Dérivation formelle de l'algorithme d'analyse syntaxique d'Earley par abstraction d'une sémantique des grammaires algébriques

Patrick Cousot

ENS

Journée de présentation des cursus en informatique

ENS Cachan

15 mai 2003

http://www.ens-cachan.fr/dptinfo/15mai.php

Grammaires algébriques

- Exemple :

  Grammaire : $A \to AA \mid a$

  formellement $\langle \{A\}, \{a\}, \{(A, AA), (A, a)\}, A \rangle$

- Grammaire

  $\langle N, T, R, A \rangle$

  ↳ axiome $A \in N$

  ↳ règles $R \subseteq N \times (N \cup T)^*$

  ↳ terminaux

  ↳ non-terminaux $N \cap T = \emptyset$

- Chaînes :

  X^* chaînes finies sur l'alphabet X (ε chaîne vide)

Points fixes

- $\langle P, \leqslant \rangle$, treillis complet (ensemble partiellement ordonné tel que tout sous-ensemble $X \subseteq P$ a une borne supérieure $\vee X$)
- $f \in P \to P$, croissante
- $\mathrm{lfp}\ f \triangleq \wedge \{x \mid f(x) \leqslant x\}$ est le plus petit point fixe de f — Tarski
- si f préserve le bornes sup. existantes $(f(\vee X) = \vee f(X))$

  $\mathrm{lfp}\ f = \bigvee_{n \in \mathbb{N}} f^n(\bot)$

  - $\bot \triangleq \vee \emptyset$ est l'infimum du treillis complet
  - $f^0(x) = x$
  - $f^{n+1}(x) = f \circ f^n(x)$, $n \in \mathbb{N}$

Exemple : sémantique de Schützenberger d'une grammaire algébrique (langage terminal engendré pour chaque non-terminal)

Exemple : $A \to AA \mid a$

$$S_G(X) = \{(A,a)\} \cup \{(A, \alpha_1\alpha_n) \mid (A,\alpha_1) \in X \wedge (A,\alpha_n) \in X\}$$

