Teoría de la computación

Unidad 4. Gramáticas Independientes del Contexto (GIC)

Dr. Christian Millán

Unidad 4. GIC

Contenido

1. Definición y notación
2. Jerarquía de Chomsky
3. Derivación izquierda y derecha
4. Lenguaje de una gramática
5. Árboles de derivación
6. Aplicaciones de las GIC
7. Ambigüedad en las gramáticas y lenguajes
8. Gramáticas regulares

1. Definición y notación

Definición de las gramáticas independientes del contexto

Existen cuatro componentes importantes en una descripción gramatical de un lenguaje:

1. Un conjunto de símbolos que forman las cadenas del lenguaje que se está definiendo. $\{0,1\}$ para los palíndromos. Denominamos a este conjunto alfabeto terminal o alfabeto de símbolos terminales.

2. Un conjunto finito de variables, denominado símbolos no terminales o categorías sintácticas. Cada variable representa un lenguaje; es decir, un conjunto de cadenas. La variable$P$, que hemos empleado para representar la clase palíndromos del alfabeto {0, 1}.

1. Definición y notación

Definición de las gramáticas independientes del contexto

3. Una de las variables representa el lenguaje que se está definiendo; se denomina símbolo inicial. Otras variables representan las clases auxiliares de cadenas que se emplean para definir el lenguaje del símbolo inicial. En el ejemplo, P es el símbolo inicial y única variable.
4. Un conjunto de producciones o reglas que representan la definición recursiva de un lenguaje. Cada producción costa de:
    a. Una variable que define (parcialmente) la producción. Esta variable a menudo se denomina cabeza de la producción.
    b. El símbolo de la producción→
    c. Una cadena formada por cero o mas símbolos terminales y variables. Esta cadena, denominada cuerpo de la producción, representa la manera de formar cadenas pertenecientes al lenguaje de la variable de la cabeza. De este modo, dejamos los símbolos terminales invariables y sustituimos cada una de las variables del cuerpo por una cadena que sabemos que pertenece al lenguaje de dicha variable.

1. Definición y notación

Definición de las gramáticas independientes del contexto

Representamos una GIC G mediante cuatro componentes, es decir,$G = (V, T, P, S)$, donde V es el conjunto de variables, T son los símbolos terminales, P es el conjunto de producciones y S es el símbolo inicial.

1. Definición y notación

Definición de las gramáticas independientes del contexto

Ejemplo:

La gramática G_{pal} para los palíndromos se representa como sigue:

$G_{pal} = (\{P\}, \{0, 1\}, A, P)$

donde A representa el conjunto de las cinco producciones:

1. $P \to \varepsilon$
2. $P \to 0$
3. $P \to 1$
4. $P \to 0P0$
5. $P \to 1P1$

1. Definición y notación

Definición de las gramáticas independientes del contexto

Ejemplo 2: Gramática independiente del contexto para expresiones simples.

