



# Parseo y Generación de Código

12 de septiembre de 2019

## **Análisis sintáctico descendente Presentación del TP 1 (Otras técnicas de análisis sintáctico)**

Licenciatura en Informática con Orientación en Desarrollo de Software  
Universidad Nacional de Quilmes

# Análisis sintáctico descendente

# Análisis sintáctico

**Objetivo:** dada una gramática independiente del contexto  $G = (N, \Sigma, P, S)$ , analizar sintácticamente una cadena  $\alpha \in \Sigma^*$ . Es decir, decidir si  $\alpha \in L(G)$  y en tal caso dar una derivación para  $\alpha$ .

## Análisis sintáctico

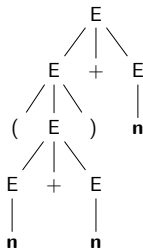
### Acciones asociadas a las producciones.

Generalmente es fácil modificar un método de análisis sintáctico para que cada producción dispare una acción. Esto puede servir, por ejemplo, para construir un AST.

**Ejemplo.** Si la gramática es  $G = (\{E\}, \{\mathbf{n}, +, (, )\}, P, E)$ :

$$E \rightarrow \mathbf{n} \mid E + E \mid (E)$$

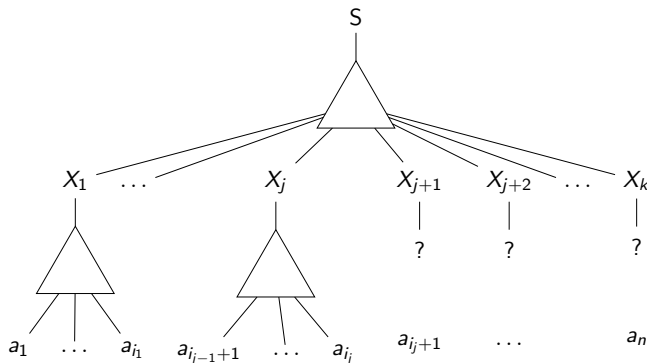
La cadena  $\alpha = (\mathbf{n} + \mathbf{n}) + \mathbf{n}$  da lugar al siguiente árbol de derivación, del que se puede extraer recursivamente un AST, ejecutando una acción en el recorrido *post-order*.



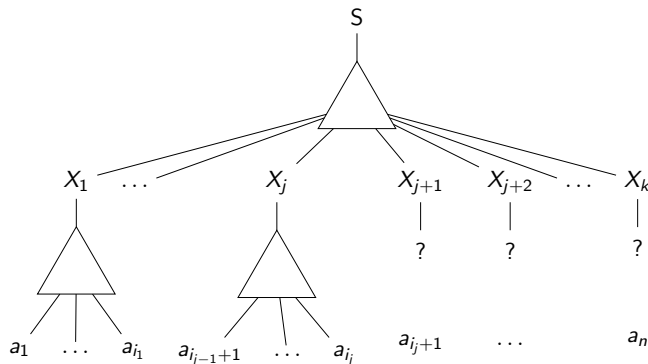
## Análisis sintáctico descendente

En el análisis sintáctico **descendente**, se empieza por la raíz  $S$  del árbol, y una cadena  $\alpha = a_1 \dots a_n$  que se quiere analizar sintácticamente.

El análisis trata de construir un árbol de derivación para  $S \Rightarrow^* a_1 \dots a_n$ . En cada paso del análisis, el árbol de derivación está parcialmente construido, y la cadena  $\alpha = a_1 \dots a_n$  ya fue parcialmente consumida.

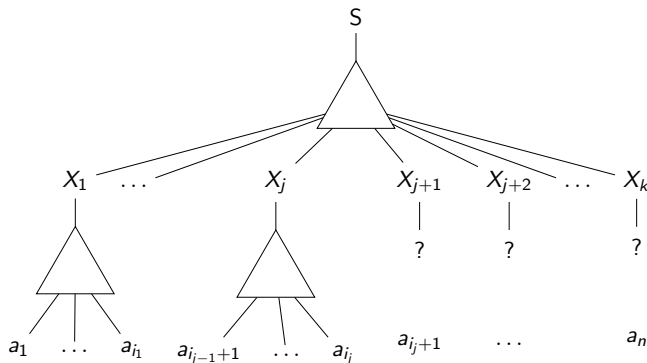


# Análisis sintáctico descendente



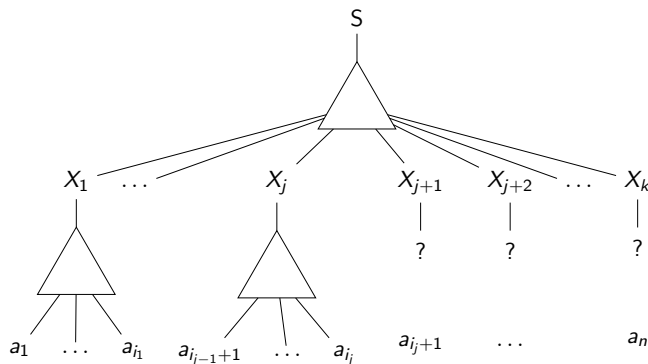
- **Caso 1:** si el símbolo  $X_{j+1}$  coincide con el símbolo terminal  $a_{i_j+1}$ , el nodo  $X_{j+1}$  pasa a ser una hoja del árbol de derivación. Se consume el símbolo  $a_{i_j+1}$  de la entrada, y se avanza a analizar el símbolo no terminal  $X_{j+2}$ .

