impairs
- Un $\mathfrak{M}$-automate pour $L$ :




avec

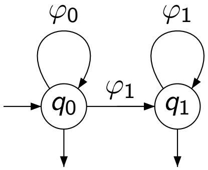
  - $\varphi_0(x)$ : $\exists y (x = y + y)$
  - $\varphi_1(x)$ : $\neg \varphi_0(x)$

# Version multi-têtes

- $\mathfrak{M}-$automate à $n$ têtes (synchrone) : reconnaît des relations $R \subseteq \Sigma^* \times \cdots \times \Sigma^*$
- Etant donnés $n$ mots $w_1, \ldots, w_n$ sur $\Sigma$, on complète avec des $\#$ de sorte à former un mot de $(\Sigma^n)^*$. On travaille avec une structure $\mathfrak{M}_\# = (\Sigma \cup \{\#\}; \ldots)$ définie à partir de $\mathfrak{M}$.

## Exemple avec $\mathfrak{M} = (\mathbb{N}; +, =)$ :

- $R = \{(u, v) \mid |u| \leq |v| \text{ et } u[i] < v[i] \text{ pour tout } i < |u|\}$
- Un $\mathfrak{M}-$automate pour $R$ :




avec

- $\varphi_0(x_1, x_2) :\ \exists y(x_2 = x_1 + y \wedge y + y \neq y)$
- $\varphi_1(x_1, x_2) :\ (x_1 = \# \wedge x_2 \neq \#)$

# Quelques propriétés

## Proposition

- Si $\mathfrak{M} = (\Sigma; (P_a)_{a\in\Sigma})$ avec $\Sigma$ fini, et $P_a(x)$ ssi $x$ porte un $a$, alors les relations $\mathfrak{M}$-reconnaissables sont les relations synchrones.
- Si $\mathfrak{M} = (\Sigma; \dots)$ avec $\Sigma$ infini alors le langage $\{aa \mid a \in \Sigma\}$ n'est pas $\mathfrak{M}$-reconnaissable.

## Proposition

La classe des langages $\mathfrak{M}$-reconnaissables est close par

- opérations booléennes
- projection

## Proposition

La décidabilité du problème "$L(\mathcal{A}) = \emptyset$ ?" est équivalente à celle de $FO(\mathfrak{M})$.

# Logique et automates

Trois formalismes classiques :

- Büchi-Elgot-Trakhtenbrot : logique du second-ordre monadique $(\mathbb{N}, <)$
- Eilenberg-Elgot-Shepherdson : logique du 1er ordre de $\mathcal{S} = (\Sigma^*; EqLength, \preceq, \{L_a\}_{a\in\Sigma})$ où
  - $EqLength(x, y)$ ssi $|x| = |y|$
  - $x \preceq y$ ssi $x$ préfixe de $y$
  - $L_a(x)$ ssi $a$ est la dernière lettre de $x$.
- Arithmétique de Büchi de base $k$ : théorie du 1er ordre de $(\mathbb{N}; +, V_k)$ où $V_k(x)$= plus grande puissance de $k$ qui divise $x$.

# Logique du second-ordre monadique (MSO)

- Extension de la logique du premier ordre
- variables du premier ordre $x, y, z, \ldots$ : (indices des) positions
- variables du second ordre $X, Y, Z, \ldots$ : ensembles de positions
- $x \in X$
- $x < y$, $Q_a(x)$ : la position $x$ porte un $a$

Exemple avec $\Sigma = \{a, b\}$ :

$$\exists x \exists y (x < y \wedge \neg \exists z (x < z < y) \wedge Q_a(x) \wedge Q_a(y))$$

définit le langage $\Sigma^* aa \Sigma^*$ ;

Théorème (Büchi-Elgot-Trakhtenbrot) :

Les langages définissables dans MSO sont les langages rationnels ;

# Généralisation de MSO

$MSO(\mathfrak{M})$ : on remplace les prédicats $Q_a(x)$ par des prédicats $\alpha_F(x)$ interprétés par “le symbole $s$ en $x-$ème position vérifie $\mathfrak{M} \models F(s)$”