$X^0 = \emptyset$ -- itérés

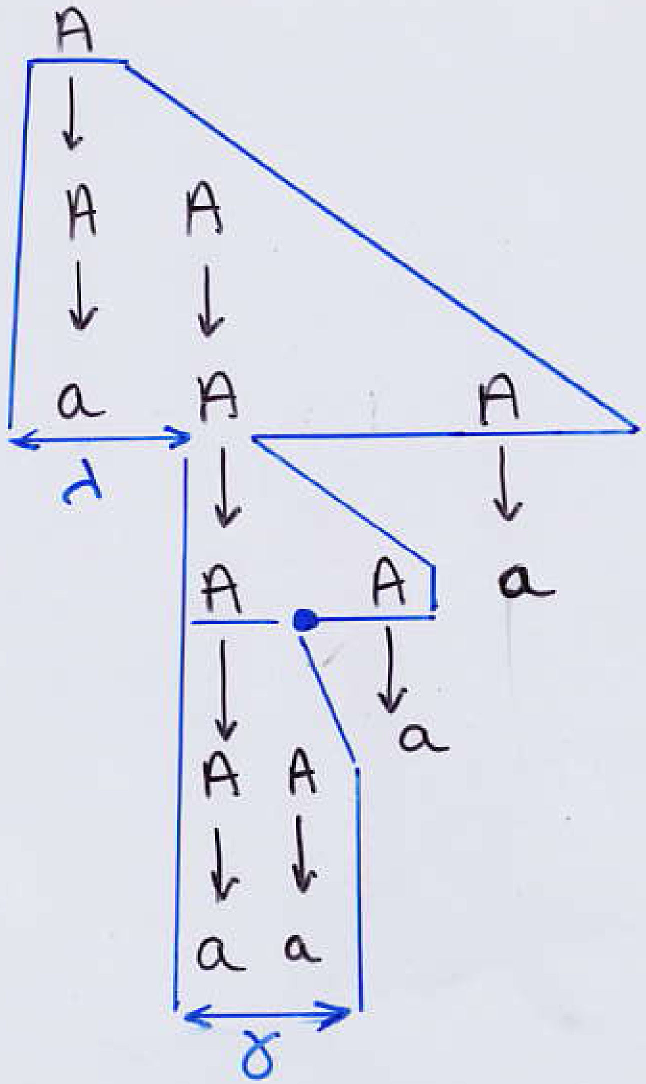
$X^1 = S_G(X^0) = \{(A,a)\}$

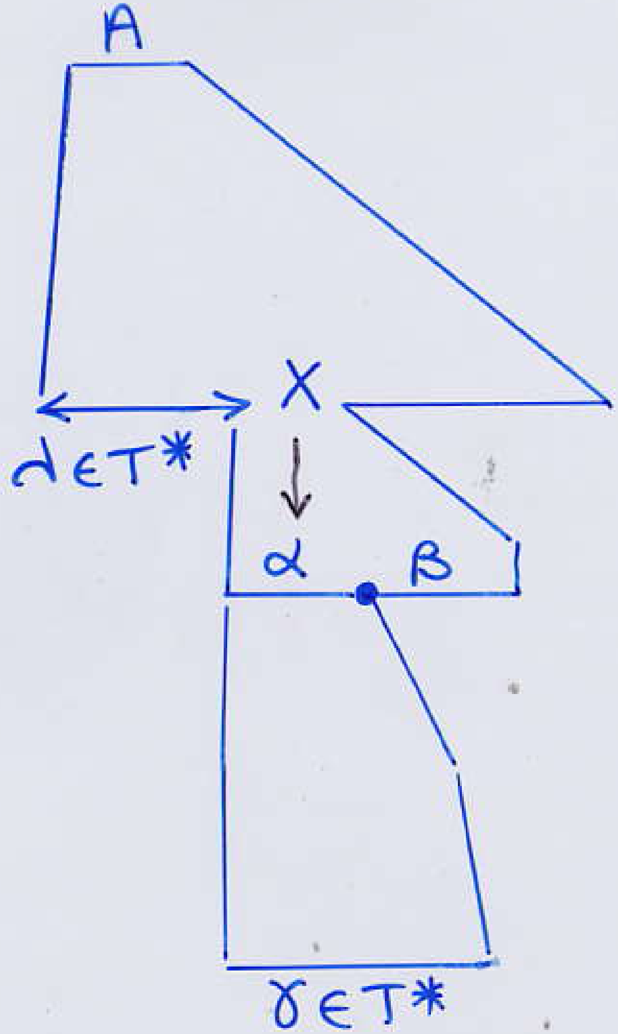
$X^n = \{(A,a^k) \mid 1 \leq k \leq 2^{n-1}\}$ -- hypothèse d'induction

$$\begin{aligned} X^{n+1} &= S_G(X^n) \\ &= \{(A,a)\} \cup \{(A, a^{k_1+k_2}) \mid 1 \leq k_1 + k_2 \leq 2^{n-1}\} \\ &= \{(A,a^k) \mid 1 \leq k \leq 2^n\} \end{aligned}$$

$$\mathrm{lfp}\ S_G = \bigcup_{n \geq 0} X^n = \{(A,a^k) \mid k \geq 1\}$$

Dérivation : exemple & notation





$X \to \alpha\beta \in R$

$[\lambda, X \to \alpha . \beta, \gamma]$ (item)

Sémantique de dérivation d'une grammaire

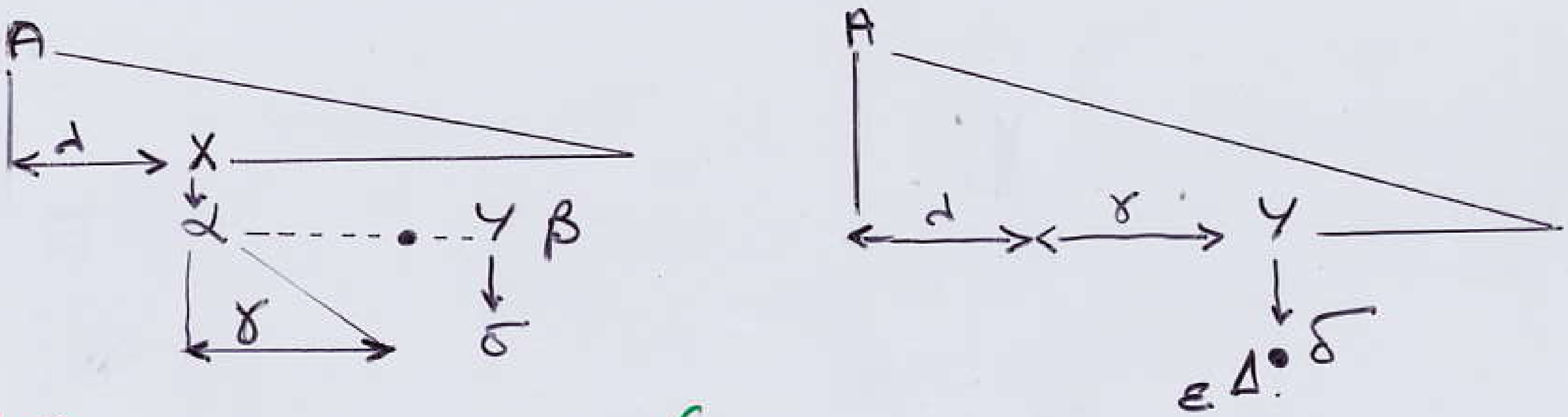
- Schéma d'axiome (initialisation)