$G = (\{E, I\}, T, P, E)$

$T = +, *, (,), a, b, 0, 1$

$P =$

$E \to I$

$E \to E + E$

$E \to E * E$

$E \to (E)$

$I \to a$

$I \to b$

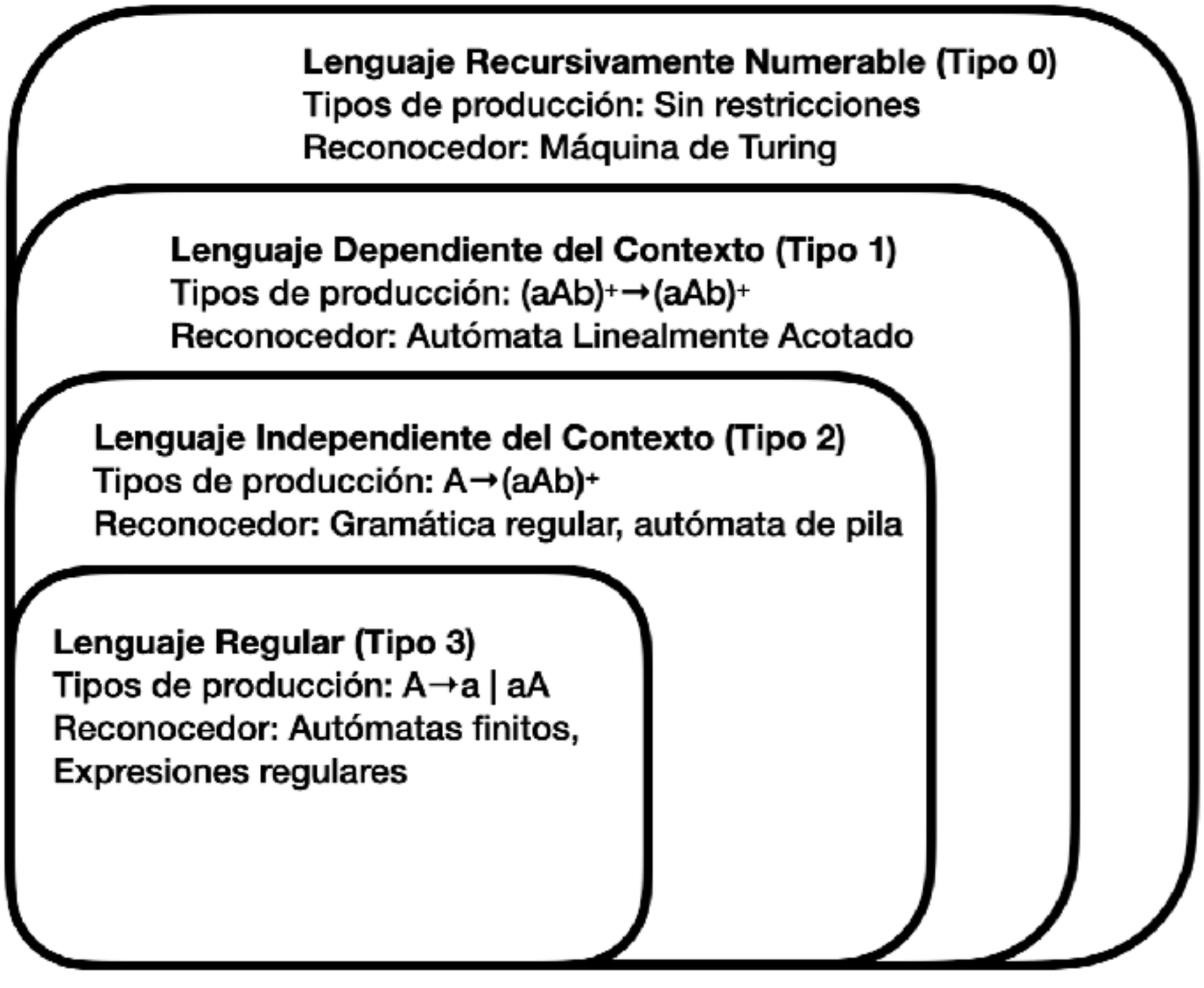
$I \to Ia$

$I \to Ib$

$I \to I0$

$I \to I1$

2. Jerarquía de Chomsky



3. Derivación izquierda y derecha

Derivación

El proceso de derivación de cadenas aplicando producciones dede la cabeza hasta el cuerpo requiere la definición de un nuevo símbolo de relación $\Rightarrow$. Supongamos que $G = (V, T, P, S)$ es un GIC. Sea $\alpha A \beta$ una cadena de símbolos terminales y variables, Siendo A una variable. Es decir, α y β son cadenas de $(V \cap T)*$ y A pertenece a V. Sea $T \rightarrow \gamma$ una producción de G. Entonces decimos que $\alpha A \beta \underset{G}{\Longrightarrow} \alpha A \beta$.

Si estamos trabajando con G, sólo podemos decir que $\alpha A \beta \Rightarrow \alpha \gamma \beta$. Una derivación reemplaza cualquier variable de cualquier parte de la cadena por el cuerpo de una de sus producciones.

