Teoría de Lenguajes

Clase Teórica 8

Gramáticas libres de contexto

Primer Cuatrimestre 2024

Bibliografía: Capítulos 6 y 7, *Introduction to Automata Theory, Languages and Computation*, J. Hopcroft, R. Motwani, J. Ullman, Second Edition, Addison Wesley, 2001.

Gramáticas Libres de Contexto

Recordemos la definición.

Definición

Una gramática $G = \langle V_N, V_T, P, S \rangle$ *es libre de contexto si las producciones en* P *son de la forma*

$$A \to \alpha, \quad \textit{con} \quad A \in V_N \quad \textit{y} \quad \alpha \in (V_N \cup V_T)^* .$$

Demostraremos que para cada gramática libre de contexto G hay un autómata de pila M que acepta el lenguaje generado por dicha gramática y viceversa.

Dada una gramática libre de contexto G, se puede reconocer si una palabra pertenece a $\mathcal{L}(G)$ en tiempo del orden cúbico de la longitud de la palabra. En casos especiales (determinismo), se puede reconocer en tiempo lineal. Lo veremos próximamente.

Lenguaje generado por una gramática

Recordemos

Definición (Derivación $\Rightarrow$)

Sea $G = (V_N, V_T, P, S)$ una gramática.
Si $\alpha, \beta, \gamma_1, \gamma_2 \in (V_N \cup V_T)^*$ y $\alpha \to \beta \in P$ entonces

$$\gamma_1 \alpha \gamma_2 \Rightarrow \gamma_1 \beta \gamma_2$$

La relación $\Rightarrow$ es un subconjunto $(V_N \cup V_T)^* \times (V_N \cup V_T)^*$ y significa derivar en un solo paso.

Las relaciones $\overset{+}{\Rightarrow}$ y $\overset{*}{\Rightarrow}$ son la clausura transitiva y reflexo-transitiva respectivamente, (más de un paso de derivación, y cero o más pasos).

Si $\alpha \in (V_N \cup V_T)^*$ y $S \overset{*}{\Rightarrow} \alpha$ decimos que α es una *forma sentencial* de G.

Definición (Lenguaje generado de una gramática G)

Dada una gramática $G = (V_N, V_T, P, S)$,

$$\mathcal{L}(G) = \{w \in T^* : S \overset{+}{\Rightarrow} w\}$$

Ejemplo

Sea $G = \langle V_N, V_T, P, S \rangle$ la gramática libre de contexto tal que $V_N = \{E\}$, $V_T = \{+, *, \mathbf{id}, (,)\}$, $S = E$ y P tiene

$$E \to E + E \mid E * E \mid (E) \mid \mathbf{id}$$

En cada paso de la de la derivación debemos elegir qué símbolo no-terminal reescribiremos y luego debemos elegir una producción que tenga ese símbolo del lado izquierdo.

Si elegimos el no-terminal más a la izquierda,
$E \overset{*}{\Rightarrow} (\mathbf{id})$, porque $E \Rightarrow (E) \Rightarrow (\mathbf{id})$

$E \overset{*}{\Rightarrow} (\mathbf{id} + \mathbf{id})$ porque $E \Rightarrow (E) \Rightarrow (E + E) \Rightarrow (\mathbf{id} + E) \Rightarrow (\mathbf{id} + \mathbf{id})$.

Si elegimos el no terminal más a la derecha,

$E \Rightarrow (E) \Rightarrow (E + E) \Rightarrow (E + \mathbf{id}) \Rightarrow (\mathbf{id} + \mathbf{id})$.

Lenguajes libres de contexto

Son exactamente los lenguajes generados por las gramáticas libres de contexto $G = (N, T, P, S)$, donde P tiene producciones de la forma $A \to \alpha$, $\alpha \in (V \cup T)^*$.

Son exactamente los lenguajes reconocibles por autómatas de pila.

Se reconocen en tiempo cúbico en longitud entrada (algoritmo CYK).