# Análisis sintáctico descendente



- **Caso 2:** si el símbolo  $X_{j+1}$  es un símbolo terminal distinto de  $a_{i_j+1}$ , el árbol de derivación construido hasta el momento no sirve para la cadena  $\alpha = a_1 \dots a_n$ :
  - Puede ser culpa del usuario (la cadena no está en el lenguaje).
  - Puede ser culpa del parser (hay otra derivación para  $\alpha$ ).

# Análisis sintáctico descendente



- **Caso 3:** si el símbolo  $X_{j+1}$  es un símbolo no terminal, se debe expandir  $X_{j+1}$  agregándole  $r$  hijos, usando alguna de las producciones  $X_{j+1} \rightarrow Y_1 \dots Y_r$ , sin consumir ningún símbolo de la entrada. Se avanza a analizar el símbolo no terminal  $Y_1$  (si  $r \geq 1$ ) o el símbolo no terminal  $X_{j+2}$  (si  $r = 0$ ).



# Análisis sintáctico descendente predictivo

Como primera aproximación al análisis sintáctico descendente, supongamos que tenemos un “oráculo” que en todo momento puede predecir cuál es la producción correcta que se debe elegir para conseguir una derivación de la cadena de entrada.

# Análisis sintáctico descendente predictivo

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$  y una cadena  $\alpha$ .

*Salida:* Una derivación más a la izquierda  $S \Rightarrow^* \alpha$   
si es que existe.

*pila* := [S]

**while** *pila* no es vacía

$X = \text{pila.pop}()$

**if**  $X$  es un símbolo no terminal

        Elegir **usando un oráculo** una producción  $X \rightarrow Y_1 \dots Y_k$ .

        Agregar  $X \rightarrow Y_1 \dots Y_k$  a la derivación.

        Apilar  $Y_1, \dots, Y_k$ , con  $Y_1$  en el tope.

**elseif**  $X$  es un símbolo terminal y coincide  
        con el próximo símbolo de la entrada  
        Consumir un símbolo de la entrada.

**else**

**return** Falla:  $\alpha$  no está en  $L(G)$ .

**end**

**end**

**if** toda la entrada fue consumida

**return** Éxito:  $S \Rightarrow^* \alpha$  con la derivación construida.

**else**

**return** Falla:  $\alpha$  no está en  $L(G)$ .

# Análisis sintáctico descendente predictivo

**Observación.** ¿Qué pasa cuando se acaba la entrada? Por ejemplo, la cadena  $\epsilon$  con la gramática  $S \rightarrow a$ .

Una práctica común para simplificar la presentación de los algoritmos es agregar un símbolo terminador al final de la entrada, es decir:

- ▶ Extender la gramática con un nuevo símbolo inicial  $S'$ , un nuevo símbolo terminal  $\$$  (el terminador) y una producción  $S' \rightarrow S\$$ .
- ▶ Analizar sintácticamente la cadena  $\alpha\$$  en lugar de la cadena  $\alpha$ .

# Análisis sintáctico descendente predictivo

**Ejercicio.** Dibujar la evolución del estado de la pila y de la entrada para un análisis sintáctico descendente de la cadena  $(\mathbf{a} \Rightarrow \mathbf{a}) \Rightarrow \mathbf{a}$  con la gramática  $G = (\{T, U\}, \{\mathbf{a}, \Rightarrow, (, )\}, P, T)$ , suponiendo que se dispone de un oráculo que elige siempre la producción correcta. Las producciones son:

$$\begin{array}{lcl} T & \rightarrow & U \\ & | & U \Rightarrow T \\ U & \rightarrow & \mathbf{a} \\ & | & (T) \end{array}$$

# Análisis sintáctico predictivo con *backtracking*

El algoritmo se puede adaptar para que haga *backtracking*, es decir, para que en caso de falla vuelva hasta el último punto en el que se eligió una producción para elegir otra alternativa.

**function** analizar\_bt( $G, \pi, \alpha$ )

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$ , una pila de símbolos  $\pi = Z_1 \dots Z_r$ , y una cadena  $\alpha$ .

*Salida:* Una derivación más a la izquierda  $Z_1 \dots Z_r \Rightarrow^* \alpha$ ,  
en caso de que dicha derivación exista y el algoritmo termine.