## Exemples avec $\mathfrak{M} = (\mathbb{N}; +)$

- l'ensemble $L_1 \subseteq \mathbb{N}^*$ des mots ne contenant que des symboles pairs est $MSO(\mathfrak{M})-$définissable par $\forall y\ \alpha_F(y)$, où $F(x)$ : $\exists z(z + z = x)$
- l'ensemble $L_2 \subseteq \mathbb{N}^*$ des mots dont les symboles sont alternativement pairs et impairs est $MSO(\mathfrak{M})-$définissable par

$$\forall x \forall y((x < y \wedge \neg \exists z\ x < z < y) \rightarrow ((\alpha_F(x) \leftrightarrow \alpha_{\neg F}(y))))$$

## Théorème

Un langage $L \subseteq \Sigma^*$ est $\mathfrak{M}-$reconnaissable ssi il est $MSO(\mathfrak{M})$-définissable.

# Structures automatiques (Hodgson, Khoussainov-Nerode)

$\rightarrow$ Structures représentables (et décidables) par automate

**Définition**

$\mathfrak{N} = (N; R_1, \ldots, R_k)$ est *automatique* s'il existe $\Sigma$ fini et une injection $\mu : N \rightarrow \Sigma^*$ tels que les images de $N, R_1, \ldots, R_k$ par $\mu$ sont des relations synchrones.

**Exemple**

$(\mathbb{N}; +)$ est automatique.

**Théorème (Hodgson)**

Si $\mathfrak{N}$ est automatique alors :

1. Toute relation définissable dans $\mathfrak{N}$ est synchrone ;
2. $FO(\mathfrak{N})$ est décidable.

# Structures $\mathfrak{M}$−automatiques

## Définition

Soit $\mathfrak{M} = (\Sigma; \dots)$. La structure $\mathfrak{N} = (N; R_1, \dots, R_k)$ est $\mathfrak{M}$−*automatique* s'il existe une injection $\mu : N \to \Sigma^*$ telle que les images de $N, R_1, \dots, R_k$ par $\mu$ sont des relations $\mathfrak{M}$−reconnaissables

## Exemple

$(\mathbb{N} \setminus \{0\}; \times)$ est $(\mathbb{N}; +)$−automatique : considérer $\mu$ qui à tout $n = p_0^{n_0} p_1^{n_1} \dots p_k^{n_k}$ associe $\mu(n) = n_0 n_1 \dots n_k$.

## Théorème

Si $\mathfrak{N}$ est $\mathfrak{M}$−automatique alors :

1. Toute relation définissable dans $\mathfrak{N}$ est $\mathfrak{M}$−reconnaissable ;
2. $FO(\mathfrak{N})$ se réduit à $FO(\mathfrak{M})$.

# Eilenberg-Elgot-Shepherdson

Soit $\mathcal{S} = (\Sigma^*; EqLength, \preceq, \{L_a\}_{a \in \Sigma})$ où

- $EqLength(x, y)$ ssi $|x| = |y|$ ;
- $x \preceq y$ ssi $x$ préfixe de $y$ ;
- $L_a(x)$ ssi $a$ est la dernière lettre de $x$.

## Théorème (EES)

1. Pour $\Sigma$ fini et $|\Sigma| \geq 2$, les relations $R \subseteq \Sigma^n$ définissables dans $\mathcal{S}$ sont les relations synchrones.
2. $FO(\mathcal{S})$ est décidable.

# Et pour un alphabet $\Sigma$ infini ?

- *Dans le cas où $\Sigma$ est infini, à quel type d'automate correspondent les relations définissables dans $\mathcal{S} = (\Sigma^*; EqLength, \preceq, \{L_a\}_{a\in\Sigma})$ ?*
- Choffrut-Grigorieff : automates avec des contraintes qui sont des combinaisons booléennes de formules de la forme $x = y$ et $x = a$
- correspondent aux $\mathfrak{M}-$automates avec $\mathfrak{M} = (\Sigma; \{P_a\}_{a\in\Sigma})$.

# Généralisation

Pour $\mathfrak{M} = (\Sigma; R_1, \ldots, R_k)$, on définit

$$S_{\mathfrak{M}} = (\Sigma^*; EqLength, \preceq, A_{R_1}, \ldots, A_{R_k}, A_=)$$

où pour tout $i$, $A_{R_i}(x_1, \ldots, x_n)$ ssi il existe $w_1, \ldots, w_n \in \Sigma^*$ et $a_1, \ldots, a_n \in \Sigma$ tels que :

- $x_i = w_i a_i$ pour tout $i$ ;
- tous les $w_i$ ont la même longueur
- $(a_1, \ldots, a_n) \in R_i^{\mathfrak{M}}$.