Podemos extender la relación $\Rightarrow$ para representar cero, uno o más pasos de derivaciones, del mismo modo que hemos extendido la función de transición δ de una automata infinito a $\hat{\delta}$. Para las derivaciones utilizaremos el símbolo * para indicar "cero o más pasos".

3. Derivación izquierda y derecha

Derivación por izquierda

Ejemplo: la inferencia de que $a * (a + b00)$ esa en el lenguaje de la variable E se puede reflejar en la derivación de dicha cadena, a partir del símbolo inicial E. Observar que en el reemplazo la política es siempre tomar la variable más a la izquierda de la cadena (derivación más a la izquierda).

$$E \underset{lm}{\Longrightarrow} E * E \underset{lm}{\Longrightarrow} I * E \underset{lm}{\Longrightarrow} a * E \underset{lm}{\Longrightarrow}$$

$$a * (E) \underset{lm}{\Longrightarrow} a * (E + E) \underset{lm}{\Longrightarrow} a * (I + E) \underset{lm}{\Longrightarrow} a * (a + E) \underset{lm}{\Longrightarrow}$$

$$a * (a + I) \underset{lm}{\Longrightarrow} a * (a + I0) \underset{lm}{\Longrightarrow} a * (a + I00) \underset{lm}{\Longrightarrow} a * (a + b00)$$

Podemos emplear la relación $E \overset{*}{\underset{lm}{\Longrightarrow}} a * (a + b00)$ para indicar la condensación de la derivación.

3. Derivación izquierda y derecha

Derivación por derecha

En una derivación más a la derecha se utiliza la misma sustitución de cada variable, pero en un orden diferente:

$$E \underset{rm}{\Longrightarrow} E * E \underset{rm}{\Longrightarrow} E * (E) \underset{rm}{\Longrightarrow} E * (E + E) \underset{rm}{\Longrightarrow}$$

$$E * (E + I) \underset{rm}{\Longrightarrow} E * (E + I0) \underset{rm}{\Longrightarrow} E * (E + I00) \underset{rm}{\Longrightarrow} E * (E + b00) \underset{rm}{\Longrightarrow}$$

$$E * (I + b00) \underset{rm}{\Longrightarrow} E * (a + b00) \underset{rm}{\Longrightarrow} I * (a + b00) \underset{rm}{\Longrightarrow} a * (a + b00)$$

Esta derivación nos permite concluir que $E \underset{rm}{\Longrightarrow} a * (a + b00)$.

Para cualquier derivación existe una derivación más a la izquierda equivalente y una derivación más a la derecha equivalente. Es decir, si w es una cadena terminal y A es una variable, entonces $A \overset{*}{\Longrightarrow} w$ si y sólo si $A \underset{lm}{\overset{*}{\Longrightarrow}} w$, y $A \overset{*}{\Longrightarrow} w$ si y sólo si $A \underset{rm}{\overset{*}{\Longrightarrow}} w$.

4. Lenguaje de una gramática

Si $G = (V, T, P, S)$ es una GIC, el lenguaje de G designado como L(G) es el conjunto de cadenas terminales que tienen derivaciones desde el símbolo inicial. Es decir,

$$L(G) = \{w \text{ pertenece a } T^* | S \overset{*}{\underset{G}{\Longrightarrow}} w\}$$

Si un lenguaje L es el lenguaje de cierta gramática independiente del contexto, entonces se dice que L es un lenguaje independiente del contexto o LIC (CFL, context-free language).

Por ejemplo, hemos dicho que la gramática de palíndromos es un lenguaje independiente del contexto. Podemos demostrar esta proposición de la forma siguiente.

5. Árboles de derivación

Existe una representación en árbol para las derivaciones, este árbol muestra cómo se agrupan los símbolos de una cadena terminal en subcadenas, que pertenecen al lenguaje de una de las variables de la gramática. Es conocido como árbol de derivación.