**case**  $\pi = \epsilon$

**if**  $\alpha = \epsilon$  **return** Éxito:  $\epsilon \Rightarrow^* \epsilon$ .

**else return** Falla.

**case**  $\pi = X\pi'$

**if**  $X$  es un símbolo no terminal

**foreach** producción  $X \rightarrow Y_1 \dots Y_k$

$resultado := \text{analizar\_bt}(G, Y_1 \dots Y_k \pi', \alpha)$

**if**  $resultado = \text{Éxito: } Y_1 \dots Y_k \pi' \Rightarrow^* \alpha$

**return** Éxito:  $X\pi' \Rightarrow Y_1 \dots Y_k \pi' \Rightarrow^* \alpha$

**end**

**end**

**return** Falla.

**elseif**  $\alpha = X\alpha'$

$resultado := \text{analizar\_bt}(G, \pi', \alpha')$

**if**  $resultado = \text{Éxito: } \pi' \Rightarrow^* \alpha'$

**return** Éxito:  $X\pi' \Rightarrow^* X\alpha'$ .

**else**

**return** Falla.

**end**

**else**

**return** Falla.

## Análisis sintáctico predictivo con *backtracking*

**Nota importante:** el algoritmo de análisis sintáctico predictivo con *backtracking* no funciona para cualquier gramática.

Por ejemplo, en general el algoritmo puede no terminar si la gramática tiene recursión a izquierda, es decir un símbolo  $A$  tal que  $A \Rightarrow^+ A\alpha$ .

Se pueden dar condiciones suficientes para que el algoritmo termine; por ejemplo, que la gramática sea  $LL(k)$ .

Además, aun si termina, puede tardar tiempo exponencial.

# Análisis sintáctico predictivo con *backtracking*

**Ejercicio.** Aplicar el algoritmo de análisis sintáctico predictivo con *backtracking* para las cadenas  $a \bullet a$  y  $a \bullet$  con las siguientes gramáticas.

1.  $G_1 = (\{S\}, \{a, \bullet\}, P, S)$  con las siguientes producciones:

$$S \rightarrow a \mid S \bullet a$$

2.  $G_2 = (\{S, S'\}, \{a, \bullet\}, P, S)$  con las siguientes producciones:

$$\begin{array}{lcl} S & \rightarrow & aS' \\ S' & \rightarrow & \epsilon \mid \bullet aS' \end{array}$$

## Análisis sintáctico predictivo con *backtracking*

**Ejercicio.** Aplicar el algoritmo de análisis sintáctico predictivo con *backtracking* para la cadena *aacc* con la siguiente gramática  $G = (\{S\}, \{a, b, c\}, P, S)$ :

$$S \rightarrow \epsilon \mid aSb \mid aSc$$



# Análisis sintáctico LL(1)

En las técnicas que vimos:

- ▶ El análisis con oráculo supone que siempre se predice la producción correcta.
- ▶ El análisis con *backtracking* prueba todas las alternativas posibles. En el peor caso no termina, y aun si termina tiene complejidad exponencial.

La técnica que veremos a continuación es el análisis LL(1):

- ▶ En general, el análisis sintáctico LL( $k$ ) predice la producción correcta mirando los siguientes  $k$  símbolos de la entrada.

# Análisis sintáctico LL(1)

**Ejercicio.** Hacer análisis sintáctico descendente de la expresión  $(n + n)/n$  con la siguiente gramática:

$G = (\{E, E', T, T', F\}, \{n, +, -, *, /, (, )\}, P, E)$ :

$$E \rightarrow TE'$$

$$E' \rightarrow \epsilon \mid +TE' \mid -TE'$$

$$T \rightarrow FT'$$

$$T' \rightarrow \epsilon \mid *FT' \mid /FT'$$

$$F \rightarrow n \mid (E)$$

Convencerse de que el oráculo podría elegir la producción siguiente usando únicamente un token de *lookahead*.

# Análisis sintáctico LL(1)

Empezaremos viendo el caso particular de los parsers LL(1).

Necesitamos dos definiciones previas:

## ► Primeros (FIRST).

Dada una cadena de símbolos terminales y no terminales  $\alpha \in (\Sigma \cup N)^*$ , el conjunto  $\text{FIRST}(\alpha) \subseteq \Sigma \cup \{\epsilon\}$  de los “primeros” de  $\alpha$  es un conjunto de símbolos terminales (y posiblemente  $\epsilon$ ) definido como sigue:

$$\text{FIRST}(\alpha) = \{a \in \Sigma \mid \alpha \Rightarrow^* a\beta \text{ para } \beta \in (N \cup \Sigma)^*\} \cup \\ \text{if } \alpha \Rightarrow^* \epsilon \text{ then } \{\epsilon\} \text{ else } \emptyset$$

## ► Siguietes (FOLLOW).

Dado un símbolo no terminal  $A \in N$ , el conjunto  $\text{FOLLOW}(A) \subseteq \Sigma \cup \{\$ \}$  de los “siguietes” de  $A$  es un conjunto de símbolos terminales (y posiblemente  $\$$ ) definido como sigue:

$$\text{FOLLOW}(A) = \{x \in \Sigma \mid S \Rightarrow^* \alpha A x \beta \text{ para } \alpha, \beta \in (N \cup \Sigma)^*\} \cup \\ \text{if } S \Rightarrow^* \alpha A \text{ then } \{\$ \} \text{ else } \emptyset$$

# Análisis sintáctico LL(1)

## Cómputo del conjunto FIRST para un símbolo.

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$ .