Formas normales
Greibach $A \to aA_1...A_n$, $n \geq 0$ (da origen a AP sin λ-transiciones)
Chomsky $A \to a$, $A \to BC$

Ejemplos
$L = \{a^n b^n : n \geq 1\}$,
$L = \{a^n b^n c^k\} \cup \{a^n b^k c^k\}$

Contraejemplo
$L = \{a^n b^n a^n\}$ (Lema de Pumping para libres de contexto)

Nota: Aun no vimos fomas normales, ni algoritmo CYK, ni Lema de Pumping para libres de contexto

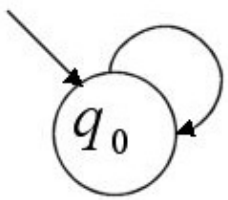
Un autómata de pila para esta gramática libre de contexto

Ejemplo

Sea $G = \langle V_N, V_T, P, S\rangle$ la gramática libre de contexto tal que $V_N = \{E\}$, $V_T = \{+, *, \mathbf{id}, (,)\}$, $S = E$ y P tiene $E \to E + E \mid E * E \mid (E) \mid \mathbf{id}$

Sea $M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, \emptyset\rangle$ con $Q = \{q_0\}$, $\Sigma = V_T$, $\Gamma = V_N \cup V_T$ y $Z_0 = S$.

$$\lambda, E \mid \mathrm{id}; \quad \lambda, E \mid (E); \quad \lambda, E \mid E + E; \quad \lambda, E \mid E * E$$



$$+, + \mid \lambda; \quad *, * \mid \lambda; \quad),) \mid \lambda; \quad (, (\mid \lambda; \quad \mathrm{id}, \mathrm{id} \mid \lambda$$

Si en el tope de la pila hay un símbolo no-terminal, el autómata M lo reemplazará en la pila por el lado derecho de alguna producción.

Si en el tope de la pila hay un símbolo terminal el autómata M constatará que es igual al próximo símbolo en la cadena de entrada y lo desapilará.

Este autómata acepta $\mathcal{L}(G)$ por pila vacía.

Teorema (Chomsky 1962, Evey 1963)

Para cada gramática G libre de contexto existe un autómata de pila M tal que $\mathcal{L}(G) = \mathcal{L}_\lambda(M)$.

Lema

$A \overset{*}{\Rightarrow} w$ *si y solo si* $(q, w, A) \overset{*}{\underset{M}{\vdash}} (q, \lambda, \lambda)$.

Demostración.

Por inducción en la cantidad de derivaciones de w, que llamamos m.

Caso base, $m = 1$. Tenemos $A \overset{1}{\Rightarrow} w$ para $w = a_1 \dots a_k$, con $k > 0$, si y solo si,

$(q, a_1 \dots a_k, A) \underset{M}{\vdash} (q, a_1 \dots a_k, a_1 \dots a_k) \overset{k}{\underset{M}{\vdash}} (q, \lambda, \lambda)$.

Caso inductivo, $m > 1$.

HI: Para todo $j < m$, $A \overset{j}{\Rightarrow} x$ si y solo si $(q, x, A) \overset{*}{\underset{M}{\vdash}} (q, \lambda, \lambda)$.

Por definición de derivación,
$A \overset{m}{\Rightarrow} w$ si y solo si $A \to X_1 \dots X_k$ está en P tal que para cada i, $X_i \overset{m_i}{\Rightarrow} x_i$, para algún $m_i < m$ y $x_1 \dots x_k = w$.

Por definición de M,
$(A \to X_1 \dots X_k) \in P$ si y solo si $(q, w, A) \underset{M}{\vdash} (q, w, X_1 \dots X_k)$.

Si $X_i \in V_N$, entonces por hipótesis inductiva, $(q, x_i, X_i) \overset{*}{\underset{M}{\vdash}} (q, \lambda, \lambda)$.

Si $X_i = x_i \in V_T^*$, entonces $(q, x_i, x_i) \underset{M}{\vdash} (q, \lambda, \lambda)$.