$$[\varepsilon, A \rightarrow \bullet\beta, \varepsilon] \qquad \text{ssi } A \rightarrow \beta \in R$$

- Schémas de règles :

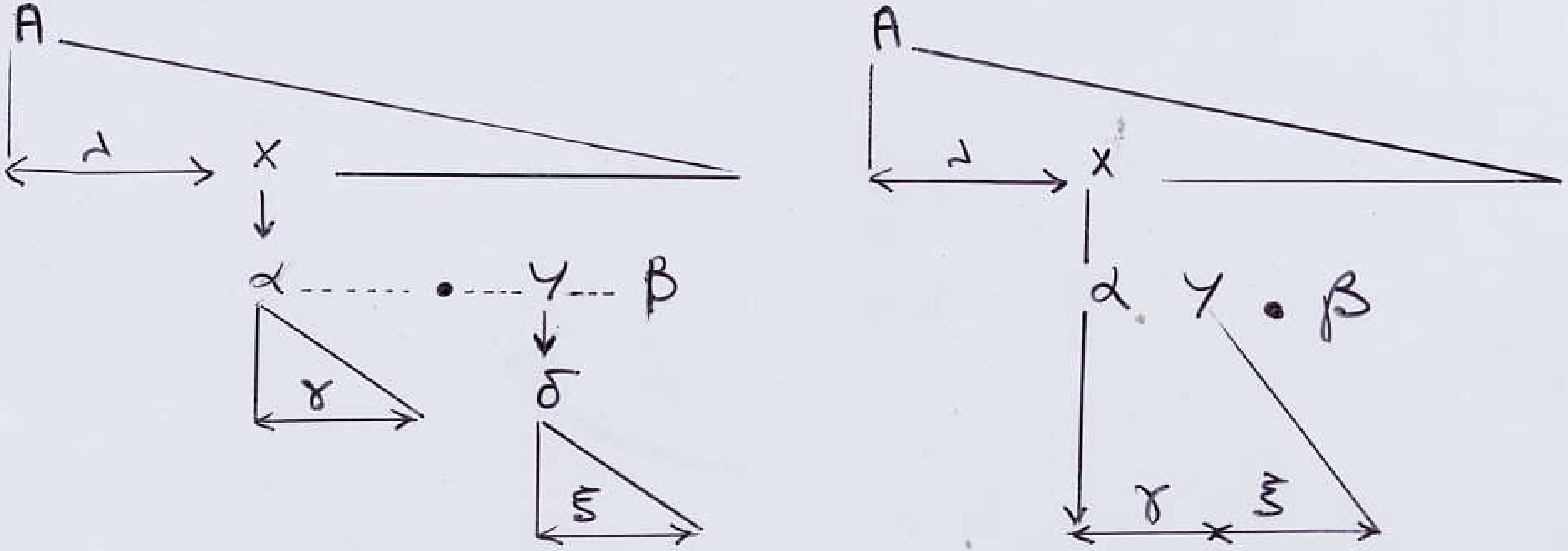
• Dérivation $(X \rightarrow \alpha Y \beta,\ Y \rightarrow \delta \in R)$

$$\frac{[\lambda, X \rightarrow \alpha \bullet Y \beta, \gamma]}{[\lambda\gamma, Y \rightarrow \bullet\delta, \varepsilon]}$$