*Salida:* El conjunto FIRST( $X$ )

para cada símbolo  $X \in N \cup \Sigma$ .

$\mathcal{F} :=$  un diccionario  $\{X \mapsto \emptyset \mid X \in N \cup \Sigma\}$

**foreach** símbolo terminal  $a \in \Sigma$

$\mathcal{F}[a] := \{a\}$

**end**

**while** hay algún cambio

**foreach** producción  $(A \rightarrow X_1 \dots X_n) \in P$

**for**  $i = 1$  to  $n$

**if**  $\epsilon \in \mathcal{F}[X_1] \cap \dots \cap \mathcal{F}[X_{i-1}]$

$\mathcal{F}[A] := \mathcal{F}[A] \cup (\mathcal{F}[X_i] \setminus \{\epsilon\})$

**end**

**end**

**if**  $\epsilon \in \mathcal{F}[X_1] \cap \dots \cap \mathcal{F}[X_n]$

$\mathcal{F}[A] := \mathcal{F}[A] \cup \{\epsilon\}$

**end**

**end**

**end**

**return**  $\mathcal{F}$

# Análisis sintáctico LL(1)

**Observación.** Notar que el algoritmo para calcular el conjunto FIRST sirve en particular para calcular el conjunto de símbolos anulables: un símbolo  $X$  es anulado si y sólo si  $\epsilon \in \text{FIRST}(X)$ .

# Análisis sintáctico LL(1)

**Ejercicio.** Calcular el conjunto FIRST para todos los símbolos de la gramática

$$G = (\{S, V, D, T, T', U\}, \{\mathbf{id}, \mathbf{int}, \mathbf{bool}, \Rightarrow, ;, (, )\}, P, S)$$

$$\begin{array}{lcl} S & \rightarrow & VS \\ & | & \epsilon \\ V & \rightarrow & D \mathbf{id}; \\ D & \rightarrow & T \\ & | & \epsilon \\ T & \rightarrow & UT' \\ T' & \rightarrow & \Rightarrow UT' \\ & | & \epsilon \\ U & \rightarrow & \mathbf{int} \\ & | & \mathbf{bool} \\ & | & (T) \end{array}$$

# Análisis sintáctico LL(1)

## Cómputo del conjunto FIRST para una cadena.

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$   
y una cadena  $X_1 \dots X_n \in (N \cup \Sigma)^*$ .

*Salida:* El conjunto  $\text{FIRST}(\alpha)$ .

Computar el diccionario  $\mathcal{F}$   
que a cada símbolo  $X \in N \cup \Sigma$  le asocia  $\text{FIRST}(X)$ .

$\mathcal{R} := \emptyset$

**for**  $i = 1$  to  $n$

**if**  $\epsilon \in \mathcal{F}[X_1] \cap \dots \cap \mathcal{F}[X_{i-1}]$

$\mathcal{R} := \mathcal{R} \cup (\mathcal{F}[X_i] \setminus \{\epsilon\})$

**end**

**end**

**if**  $\epsilon \in \mathcal{F}[X_1] \cap \dots \cap \mathcal{F}[X_n]$

$\mathcal{R} := \mathcal{R} \cup \{\epsilon\}$

**end**

**return**  $\mathcal{R}$

# Análisis sintáctico LL(1)

## Cómputo del conjunto FOLLOW.

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$ .

*Salida:* El conjunto FOLLOW( $A$ )

para cada símbolo no terminal  $A \in N$ .