## Théorème

- Pour toute relation $R \subseteq \Sigma^n$, $R$ est $\mathfrak{M}-$reconnaissable ssi $R$ est définissable dans $S_{\mathfrak{M}}$ ;
- $FO(S_{\mathfrak{M}})$ se réduit à $FO(\mathfrak{M})$.

# Un peu mieux

Pour $\mathfrak{N} = (\Sigma^*; \preceq, \dots)$, $MSO^{ch}(\mathfrak{N})$ : restriction de $MSO(\mathfrak{N})$ où les variables du second ordre sont interprétées comme des chaînes pour $\preceq$.

## Théorème (Kuske-Lohrey 2006)

Soient

$$\mathfrak{M} = (\Sigma; R_1, \dots, R_k)$$

et

$$\mathfrak{M}^* = (\Sigma^*; \preceq, A'_{R_1}, \dots, A'_{R_k}, A'_=)$$

où $A'_{R_i} = \{(ua_1, \dots, ua_n) \mid (a_1, \dots, a_n) \in R_i\}$.
Alors $MSO^{ch}(\mathfrak{M}^*)$ se réduit à $FO(\mathfrak{M})$.

## Théorème (Thomas-Bès)

$MSO^{ch}(\Sigma^*; EqLength, \preceq, A_{R_1}, \dots, A_{R_k}, A_=)$ se réduit à $FO(\mathfrak{M})$.

# $\mathfrak{M}-$automates et produits de structures

## Produits de structures en logique (Mostowski, Feferman-Vaught, Shelah...)

- $I$ un ensemble, $(\mathfrak{M}_i)_{i \in I}$ une famille de structures sur le même langage $\rightarrow$ définitions diverses de $\prod_{i \in I} \mathfrak{M}_i$
- Théorèmes de (dé)composition : la théorie du 1er ordre du produit se réduit à celle des facteurs et à celle d'une structure *Ind* liée à l'ensemble des parties de $I$

# Cas particulier intéressant de produit $\prod_{i \in I} \mathfrak{M}_i$

- Tous les $\mathfrak{M}_i$ sont égaux à $\mathfrak{M}_\# = (\Sigma \cup \{\#\}; \dots)$
- $I = \mathbb{N}$, et $Ind = (PartiesFinies(\mathbb{N}); \subseteq, \ll)$ où $X \ll Y$ ssi $X = \{x\}, Y = \{y\}$ avec $x < y$ (variante de $\mathrm{WMSO}(\mathbb{N}; <)$).
- Une *puissance faible généralisée* (Feferman-Vaught) de $\mathfrak{M}_\#$ relativement à $Ind$ est (presque) par définition une structure dont le domaine et les relations de base sont $MSO(\mathfrak{M})$-définissables, i.e. une structure $\mathfrak{M}-$automatique.

## Relecture des résultats précédents

- clôture des relations $\mathfrak{M}-$reconnaissables $\leftrightarrow$ clôture par définissabilité
- Si $\mathfrak{N}$ est $\mathfrak{M}-$automatique alors $FO(\mathfrak{N})$ se réduit à $FO(\mathfrak{M}) \leftrightarrow$ $FO(\mathfrak{N})$ se réduit à $FO(Ind)$ et $FO(\mathfrak{M})$

# Une extension du formalisme logique

- Le modèle manque d'expressivité : si $\mathfrak{M} = (\Sigma; \dots)$ avec $\Sigma$ infini alors $\{aa \mid a \in \Sigma\}$ n'est pas $\mathfrak{M}$-reconnaissable
- Idée : ajouter à $MSO(\mathfrak{M})$ des prédicats $\alpha_F(x_1, \dots, x_n)$ qui expriment des relations entre les symboles associés aux positions $x_1, \dots, x_n$ (exemple : égalité)
- Notation : $MSO^+(\mathfrak{L})$
- Problème : trop expressif $\rightarrow$ indécidabilité (même avec $FO$ et $\mathfrak{M} = (\mathbb{N}; =)$)