Por lo tanto,

$$(q, w, A) \underset{M}{\vdash} (q, x_1 \dots x_k, X_1 \dots X_k) \underset{M}{\vdash} (q, x_2 \dots x_k, X_2 \dots X_k) \underset{M}{\vdash} \dots$$
$$\underset{M}{\vdash} (q, x_k, X_k) \underset{M}{\vdash} (q, \lambda, \lambda). \qquad \square$$

Demostración del Teorema. Sea GLC $G = \langle V_N, V_T, P, S\rangle$. Definimos AP $M = \langle \{q\}, V_T, V_N \cup V_T, \delta, q, S, \phi\rangle$ donde
$\delta : Q \times (V_T \cup \{\lambda\}) \times (V_N \cup V_T) \to \mathcal{P}(Q \times (V_N \cup V_T)^*)$ es tal que
- si $(A \to \alpha) \in P$, entonces $(q, \alpha) \in \delta(q, \lambda, A)$.
- $\forall a \in V_T,\ \delta(q, a, a) = \{(q, \lambda)\}$.

Queremos ver que

$S \overset{+}{\Rightarrow} w$ si y solo si $(q, w, S) \overset{*}{\underset{M}{\vdash}} (q, \lambda, \lambda)$.

El Lema dice que para cualquier A en V_N, $w \in V_T^*$

$$A \overset{*}{\Rightarrow} w \text{ si y solo si } (q, w, A) \overset{*}{\underset{M}{\vdash}} (q, \lambda, \lambda).$$

Luego, para cualquier $w \in V_T^*$,

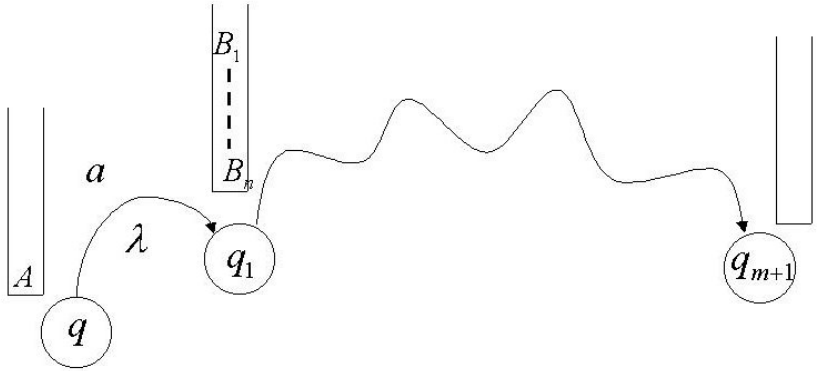
$$S \overset{+}{\Rightarrow} w \text{ si y solo si } (q, w, S) \overset{+}{\underset{M}{\vdash}} (q, \lambda, \lambda).$$

por lo tanto $\mathcal{L}(G) = \mathcal{L}_\lambda(M)$. □

Teorema (Chomsky 1962, Evey 1963)

Si M es un autómata de pila entonces $\mathcal{L}_\lambda(M)$ es libre de contexto.

Demostración del Teorema.



Dado AP $M = (Q, \Sigma, \Gamma, \delta, q_0, Z_0, \emptyset)$ definamos $G = \langle V_N, V_T, P, S\rangle$ donde S es símbolo nuevo, $V_N = \{[q, A, p] : q \in Q, A \in \Gamma, p \in Q\} \cup \{S\}$, $V_T = \Sigma$ y P:

- $S \to [q_0, Z_0, q]$ en P, para cada q en Q.
- $[q, A, q_1] \to a$ en P si y solo si $(q_1, \lambda) \in \delta(q, a, A)$.
- $[q, A, q_1] \to \lambda$ en P si y solo si $(q_1, \lambda) \in \delta(q, \lambda, A)$.

Para cada $q, q_1, q_2, \ldots, q_{m+1} \in Q$, $a \in \Sigma$ y $A, B_1, \ldots, B_m \in \Gamma$,

- $[q, A, q_{m+1}] \to a[q_1, B_1, q_2] \ldots [q_m, B_m, q_{m+1}]$ en P si y solo si $(q_1, B_1 \ldots B_m) \in \delta(q, a, A)$.
- $[q, A, q_{m+1}] \to [q_1, B_1, q_2] \ldots [q_m, B_m, q_{m+1}]$ en P si y solo si $(q_1, B_1 \ldots B_m) \in \delta(q, \lambda, A)$.

(G es tal que su derivación más a la izquierda es una simulación de M).