$\mathcal{W} :=$  un diccionario  $\{X \mapsto \emptyset \mid X \in N \cup \Sigma\}$

$\mathcal{W}[S] := \{\$ \}$

**foreach** producción  $A \rightarrow \alpha B \beta$

$\mathcal{W}[B] := \mathcal{W}[B] \cup (\text{FIRST}(\beta) \setminus \{\epsilon\})$

**end**

**while** hay algún cambio

**foreach** producción  $A \rightarrow \alpha B \beta$  tal que  $\epsilon \in \text{FIRST}(\beta)$

$\mathcal{W}[B] := \mathcal{W}[B] \cup \mathcal{W}[A]$

**end**

**end**

**return**  $\mathcal{W}$



## Análisis sintáctico LL(1)

**Ejercicio.** Calcular el conjunto FOLLOW para todos los símbolos no terminales de la gramática

$$G = (\{S, V, D, T, T', U\}, \{\mathbf{id}, \mathbf{int}, \mathbf{bool}, \Rightarrow, ;, (, )\}, P, S)$$

$S \rightarrow VS$

$\mid \epsilon$

$V \rightarrow D \mathbf{id};$

$D \rightarrow T$

$\mid \epsilon$

$T \rightarrow UT'$

$T' \rightarrow \Rightarrow UT'$

$\mid \epsilon$

$U \rightarrow \mathbf{int}$

$\mid \mathbf{bool}$

$\mid (T)$

## Análisis sintáctico LL(1)

Dada una gramática  $G = (N, \Sigma, P, S)$ , una **tabla de análisis sintáctico LL(1)** es un diccionario  $\mathcal{T}$  que a cada par  $(A, b) \in N \times (\Sigma \cup \{\$\})$  le asocia un conjunto de producciones  $(C \rightarrow \alpha) \in P$ . Se construye con el siguiente método:

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$ .

*Salida:* La tabla de análisis sintáctico LL(1) para  $G$ .

Poner  $\mathcal{T}[A, b] := \emptyset$  para todo  $(A, b) \in N \times \Sigma$ .

```
foreach producción  $(A \rightarrow \alpha) \in P$ 
    foreach símbolo terminal  $x \in (\text{FIRST}(\alpha) \setminus \{\epsilon\})$ 
         $\mathcal{T}[A, x] := \mathcal{T}[A, x] \cup \{A \rightarrow \alpha\}$ 
    end
    if  $\epsilon \in \text{FIRST}(\alpha)$ 
        foreach símbolo  $x \in \text{FOLLOW}(A)$ 
             $\mathcal{T}[A, x] := \mathcal{T}[A, x] \cup \{A \rightarrow \alpha\}$ 
        end
    end
end
return  $\mathcal{T}$ 
```

# Análisis sintáctico LL(1)

Si la tabla de análisis sintáctico LL(1) verifica que todos los conjuntos  $\mathcal{T}[A, b]$  tienen a lo sumo una entrada, se dice que la gramática es **LL(1)**.

En ese caso, la tabla de análisis sintáctico LL(1) funciona como oráculo para el análisis sintáctico predictivo.

## Análisis sintáctico LL(1)

**Ejercicio.** Calcular la tabla de análisis sintáctico LL(1) para la gramática

$$G = (\{S, V, D, T, T', U\}, \{\mathbf{id}, \mathbf{int}, \mathbf{bool}, \Rightarrow, ;, (, )\}, P, S)$$

$$S \rightarrow VS \mid \epsilon$$

$$V \rightarrow D \mathbf{id};$$

$$D \rightarrow T \mid \epsilon$$

$$T \rightarrow UT'$$

$$T' \rightarrow \Rightarrow UT' \mid \epsilon$$

$$U \rightarrow \mathbf{int} \mid \mathbf{bool} \mid (T)$$

Recordar que:

	FIRST	FOLLOW
$S$	$\epsilon \mathbf{id} \mathbf{int} \mathbf{bool} ($	$\$$
$V$	$\mathbf{id} \mathbf{int} \mathbf{bool} ($	$\mathbf{id} \mathbf{int} \mathbf{bool} ( \$$
$D$	$\epsilon \mathbf{int} \mathbf{bool} ($	$\mathbf{id}$
$T$	$\mathbf{int} \mathbf{bool} ($	$\mathbf{id} )$
$T'$	$\Rightarrow \epsilon$	$\mathbf{id} )$
$U$	$\mathbf{int} \mathbf{bool} ($	$\Rightarrow \mathbf{id} )$

# Análisis sintáctico LL(1)

**Ejercicio.** Usando la tabla de análisis sintáctico LL(1) para la gramática anterior, analizar sintácticamente **int**  $\Rightarrow$  **int id**;

# Análisis sintáctico LL(1)

**Ejercicio.** Mostrar que la gramática

$G = (\{S, E\}, \{\text{if, then, else, cmd, exp}\}, P, S)$  no es LL(1):

$S$	$\rightarrow$	<b>if <math>E</math> then <math>S</math></b>
		<b>if <math>E</math> then <math>S</math> else <math>S</math></b>
		<b>cmd</b>
$E$	$\rightarrow$	<b>exp</b>

# Análisis sintáctico LL(1)

- ▶ Un símbolo  $X \in N \cup \Sigma$  es **inútil** si no hay ninguna derivación de la forma  $S \Rightarrow^* \alpha X \beta \Rightarrow^* \gamma \in \Sigma^*$ , es decir,  $X$  nunca puede aparecer en la derivación de una cadena  $\gamma \in \Sigma^*$ .