Los árboles de derivación están extremadamente relacionados a la existencia de derivaciones y las inferencias recursivas.

5. Árboles de derivación

Construcción de árboles de derivación

Sea $G = (V, T, P, S)$ es una grámatica. Los árboles de derivación para G son aquellos árboles que cumplen con las condiciones siguientes:

1. Cada nodo está etiquetado con una variable de V
2. Cada hoja está etiquetada bien con una variable, un símbolo terminal o ε. Sin embargo, si la hoja está etiquetada con ε, entonces tiene que ser el único hijo de su padre.
3. Si un nodo interior está etiquetado como A y sus hijos están etiquetados como:

$X_1, X_2, \ldots, X_k$

respectivamente, comenzando por la izquierda, entonces $A \to X_1 X_2 \cdots X_k$ es una producción de P. Observe que el único caso en que una de las X puede reemplazarse por ε es cuando es la etiqueta del único hijo y $A \to \varepsilon$ es una producción de G.

5. Árboles de derivación

Resultado de un árbol de derivación

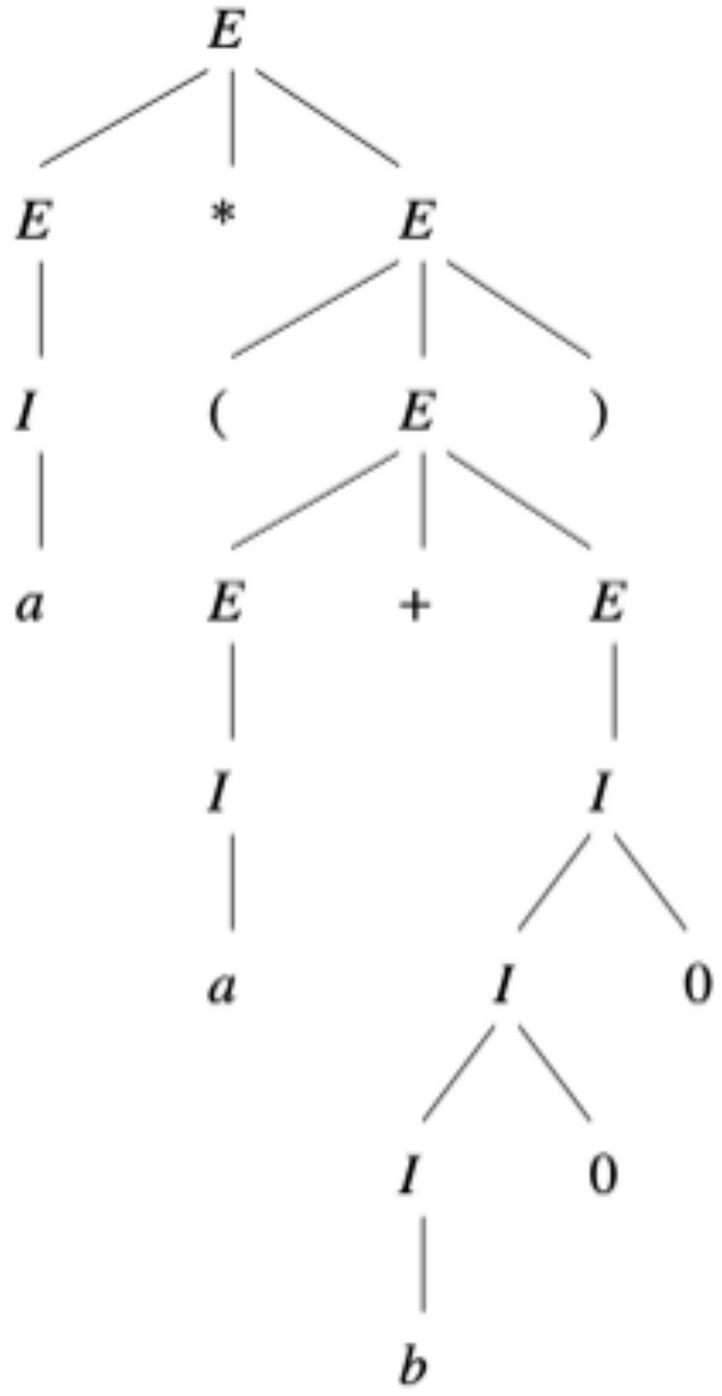
Al concatenar las hojas de un árbol de derivación por la izquierda obtenemos una cadena denominada resultado del árbol.