Lema

Para todo $q \in Q, A \in \Gamma, p \in Q$,

$$(q, x, A) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda) \quad \textit{si y solo si} \quad [q, A, p] \overset{*}{\underset{G}{\Rightarrow}} x.$$

Demostración del Lema.

Veamos primero de autómta M a gramática G

Veamos por inducción que para todo $i \geq 1$,

$$\text{Si } (q, x, A) \overset{i}{\underset{M}{\vdash}} (p, \lambda, \lambda) \quad \text{entonces } [q, A, p] \overset{*}{\underset{G}{\Rightarrow}} x.$$

Escribimos a para denotar un símbolo de Σ o λ.

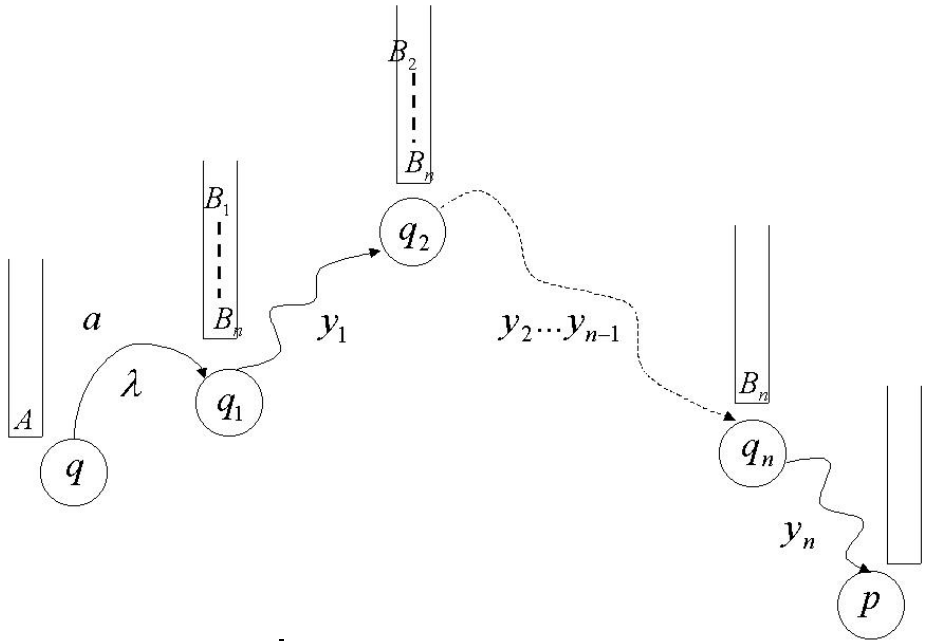
Caso $i = 1$. Tenemos $(q, a, A) \overset{1}{\underset{M}{\vdash}} (p, \lambda, \lambda)$. Entonces, $(p, \lambda) \in \delta(q, a, A)$.

Y por definición de G, $[q, A, p] \to a$. Por lo tanto, $[q, A, p] \underset{G}{\Rightarrow} a$.

Caso $i > 1$. Tenemos $x = ay$ con $y \in \Sigma^*$, $(q, x, A) \overset{i}{\underset{M}{\vdash}} (p, \lambda, \lambda)$

Existen $B_1, .., B_n$ en Γ tales $(q, ay, A) \underset{M}{\vdash} (q_1, y, B_1, \ldots, B_n) \overset{i-1}{\underset{M}{\vdash}} (p, \lambda, \lambda)$.

Necesariamente y se decompone como $y = y_1 \ldots y_n$, tales que para $1 \leq j \leq n$, $y_1 \ldots y_j$ hacen que B_j quede en el tope de pila.



Existen $q_2, \ldots, q_{n+1}$ tales que, para $1 \leq j \leq n$, en menos de i transiciones, $(q_j, y_j, B_j) \overset{*}{\underset{M}{\vdash}} (q_{j+1}, \lambda, \lambda)$.

Por hipótesis inductiva, usando menos que i pasos, para cada $1 \leq j \leq n$,

Si $(q_j, y_j, B_j) \overset{*}{\underset{M}{\vdash}} (q_{j+1}, \lambda, \lambda)$ entonces $[q_j, B_j, q_{j+1}] \overset{*}{\underset{G}{\Rightarrow}} y_j$.