Una gramática recursiva a izquierda sin símbolos inútiles no puede ser LL(1). (¿Por qué?).

- ▶ Una gramática LL(1) no puede ser ambigua. (¿Por qué?).
- ▶ ¿Cuánto tiempo toma un parser descendente predictivo en analizar sintácticamente una cadena  $\alpha$  si la gramática es LL(1) y se dispone de la tabla de análisis sintáctico LL(1)?

# Análisis sintáctico $LL(k)$

El análisis sintáctico  $LL(1)$  se puede generalizar al análisis  $LL(k)$  en el que el parser puede contemplar los siguientes  $k$  símbolos de la entrada.

- ▶ La construcción es bastante más complicada.
- ▶ Se consideraba impracticable hasta mediados de los 1990.  
(Observar que el tamaño de la tabla es  $|N| \cdot |\Sigma|^k$ . Para números típicos como  $|N| = 50$ ,  $|\Sigma| = 128$ ,  $k = 4$  *a priori* la tabla podría tener 13.421.772.800 entradas).
- ▶ La ventaja de los parsers  $LL(k)$  es su expresividad: hay gramáticas que son  $LL(k+1)$  pero no son  $LL(k)$ .



# Técnicas surtidas de análisis sintáctico

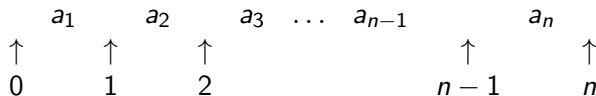
## Algoritmo de Earley

# Algoritmo de Earley

El **algoritmo de Earley** es un algoritmo general de análisis sintáctico. Su interés es que funciona para gramáticas independientes del contexto **arbitrarias**.

# Algoritmo de Earley

- ▶ Si  $A \rightarrow \alpha\beta$  es una producción, la notación  $A \rightarrow \alpha \bullet \beta$  representa una situación en la que se consumió  $\alpha$  y se espera leer  $\beta$ .
- ▶ Si la entrada que se quiere analizar es  $a_1 \dots a_n$  hay  $n + 1$  posiciones en la entrada:



- ▶ Para cada posición de la entrada  $0 \leq k \leq n$ , el parser genera un conjunto de estados  $\mathcal{S}[k]$ . Cada estado es de la forma  $(A \rightarrow \alpha \bullet \beta, i)$  donde:
  - ▶ El número  $0 \leq i \leq n$  representa la posición en la que comenzó la aplicación de la producción  $A \rightarrow \alpha\beta$ .
  - ▶ La cadena  $\alpha$  es el fragmento de la producción  $A \rightarrow \alpha\beta$  que ya fue consumido hasta la posición  $k$ .
  - ▶ La cadena  $\beta$  es el fragmento aún no consumido.

# Algoritmo de Earley

*Entrada:* Una gramática  $G = (N, \Sigma, P, S)$ ,

una cadena de entrada  $\alpha = a_1 \dots a_n$ .

*Salida:* Un booleano indicando si  $\alpha \in L(G)$ .

Crear un arreglo  $S[0..n]$  de conjuntos de estados.

Poner  $S[k] = \emptyset$  para cada  $0 \leq k \leq n$ .

$S[0] := \{(S' \rightarrow \bullet S, 0)\}$

**for**  $k = 0$  **to**  $n$

**while** hay algún estado de  $S[k]$  no visitado

        Sea  $(A \rightarrow \alpha \bullet \beta, j)$  un estado no visitado de  $S[k]$ .

        Marcar  $(A \rightarrow \alpha \bullet \beta, j)$  como visitado.

**if**  $\beta$  comienza con un símbolo no terminal

            predecir( $(A \rightarrow \alpha \bullet \beta, j), k$ )

**elseif**  $\beta$  comienza con un símbolo terminal

            avanzar( $(A \rightarrow \alpha \bullet \beta, j), k$ )

**else** /\*  $\beta$  es vacía \*/

            completar( $(A \rightarrow \alpha \bullet \beta, j), k$ )

**end**

**end**

**end**

**return**  $(S' \rightarrow S \bullet, 0) \in S[n]$

# Algoritmo de Earley

```
// Si lo que sigue es un símbolo no terminal.
function predecir( $(A \rightarrow \alpha \bullet B \beta, j)$ ,  $k$ )
    foreach producción  $B \rightarrow \gamma$ 
        Agregar  $(B \rightarrow \bullet \gamma, k)$  a  $S[k]$ .
    end
end

// Si lo que sigue es un símbolo terminal.
function avanzar( $(A \rightarrow \alpha \bullet b \beta, j)$ ,  $k$ )
    if  $b = a_k$ 
        Agregar  $(A \rightarrow \alpha b \bullet \beta, j)$  a  $S[k+1]$ .
    end
end

// Si la producción termina.
function completar( $(A \rightarrow \alpha \bullet, j)$ ,  $k$ )
    foreach estado  $(B \rightarrow \beta \bullet A \gamma, i) \in S[j]$ 
        Agregar  $(B \rightarrow \beta A \bullet \gamma, i)$  a  $S[k]$ .
    end
end
```