Los árboles de derivación:

1. El resultado es una cadena terminal. Es decir, todas las hoja están etiquetadas con un símbolo terminal o con ε
2. La raíz está etiquetada con el símbolo inicial

Ejemplo de árbol de derivación

5. Árboles de derivación



6. Aplicaciones de las GIC

Las gramáticas independientes de contexto fueron concebidas por N. Chomsky como una forma de describir los lenguajes naturales. Esta posibilidad no ha llegado a cumplirse.

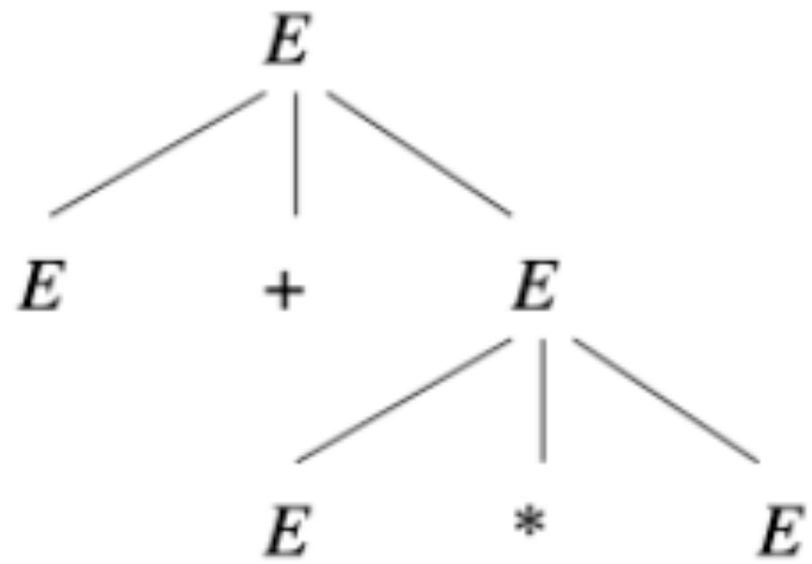
1. Las gramáticas se utilizan para describir lenguajes de programación

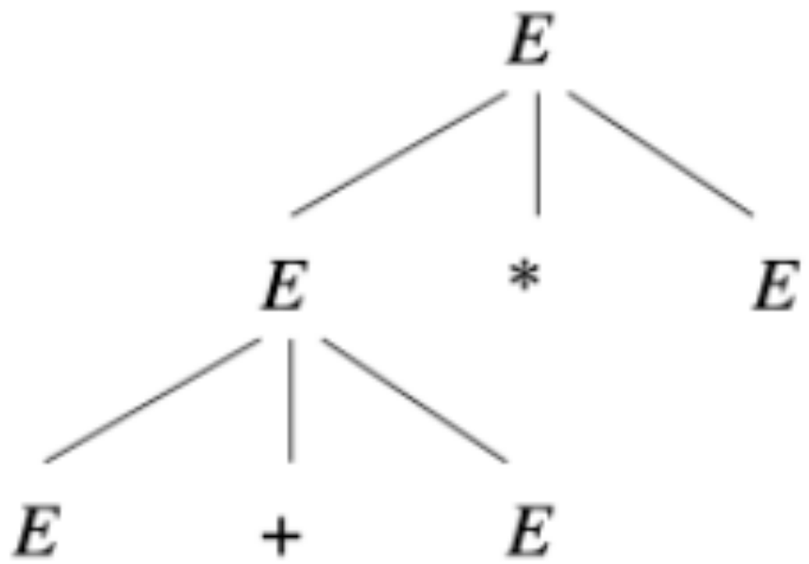
2. El desarrollo de XML (Extensible Markup Language) facilitará el comercio electrónico permitiendo a los participantes compartir convenios, independientemente de los pedidos, descripciones de los productos y de otros muchos tipos de documentos. Una parte fundamental de XML es la DTD (Document Type Definition, definición de tipo de documentos), que principalmente es una gramática independiente del contexto que describe las etiquetas permitidas y las formas que dichas etiquetas pueden anidarse. Las etiquetas XML no se encargan de dar formato al texto, sino del significado del mismo.