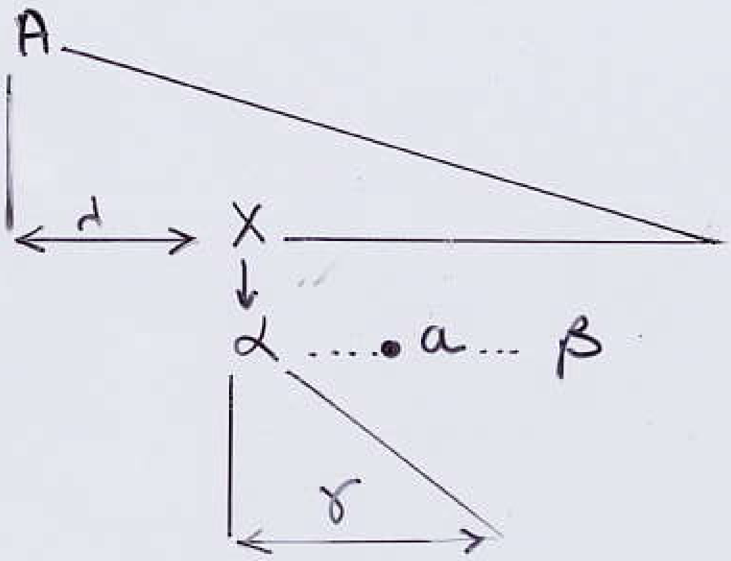
- Réduction $(X \to \alpha Y \beta,\ Y \to \gamma \in R)$:

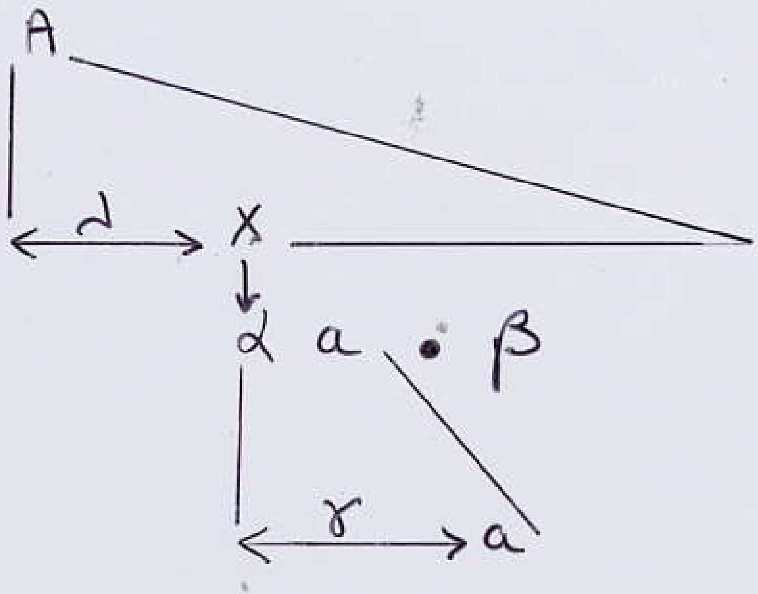
$$\frac{[\iota, X \to \alpha \bullet Y \beta, \gamma] \quad [\iota\gamma, Y \to \delta \bullet, \xi]}{[\iota, X \to \alpha Y \bullet \beta, \gamma\xi]}$$



- Avancement $(X \to \delta a \beta \in R,\ a \in T)$:

$$\frac{[\delta,\ X \longrightarrow \alpha \bullet a \beta,\ \gamma]}{[\delta,\ X \longrightarrow \alpha a \bullet \beta,\ \gamma a]}$$





Ensemble défini par des règles :

- $\mathcal{U}$ univers
- Règles $R = \left\{ \dfrac{P_i}{c_i} \;\middle|\; i \in \Delta \right\}$ Premisse $P_i \subseteq \mathcal{U}$ ($\emptyset$ pour axiome) Conclusion $c_i \in \mathcal{U}$
- $F_R(X) = \left\{ c \;\middle|\; \exists \dfrac{P}{c} \in R : P \subseteq X \right\}$

  (conclusions que l'on peut tirer de X par R)
- F_R est croissante sur le treillis complet $(\mathcal{P}(\mathcal{U}), \subseteq)$
- $\mathrm{lfp}\ F_R$ existe (Tarski)

  $= \bigcup_{n \geq 0} F_R^{\,n}(\emptyset)$ $\qquad f^0(x) = x \quad f^{n+1}(x) = f(f^n(x))$

  (conclusions successives que l'on peut tirer de R sans hypothèse initiale).