# Un fragment syntaxique décidable

### Définition

$MSO_R^+(\mathfrak{L})$ : fragment de $MSO^+(\mathfrak{L})$ constitué des énoncés

$$\exists x_1 \ldots \exists x_m \psi$$

où $\psi$ formule de $MSO^+(\mathfrak{L})$ où les $\alpha_F$ sont de la forme $\alpha_F(x_1, \ldots, x_m, y)$

### Proposition

Le problème de la satisfaisabilité pour $MSO_R^+(\mathfrak{M})$ se réduit à la décidabilité de $FO(\mathfrak{M})$

## Exemples avec $\mathfrak{M} = (\mathbb{N}; +)$

- $L_1$ : ens. mots $u$ sur $\mathbb{N}$ tels qu'un symbole $s \in \Sigma$ apparaît au moins deux fois dans $u$

  $$\exists x \exists y (x < y \wedge \alpha_F(x, y))$$

  où

  $$F(x_1, x_2) : x_1 = x_2.$$

- $L_2$ : ens. mots $u = s_0 \dots s_m$ tels qu'il existe $j$ avec $s_k \geq 2s_j$ pour tout $k > j$ :

  $$\exists x (\exists x' (x < x') \wedge \forall y (x < y \rightarrow \alpha_G(x, y))$$

  où

  $$G(x_1, x_2) : \exists z (x_2 = x_1 + x_1 + z)$$

  exprime $x_2 \geq 2x_1$.

## Exemple avec $\mathfrak{M} = (\Sigma^*; EqLength, \preceq, \{L_a\}_{a\in\Sigma})$

- $L$ : ens. des mots $w = s_0 \dots s_m$ sur $\Gamma = \Sigma^*$ tels que toutes les positions paires portent le même symbole, et toutes les positions impaires portent un symbole qui est préfixe de $s_0$

  $$\exists X[EvenPositions(X) \wedge$$
  $$\wedge \exists x \in X \; \forall y \in X \; \alpha_{F_1}(x,y) \wedge \exists z (\forall t \; \neg t < z \wedge \forall y \notin X \; \alpha_{F_2}(y,z))]$$

  où $EvenPositions(X)$ exprime que $X$ contient les positions paires et

  $$F_1(v_1, v_2): \; v_1 = v_2 \; ;$$
  $$F_2(v_1, v_2): \; v_1 \preceq v_2 .$$

# Perspectives

- Extension aux mots infinis, aux arbres
- Fragments de complexité raisonnable
- Version automates du résultat précédent
- Extensions décidables de MSO
- Nouvelles techniques de composition de structures

# Annexe - Quelques détails sur les produits de structures

## Produits de structures en logique

- $I$ un ensemble, $(\mathfrak{M}_i)_{i\in I}$ une famille de structures sur le même langage $\rightarrow$ définitions diverses de $\prod_{i\in I} \mathfrak{M}_i$
- Théorèmes de composition : la théorie du 1er ordre du produit se réduit à celle des facteurs et à celle d'une structure liée à l'ensemble des parties de $I$

## Puissance faible d'une structure (Mostowski)

- $\mathfrak{M} = (\Sigma \cup \{\#\}; R_1, \ldots, R_n)$ une structure relationnelle
- La puissance faible de $\mathfrak{M}$ relativement à $\mathbb{N}$ et $\#$ est la structure $\mathfrak{N} = (D; R_1^{\mathfrak{N}}, \ldots, R_k^{\mathfrak{N}})$ telle que
  - $D$ : fonctions $f : \mathbb{N} \rightarrow \Sigma \cup \{\#\}$ presque toujours égales à $\#$
  - $\mathfrak{N} \models R_i^{\mathfrak{N}}(f_1, \ldots, f_k)$ ssi $\mathfrak{M} \models R_i^{\mathfrak{N}}(f_1(i), \ldots, f_k(i))$ pour tout $i \in \mathbb{N}$

# Un exemple de puissance faible

- $\mathfrak{M} = (\mathbb{N}; +)$ où 0 est l'élément distingué
- $\mathfrak{N} = (D; +^{\mathfrak{N}})$
  - $D$ : fonctions $f : \mathbb{N} \to \mathbb{N}$ presque toujours nulles
  - $\mathfrak{N} \models +^{\mathfrak{N}}(f_1, f_2, f_3)$ ssi $f_1(i) + f_2(i) = f_3(i)$ pour tout $i$
  - $\mathfrak{N}$ est isomorphe à $(\mathbb{N} \setminus \{0\}; \times)$