7. Ambigüedad en las gramáticas y lenguajes

La suposición tácita era que una gramática determina de manera unívoca una estructura sea única para cada cadena del lenguaje. Sin embargo, no todas las gramáticas proporcionan estructuras únicas. Es posible en algunas ocasiones rediseñar la gramática para hacer que la estructura sea única para cada cadena de lenguaje. En algunas ocasiones esto no será posible porque existe algunos lenguajes independientes del contexto que son inherentemente ambiguos.

7. Ambigüedad en las gramáticas y lenguajes





8. Gramáticas regulares

Las gramáticas regulares (aquellos reconocidos por un autómata finito). Son las gramáticas más restrictivas. El lado derecho de una producción debe contener un símbolo terminar y cómo máximo un símbolo no terminal.

8. Gramáticas regulares

Estas gramáticas pueden ser:

- Lineales a la derecha, si todas las producciones son de la forma: $A \to aB$ o $A \to a$; $A \in N \cup S; B \in N; a \in T$. En el lado derecho de las producciones el símbolo no terminal parece a la derecha del símbolo termina.
- Lineales a la izquierda, si todas las producciones son de la forma: $A \to Ba$ o $A \to a$; $A \in N \cup S; B \in N; a \in T$. En el lado derecho de las producciones el símbolo no terminal parece a la izquierda del símbolo termina.

8. Gramáticas regulares

En ambos casos, se puede incluir la producción $S \to \varepsilon$, si el lenguaje que se quiere generar contiene la cadena vacía.

Por ejemplo las siguientes gramáticas G_1 y G_2, son gramáticas regulares lineales a la derecha y lineales a la izquierda respectivamente, que generan el lenguaje $L = \{a^{2n} | n \geq 0\}$

$G_1 = (\{A, B\}, \{a\}, P_1, S_1)$

donde P_1 es el conjunto:

$S_1 \to \varepsilon$

$S_1 \to aA$

$A \to aB$

$A \to a$

$B \to aA$

$G_2 = (\{C, D\}, \{a\}, P_2, S_2)$

donde P_2 es el conjunto:

$S_2 \to \varepsilon$

$S_2 \to Ca$

$C \to Da$

$C \to a$

$D \to Ca$

8. Gramáticas regulares

Algoritmo para obtener la gramática regular desde el autómata finito

Existe algún algoritmo que permite obtener una gramática regular que genera un lenguaje regular dado a partir del autómata finito que reconoce ese lenguaje. Los pasos a seguir son los siguientes:

1. Asociar al estado inicial el símbolo distinguido S.
2. Asociar a cada estado del autómata (menos al estado inicial) un símbolo no terminal. Si al estado inicial llega algún arco asociar también un símbolo no terminal (además del símbolo distinguido). No asociar símbolo no termina a aquellos estados finales de los que no salen arcos.
3. Para cada transición definida $\delta(e_i, a) = e_j$, agregar al conjunto de producciones, la producción $A \to aB$, siendo A y B los símbolos no terminales asociados a e_i y e_j respectivamente. Si e_j es una estado final, agregar también la producción $A \to a$. Si e_j es el estado inicial (tiene dos símbolos asociados, el distinguido y un no terminal), utilizar el símbolo no terminal (de esta manera se evita que el símbolo distinguido aparezca a la derecha de una producción).
4. Si el estado inicial es también final agregar la producción $A \to \varepsilon$.