Sémantique de dérivation d'une grammaire algébrique

$G = \langle N, T, R, A \rangle$

$\mathrm{lfp}\ F_G$ où

$$F_G(X) = \{[\varepsilon, A \to \bullet\beta, \varepsilon] \mid A \to \beta \in R\}$$
$$\cup \{[\alpha\gamma, Y \to \bullet\delta, \varepsilon] \mid [\alpha, X \to \alpha \bullet Y\beta, \gamma] \in X \wedge Y \to \delta \in R\}$$
$$\cup \{[\alpha, X \to \alpha Y \bullet \beta, \gamma\xi] \mid [\alpha, X \to \alpha \bullet Y\beta, \gamma] \in X \wedge [\alpha\gamma, Y \to \delta\bullet, \xi] \in X\}$$
$$\cup \{[\alpha, X \to \alpha a \bullet \beta, \gamma a] \mid [\alpha, X \to \alpha \bullet a\beta, \gamma] \in X\}$$

Abstraction par une correspondance de Galois

- $(P, \leqslant)$ ensemble partiellement ordonné
- $(Q, \sqsubseteq)$ ensemble partiellement ordonné
- $\alpha \in P \to Q$, $\gamma \in Q \to P$ tels que :

$$\forall x \in P : \forall y \in Q : \alpha(x) \sqsubseteq y \iff x \leqslant \gamma(y)$$

$\leftrightarrow$ $\langle (P, \leqslant), (Q, \sqsubseteq), \alpha, \gamma \rangle$ est une correspondance de Galois.

Exemple :

$$f : A \to B$$

$$\alpha \in \wp(A) \to \wp(B)$$

$$\alpha(X) = \{f(x) \mid x \in X\}$$

$$\gamma \in \wp(B) \to \wp(A)$$

$$\gamma(Y) = \{x \in A \mid f(x) \in Y\}$$

Abstraction de point fixe par une correspondance de Galois

- $\langle P, \leq \rangle$ et $\langle Q, \sqsubseteq \rangle$, treillis complets
- $\langle P, \leq \rangle \underset{\alpha}{\overset{\gamma}{\leftrightarrows}} \langle Q, \sqsubseteq \rangle$ correspondance de Galois
- $f \in P \to P$, croissante
- $g \in Q \to Q$, croissante
- $\alpha \circ f = g \circ \alpha$

$$\Longrightarrow \alpha(\mathrm{lfp}\, f) = \mathrm{lfp}\, g$$

preuve

Si g(x) <=x alors alpha o gamma (x) <= x et g croissante implique g o alpha o gamma(x) <= x
donc alpha o f o gamma (x) <= x (car alpha o f = g o alpha)
et donc f o gamma(x) <= gamma(x) par def. correspondance de Galois.
Il vient {y|f(y) <= y} contient {gamma(x)|f(gamma(x)) <= gamma(x)} qui contient {gamma(x) | g(x) <= x}
et donc /\{y|f(y) <= y} <= /\{gamma(x) | g(x) <= x}
mais /\{gamma(x) | g(x) <= x}= gamma(/\{x| g(x) <= x}) (dans la correspondance de Galois, gamma preserve les intersections).
D'après Tarski, on lfp f <= gamma(lfp g)
soit alpha(lfp f) <= lfp g, par définition d'une correspondance de Galois.

– $g(\alpha(\mathrm{lfp}\ f))$
$= \alpha(f(\mathrm{lfp}\ f))$ -- $\alpha \circ f = g \circ \alpha$
$= \alpha(\mathrm{lfp}\ f)$ -- lfp f est un point fixe (le plus petit)
$\Rightarrow \mathrm{lfp}\ g \sqsubseteq \alpha(\mathrm{lfp}\ f)$ -- lfp g est le plus petit point fixe de g

– $\mathrm{lfp}\ g = \alpha(\mathrm{lfp}\ f)$ -- antisymétrie

$\square$

Abstraction de la sémantique de dérivation de $G = \langle N, T, R, A \rangle$ en la sémantique de Schützenberger (langage fini engendré pour chaque non-terminal)

$$\alpha_s(I) = \{\langle X, \gamma \rangle \mid \exists \lambda \in T^* : [\lambda, X \to \alpha \bullet, \gamma] \in I\}$$

$$\alpha_s(F_G(I)) =$$

--- -- 2 pages de calcul

$$S_G(\alpha_s(I))$$

$$S_G(L) = \{(X, \alpha) \mid X \to \alpha \in R \wedge \alpha \in T^*\}$$
$$\cup \{(X, \alpha_1 \ldots \alpha_n) \mid X \to X_1 \ldots X_n \in R \wedge \forall i \in [1, n] : (X_i = \alpha_i \in T) \vee ((X, \alpha_i) \in L)\}$$

$$\alpha_s(\mathrm{lfp}\, F_G) = \mathrm{lfp}\, S_G$$ sémantique de Schützenberger