Evaluation des formules dans $\mathfrak{N}$ :

- On a $\mathfrak{N} \models +^{\mathfrak{N}}(f_1, f_2, f_3)$ ssi $f_1(i) + f_2(i) = f_3(i)$ pour tout $i \in \mathbb{N}$
- ou encore : $(\mathcal{P}_f(\mathbb{N}); \subseteq) \models \forall Y \; X \subseteq Y$ avec $X = \{i \mid \mathfrak{M} \models f_1(i) + f_2(i) \neq f_3(i)\}$
- On a ainsi trouvé des formules $G$ et $\theta$ telles que $\mathfrak{N} \models +^{\mathfrak{N}}(f_1, f_2, f_3)$ ssi $(\mathcal{P}_f(\mathbb{N}); \subseteq) \models G(X)$ avec $X = \{i \mid \mathfrak{M} \models \theta(f_1(i), f_2(i), f_3(i))\}$
- $(G, \theta)$ : *suite de réduction* pour la formule $+^{\mathfrak{N}}(x, y, z)$

# Théorème de composition pour les puissances faibles

### Théorème (Mostowski)

- A toute formule $F(f_1, \ldots, f_k)$ dans $\mathfrak{N}$ on peut associer de manière effective une suite de réduction $(G, \theta_1, \ldots, \theta_n)$
- $FO(\mathfrak{N})$ se réduit à $FO(\mathfrak{M})$ et $FO(\mathcal{P}_f(\mathbb{N}); \subseteq)$.
- $FO(\mathcal{P}_f(\mathbb{N}); \subseteq)$ est décidable

### Exemple (suite)

$FO(\mathfrak{N})$ se réduit à $FO(\mathbb{N}; +)$ qui est décidable. Donc $FO(\mathfrak{N})$ (et $FO(\mathbb{N} \setminus \{0\}; \times)$) sont décidables.

# Puissances faibles généralisées (Feferman-Vaught)

On part de

$$\mathfrak{M} = (\Sigma \cup \{\#\}; R_1, \ldots, R_n)$$

et d'une structure

$$S = (\mathcal{P}_f(\mathbb{N}); \subseteq, S_1, \ldots, S_p))$$

*Puissance faible généralisée de* $\mathfrak{M}$ *relativement à* $S$ : toute structure $\mathfrak{N}$ de domaine $D$ et dont toutes les relations de base admettent une suite de réduction relativement à $M$ et $S$.

## Théorème (Feferman-Vaught)

- A toute formule $F(f_1, \ldots, f_k)$ dans $\mathfrak{N}$ on peut associer de manière effective une suite de réduction $(G, \theta_1, \ldots, \theta_n)$ ;
- $FO(\mathfrak{N})$ se réduit à $FO(\mathfrak{M})$ et $FO(S)$.

# Puissances faibles généralisées et $\mathfrak{M}$-automates

- Cas particulier intéressant : $S = (\mathcal{P}_f(\mathbb{N}); \subseteq, \ll))$ où $X \ll Y$ ssi $X = \{x\}$, $Y = \{y\}$, avec $x < y$. $FO(S)$ est une variante de $WMSO(\mathbb{N}; <)$.
- Soient $\mathfrak{M} = (\Sigma; \dots)$, et $\mathfrak{N}$ une puissance faible généralisée de $\mathfrak{M}_{\#}$ relativement à $S$.
- A tout mot $w \in \Sigma^*$ on fait correspondre (en complétant par des $\#$) un mot $c(w) \in (\Sigma \cup \{\#\})^\omega$, i.e un élément du domaine de $\mathfrak{N}$.

## Proposition

- Une relation $R$ est $\mathfrak{M}-$reconnaissable ssi $c(R)$ admet une suite de réduction relativement à $\mathfrak{M}$ et $S$
- $\mathfrak{N}$ est une structure $\mathfrak{M}$-automatique
- clôture des relations $\mathfrak{M}$-reconnaissables $\leftrightarrow$ clôture par définissabilité