Pero en G tenemos la producción

$[q, A, q_{n+1}] \to a[q_1, B_1, q_2] \ldots [q_n, B_n, q_{n+1}]$

Usando que para cada j, $[q_j, B_j, q_{j+1}] \overset{*}{\underset{G}{\Rightarrow}} y_j$, obtenemos

$[q, A, q_{n+1}] \overset{*}{\Rightarrow} a y_1 \ldots y_n = x$.

Veamos ahora de gramática G a autómata M

Veamos por inducción sobre i que para todo $i \geq 1$,

Si $[q, A, p] \underset{G}{\overset{i}{\Rightarrow}} x$ entonces $(q, x, A) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda)$.

Escribimos a para denotar un símbolo de Σ o λ.

Para $i = 1$. Supongamos $[q, A, p] \underset{G}{\overset{1}{\Rightarrow}} a$ Entonces, $[q, A, p] \underset{G}{\overset{1}{\to}} a$ es producción de G y por definición de M, $(p, \lambda) \in \delta(q, a, A)$.

Para $i > 1$. Supongamos $[q, A, p] \underset{G}{\overset{i}{\Rightarrow}} x$. Entonces,

$[q, A, p] \underset{G}{\Rightarrow} a[q_1, B_1, q_2] \dots [q_n, B_n, p] \underset{G}{\overset{i-1}{\Rightarrow}} x$.

Descomponemos x como $x = ax_1 \dots x_n$ tal que para cada $1 \leq j \leq n$, cada derivación toma menos de i pasos, $[q_j, B_j, q_{j+1}] \underset{G}{\overset{*}{\Rightarrow}} x_j$

Por hipótesis inductiva, usando menos que i pasos, para cada $1 \leq j \leq n$,

Si $(q_j, x_j, B_j) \overset{*}{\underset{M}{\vdash}} (q_{j+1}, \lambda, \lambda)$ entonces $(q_j, x_j, B_j ... B_n) \overset{*}{\underset{M}{\vdash}} (q_{j+1}, \lambda, B_{j+1} ... B_n)$.

Partimos de $[q, A, p] \underset{G}{\Rightarrow} a[q_1, B_1, q_2] \dots [q_n, B_n, p]$.

Por definicion de M, $(q, a, A) \underset{M}{\vdash} (q_1, \lambda, B_1 \dots B_n)$.

Llamando p al q_{n+1}, obtenemos

$(q, ax_1 ... x_n, A) \underset{M}{\vdash} (q_1, x_1 ... x_n, B_1 ... B_n) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda)$.

□

Continuación de Demostración del Teorema.

Por el Lema, para todo $q \in Q, A \in \Gamma, p \in Q$,

$$(q, x, A) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda) \qquad \text{si y solo si} \qquad [q, A, p] \overset{*}{\underset{G}{\Rightarrow}} x.$$

Tomando $q = q_0$ y $A = Z_0$,

$$(q_0, x, Z_0) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda) \quad \text{si y solo si} \quad [q_0, Z_0, p] \overset{*}{\underset{G}{\Rightarrow}} x.$$

Por la definición de G, $S \to [q_0, Z_0, p]$ está en P, entonces,

$$(q_0, x, Z_0) \overset{*}{\underset{M}{\vdash}} (p, \lambda, \lambda) \quad \text{si y solo si} \quad S \overset{*}{\underset{G}{\Rightarrow}} x.$$

O, lo que es lo mismo

$$x \in \mathcal{L}_\lambda (M) \quad \text{si y solo si } x \in \mathcal{L} (G).$$

□

Cota en la cantidad de pasos en una derivación

Una gramática no es recursiva a izquierda si no tiene derivaciones $A \underset{L}{\overset{i}{\Rightarrow}} A\alpha$.

Lema (Lema 4.1 Aho-Ullman vol. 1)

Sea $G = (N, T, P, S)$ libre de contexto y no recursiva a izquierda. Existe una constante c tal que si $A \underset{L}{\overset{i}{\Rightarrow}} wB\alpha$ y $|w| = n$ entonces $i \leq c^{n+2}$.

Se puede demostrar un resultado mucho más ajustado, con i lineal en n, pero la misma constante.