Analyse syntaxique

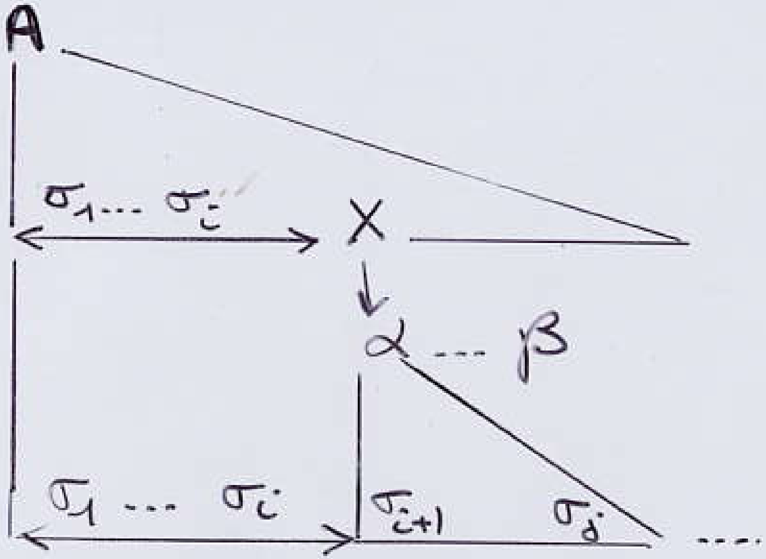
Etant donnée une grammaire $G = (N, T, R, A)$ et une phrase terminale $\sigma \in T^*$, répondre à la question :

- Est-ce que σ est engendrée par G ?
- i.e. $\sigma \in \mathrm{lfp}\ S_G$ (sémantique de Schützenberger)

(et déterminer sa structure syntaxique)

Abstraction de la sémantique de dérivation en l'algorithme d'analyse syntaxique d'Earley.

- $\sigma = \sigma_1 \dots \sigma_n \in T^*$, phrase à analyser
- $\alpha_E(I) = \{ (X \to \alpha \cdot \beta, i, j) \mid 0 \leq i \leq j \leq n \ \wedge$
  $[\sigma_1 \dots \sigma_i , X \to \alpha \cdot \beta , \sigma_{i+1} \dots \sigma_j] \in I \}$



Dérivation de l'algorithme de Earley

$\alpha_E(F_G(I)) =$

$\cdots$ — 2 pages de calculs

$E_G(\alpha_E(I))$

$$
\begin{aligned}
E_G(I) = & \{ (A \to \bullet \gamma, 0, 0) \mid A \to \gamma \in R \} \\
\cup & \{ (Y \to \bullet \gamma, j, j) \mid (X \to \alpha \bullet Y \beta, i, j) \in I \wedge Y \to \gamma \in R \} \\
\cup & \{ (X \to \alpha Y \bullet \beta, k, j) \mid (X \to \alpha \bullet Y \beta, k, i) \in I \wedge \\
& \qquad (Y \to \gamma \bullet, i, j) \in I \} \\
\cup & \{ (X \to \alpha \sigma_j \bullet \beta, i, j) \mid (X \to \alpha \bullet \sigma_j \beta, i, j-1) \in I \}
\end{aligned}
$$

$\mathrm{lfp}\, E_G$ est fini, donc calculable itérativement.

$\langle A \to \gamma \bullet, 0, n \rangle \in \mathrm{lfp}\, E_G$ est la réponse.

Correction de l'algorithme d'Earley

$$
\begin{array}{ll}
& (A \to \gamma., 0, n) \in \mathrm{lfp}\, E_G \\
\Longleftrightarrow & (A \to \gamma., 0, n) \in \alpha_E(\mathrm{lfp}\, F_G) \\
\Longleftrightarrow & (\varepsilon, A \to \gamma., \sigma) \in \mathrm{lfp}\, F_G \\
\Longleftrightarrow & (A, \sigma) \in \alpha_S(\mathrm{lfp}\, F_G) \\
\Longleftrightarrow & (A, \sigma) \in \mathrm{lfp}\, S_G
\end{array}
$$

-- sémantique de Schützenberger

i.e. la phrase σ est engendrée par la grammaire G.

Pour connaitre tous les détails :

P. Cousot & R. Cousot

Parsing as abstract interpretation of grammars

Theoret. Comput. Sci. 290 : 531 – 544, 2003

http://www.di.ens.fr/~cousot/papers/TCS03-parsing.shtml

— MERCI DE VOTRE ATTENTION —