# Algoritmo de Earley

**Ejercicio.** Analizar la cadena  $(\mathbf{n} + \mathbf{n})$  con la gramática  $G = (\{E, T\}, \{\mathbf{n}, +, (, )\}, P, E)$ , con producciones:

$$\begin{array}{lcl} E & \rightarrow & T \\ & | & E + T \\ T & \rightarrow & \mathbf{n} \\ & | & (E) \end{array}$$

## Análisis sintáctico con combinadores



# Análisis sintáctico con combinadores

Un analizador sintáctico que procesa la entrada de izquierda a derecha se puede pensar como una función que:

- ▶ Recibe una entrada.
- ▶ Analiza sintácticamente algún prefijo de la entrada.
- ▶ Devuelve el resultado de hacer el análisis sintáctico y el sufijo de la entrada que todavía no fue analizado.  
(Alternativamente, puede fallar).

# Análisis sintáctico con combinadores

Por ejemplo, en Haskell:

```
data ErrorMsg = String
type Parser a = String -> Either ErrorMsg (a, String)

--
-- match1 c    Es un parser básico que acepta
--              el lenguaje {c}
--
match1 :: Char -> Parser Char
match1 c [] = Left ("Esperaba " ++ show c ++
                  " pero se encuentra EOF")
match1 c (c' : s')
  | c == c'    = Right (c, s')
  | otherwise  = Left ("Esperaba " ++ show c ++
                  " pero se encuentra " ++ show c')
```

# Análisis sintáctico con combinadores

Se pueden construir **combinadores** de parsers: son funciones que reciben parsers y devuelven parsers.

Combinador de secuenciación binaria:

```
--  
-- seqP p1 p2    Parser que acepta la concatenación  
--               de los lenguajes aceptados por p1 y p2.  
--  
seqP :: Parser a -> Parser b -> Parser (a, b)  
seqP p1 p2 s =  
  case p1 s of  
    Left errmsg -> Left errmsg  
    Right (a, s') ->  
      case p2 s' of  
        Left errmsg      -> Left errmsg  
        Right (b, s'') -> Right ((a, b), s'')
```

# Análisis sintáctico con combinadores

Combinador de alternativa binaria:

```
--  
-- altP p1 p2    Parser que acepta la unión  
--               de los lenguajes aceptados por p1 y p2.  
--  
-- Nota: hace backtracking.  
--  
altP :: Parser a -> Parser b -> Parser (Either a b)  
altP p1 p2 s =  
  case p1 s of  
    Right (a, s') -> Right (Left a, s')  
    Left errmsg ->  
      case p2 s of  
        Left errmsg -> Left ("La entrada no está en " ++  
                              "ninguno de los lenguajes.")  
        Right (b, s') -> Right (Right b, s')
```

# Análisis sintáctico con combinadores

Combinador de repetición:

```
--  
-- repP p    Parser que acepta la "repetición"  
--           del lenguaje aceptado por p.  
--  
-- Nota: no acepta exactamente la clausura de Kleene,  
-- porque usa la regla de "maximal munch", es decir,  
-- trata de consumir el prefijo de la entrada más  
-- largo posible.  
--  
repP :: Parser a -> Parser [a]  
repP p s =  
  case p s of  
    Left errmsg    -> Right []  
    Right (a, s') ->  
      case repP p s' of  
        Left errmsg -> Left errmsg  
        Right (as, s'') -> Right (a : as, s'')
```

# Análisis sintáctico con combinadores

Se pueden agregar acciones para construir distintos tipos de resultados:

```
mapP :: (a -> b) -> Parser a -> Parser b
mapP f p s =
  case p s of
    Left errmsg -> Left errmsg
    Right (a, s') -> Right (f a, s')
```

## Análisis sintáctico con combinadores

Por conveniencia se pueden crear combinadores de secuenciación o alternativa  $n$ -arios, por ejemplo:

```
seq4P :: Parser a -> Parser b -> Parser c -> Parser d ->
        Parser (a, b, c, d)
seq4P p1 p2 p3 p4 =
    mapP (\ (a, (b, (c, d))) -> (a, b, c, d))
        (seqP p1 (seqP p2 (seqP p3 p4)))
```

## Análisis sintáctico con combinadores

Se pueden crear combinadores de repetición con distintas restricciones sobre la cantidad de repeticiones, p.ej. repetición al menos una vez:

```
rep1P :: Parser a -> Parser [a]
rep1P p = mapP (\ (a, as) -> a : as) (seqP p (repP p))
```