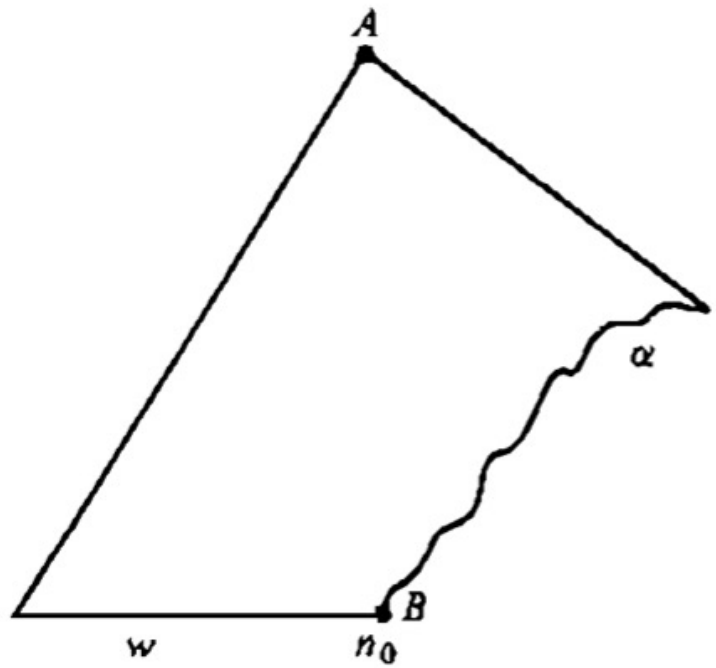
En particular, el resultado anterior vale para w igual a la cadena vacía. Así obtenemos el resultado que necesitamos para la demostración del teorema.

Corolario

Sea G libre de contexto y no recursiva. Existe una constante c tal que para todo par de no terminales A, B, si $A \underset{L}{\overset{i}{\Rightarrow}} B\alpha$ entonces $i \leq c^2$.

Demostración del lema

Llamemos k a la cantidad de símboos no-terminales.
Sea $\mathcal{A}$ el ábol de la derivación más a la izquierda para $A \underset{L}{\overset{i}{\Rightarrow}} wB\alpha$.



Sea n_0 el nodo con etiqueta B en la derivación $A \underset{L}{\overset{i}{\Rightarrow}} wB\alpha$. Notemos que, por tratarse de la derivación más a la izquierda, todos los caminos a la derecha del camino desde la raíz a n_0 son más cortos, o del mismo largo.

Supongamos que hay un camino de longitud mayor o igual que $k(n+2)$ arcos de la raíz a la hoja,

$$A = \alpha_0 \underset{L}{\Rightarrow} \alpha_1 \underset{L}{\Rightarrow} \ldots \underset{L}{\Rightarrow} \alpha_{k(n+2)-1} \underset{L}{\Rightarrow} \alpha_{k(n+2)} = wB\alpha$$

Visualicemos esta derivación por segmentos así:

$$\alpha_0 \underset{L}{\Rightarrow} \alpha_1 \underset{L}{\Rightarrow} \ldots \underset{L}{\Rightarrow} \alpha_k$$

$$\alpha_k \underset{L}{\Rightarrow} \ldots \underset{L}{\Rightarrow} \alpha_{2k}$$

$$\ldots$$

$$\alpha_{(n+1)k} \underset{L}{\Rightarrow} \ldots \underset{L}{\Rightarrow} \alpha_{(n+2)k}.$$

Son $(n+2)$ segmentos de derivaciones. Es imposible que cada uno de estos produzca uno o más símbolos de wB, porque $|wB| = n+1$. Entonces ¡hay al menos uno de estos segmentos que no produce ningun símbolo!

Entonces en el árbol de derivación $\mathcal{A}$ hay un segmento, digamos el iésimo,

$$\alpha_{ik} \underset{L}{\Rightarrow} \cdots \underset{L}{\Rightarrow} \alpha_{(i+1)k}$$

que no produce ningún símbolo de wB.