## Análisis sintáctico con combinadores

Se pueden generalizar los combinadores de secuencia y alternativa para combinar listas de parsers. Por simplicidad supondremos que las listas no están vacías.

```
seqsP :: [Parser a] -> Parser [a]
seqsP [p]          = mapP (\ a -> [a]) p
seqsP (p : ps) = mapP (\ (a, as) -> a : as)
                  (seqP p (seqsP ps))
```

```
altsP :: [Parser a] -> Parser a
altsP [p]          = p
altsP (p : ps) = mapP f (altP p (altsP ps))
  where f (Left a)  = a
        f (Right a) = a
```

# Análisis sintáctico con combinadores

Ejemplo de un parser sencillo:

```
digitP = altsP (map match1 ['0','1','2','3','4',  
                             '5','6','7','8','9'])  
numP   = mapP (\ s -> read s :: Integer)  
        (rep1P digitP)  
spaceP = altsP (map match1 [' ', '\t', '\r', '\n'])  
spacesP = repP spaceP  
exprP  = altsP [  
    mapP (\ (a, _, _, _, b) -> a + b)  
        (seq5P termP spacesP (match1 '+') spacesP exprP),  
    termP ]  
termP  = altsP [  
    mapP (\ (a, _, _, _, b) -> a * b)  
        (seq5P factorP spacesP (match1 '*') spacesP termP),  
    factorP ]  
factorP = altsP [  
    numP,  
    mapP (\ (_, a, _) -> a)  
        (seq3P (match1 '(') exprP (match1 ')')) ]
```

# Análisis sintáctico con combinadores

**Nota:** el análisis sintáctico basado en combinadores no es una técnica propia de los lenguajes funcionales. Por ejemplo los parsers y combinadores de parsers se pueden representar como objetos:

```
class Match1(object):

    def __init__(self, c):
        self.c = c

    def analizar(self, entrada):
        if entrada.peek() == self.c:
            entrada.read()
            return self.c
        else:
            raise Exception("Esperaba '%c'" % (self.c,))
```

Análisis con tablas de precedencia

# Análisis con tablas de precedencia

Supongamos que tenemos una tabla de precedencia de operadores:

Nivel	Operador	Asociatividad
1	$\theta_1$	$assoc_1$
$\vdots$	$\vdots$	
$M$	$\theta_M$	$assoc_M$

donde  $\theta_1, \dots, \theta_M \in \Sigma$  son símbolos terminales,  $\theta_i$  tiene menor precedencia que  $\theta_{i+1}$ , es decir, se espera que  $E_1 \theta_i E_2 \theta_{i+1} E_3$  se analice como  $E_1 \theta_i (E_2 \theta_{i+1} E_3)$ , y la asociatividad  $assoc_i$  está en  $\{\text{left, right, prefix, suffix}\}$ .

# Análisis con tablas de precedencia

```
function analizar_expresión(nivel)  
  if nivel = M  
    if proximo_token() = "("  
      consumir_token("(")  
      e := analizar_expresión(1)  
      consumir_token(")")  
      return e  
    else  
      return analizar_expresión_atómica() // (TODO)  
    end  
  else  
    case assocnivel = left  
      return analizar_izquierdo(nivel)  
    case assocnivel = right  
      return analizar_derecho(nivel)  
    case assocnivel = prefix  
      return analizar_prefijo(nivel)  
    case assocnivel = suffix  
      return analizar_sufijo(nivel)  
    end  
  end  
end
```

# Análisis con tablas de precedencia

```
function analizar_izquierdo(nivel)  
  e := analizar_expresión(nivel + 1)  
  while siguiente_token =  $\theta_{nivel}$   
    consumir_token(" $\theta_{nivel}$ ")  
    e := new AST(" $\theta_{nivel}$ ", e, analizar_op(nivel + 1))  
  end  
  return e  
end
```

```
function analizar_derecho(nivel)  
  lista := []  
  lista.push(analizar_op(nivel + 1))  
  while siguiente_token =  $\theta_{nivel}$   
    consumir_token(" $\theta_{nivel}$ ")  
    lista.push(analizar_op(nivel + 1))  
  end  
  e := lista[lista.size - 1];  
  for i = lista.size - 2 downto 0  
    e := new AST(" $\theta_{nivel}$ ", lista[i], e)  
  end
```

# Análisis con tablas de precedencia

```
function analizar_prefijo(nivel)  
  if siguiente_token =  $\theta_{nivel}$   
    consumir_token( $\theta_{nivel}$ )  
    e := analizar_expresión(nivel)  
    return new AST(" $\theta_{nivel}$ ", e)  
  else  
    return analizar_expresión(nivel + 1)  
  end  
end
```

```
function analizar_sufijo(nivel)  
  e := analizar_expresión(nivel + 1)  
  while siguiente_token =  $\theta_{nivel}$   
    consumir_token( $\theta_{nivel}$ )  
    e := new AST(" $\theta_{nivel}$ ", e)  
  end  
  return e  
end
```