Llamemos $v_0, .., v_k$ a los vértices cuyas etiquetas son $\alpha_{ik}, .., \alpha_{(i+1)k}$. Entonces, el subarbol $v_0, .., v_k$ deriva solamente λ. Como cada uno de $\alpha_{ik}, .., \alpha_{(i+1)k}$ es una cadena de símbolos no terminales y son en total $k+1$, entonces, necesariamente hay dos que empiezan con el mismo símbolo. Pero esto contradice que la gramática no es recursiva a izquierda. Entonces nuestra suposición de que el árbol de derivación $\mathcal{A}$ tiene un camino de longitud mayor igual que $k(n+2)$ es imposible.

Sea ℓ el máximo número de símbolos en la parte derecha de una producción de la gramática. La cantidad de nodos del árbol de derivación $\mathcal{A}$ es a lo sumo

$$\ell^{k(n+2)}$$

Por lo tanto, si $A \overset{i}{\underset{L}{\Rightarrow}} wB\alpha$, entonces $i \leq \ell^{k(n+2)}$.

Para finalizar la demostración basta tomar $c = \ell^k$. □

¿Qué hace este algoritmo?

Input $G = (N, T, P, S)$ libre de contexto
$N_0 = \emptyset$
repetir
$$i = i + 1$$
$$N_i = \{A : A \to \alpha \in P, \alpha \in (N_{i-1} \cup T)^*\} \cup N_{i-1}$$
hasta que $N_i = N_{i-1}$
Si $S \in N_i$, output **sí**.

Propiedades de clausura libres de contexto

union (gramática)

Supongamos $G_1 = (N_1, T_1, P_1, S_1)$ *y* $G_2 = (N_2, T_2, P_2, S_2)$.
Definimos $G = (N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, P, S)$ *donde*
$P = P_1 \cup P_2 \cup \{S \to S_1 | S_2\}$

concatenacion (gramática)

Supongamos $G_1 = (N_1, T_1, P_1, S_1)$ *y* $G_2 = (N_2, T_2, P_2, S_2)$.
Definimos $G = (N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, P, S)$ *donde*
$P = P_1 \cup P_2 \cup \{S \to S_1 S_2\}$

clausura Kleene (gramática)

Supongamos $G = (N, T, P, S)$
Definimos $G' = (N \cup \{S'\}, T, P', S')$ *donde*
$P' = P \cup \{S' \to SS' | \lambda\}$

interseccion con lenguaje regular (autómata de pila para lenguaje intersección)

Propiedades de clausura libres de contexto

reversa (gramática invertiendo cuerpo gramática)

Supongamos $G = (V, T, P, S)$ *es libre de contexto,* $L = L(G)$.
Construimos $G' = (V, T, P', S)$ *para* L^R *así:*

Para cada producción $X \to \alpha$ *en* P *ponemos* $X \to \alpha^R$ *en* P'.

Supongamos P *tiene* $S \to uXv$ *y* $X \to \alpha$.
Entonces P' *tiene* $S \to v^R X u^R$ *y* $X \to \alpha^R$.
Luego $S \underset{G}{\Rightarrow} u\alpha v$ *y* $S \underset{G'}{\Rightarrow} v^R \alpha^R u^R$.
Dado que $v^R \alpha^R u^R = (u\alpha v)^R$, *tenemos* $S \underset{G'}{\Rightarrow} (u\alpha v)^R$.

Propiedades de clausura libres de contexto

No están clausurados por :

intersección

Sean los lenguajes libres de contexto
$L_1 = \{a^i b^j c^j\}$ *y* $L_2 = \{a^i b^i c^j\}$.
Notemos que $L_1 \cap L_2 = \{a^i b^i c^i\}$ *no es libre de contexto.*

complemento

Supongamos que el complemento fuera libre de contexto.
Entonces
$L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$ *sería libre de contexto.*

diferencia

Si lo fuera entonces $\Sigma^* - L$ *deberia ser libre de contexto*

Decisión de lenguajes libres de contexto

Sea G gramática libre de contexto y sea $L = L(G)$.

Hay algoritmos para :
$L = \emptyset$?
L finito?
L infinito?
$w \in L$? en tiempo cúbico en la longitud de w (algoritmo CYK).

No hay algoritmos para:
L es regular?
G libre de contexto es ambigua?
$L_1 = L_2$?
$L = \Sigma^*$?
$L_1 \subseteq L_2$?
$L_1 \cap L_2 = \emptyset$?