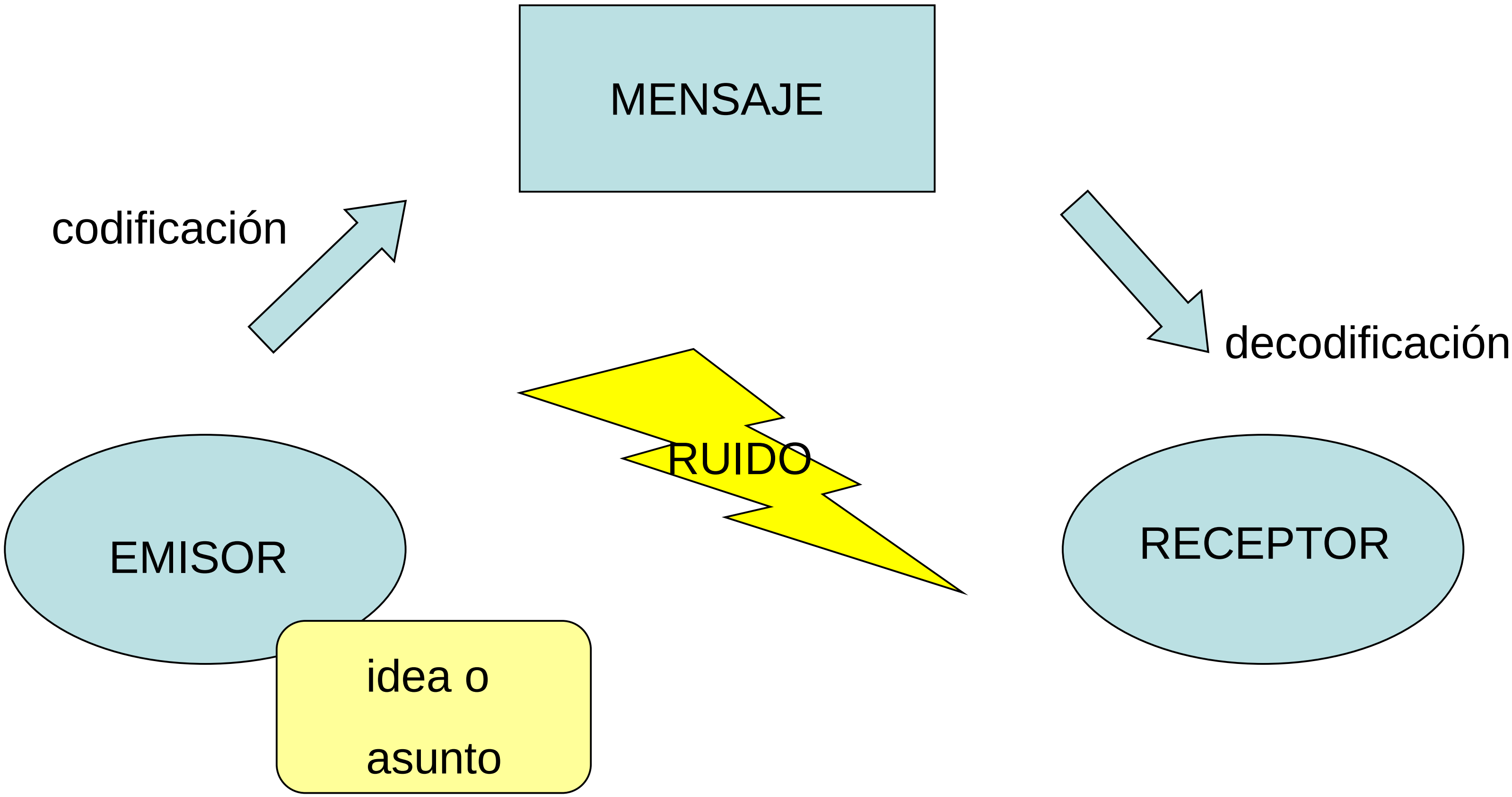
Lógica y Matemática Computacional
Licenciatura en Sistemas de Información

Lenguajes Formales – Gramáticas
Jerarquía de Chomsky

Ing. JULIO C. ACOSTA

Comunicación



LENGUAJES

Lenguaje es vehículo de comunicación entre

personas - personas

personas – máquinas   máquinas - personas

personas – animales ??

LENGUAJES

- Lenguaje Vulgar

- Lenguaje de Ciencia

- Lenguaje Formal

Lenguajes Formales

Un lenguaje formal está conformado por cadenas de símbolos y cada cadena debe tener un significado “preciso e inequívoco”.

Un lenguaje posee: Vocabulario y Gramática

Vocabulario son los “bloques o signos” a partir de los cuales se construyen las cadenas del lenguaje

Ejemplos de cadenas de lenguajes formales

L(1)={a,b}

L(2)={a,b,ab,ba,aa,bb,aba,bab}

L(3)={aaa,bbb,abb,baa,aab,bba,aba,bab}

L(1)={a,b,ab,ba}   es un lenguaje de 4 palabras (cadenas)

L(2)={a,b,ab,ba,aa,bb,aba,bab}

L(3)={aaa,bbb,abb,baa,aab,bba,aba,bab}

El lenguaje L(3) se puede escribir L(3) = $\{a,b\}^3$ que significa lenguaje conformado por todas las cadenas de tres símbolos que pueden formarse con los terminales a y b

Longitud de una cadena es la cantidad de símbolos que contiene la misma.

Cada cadena tiene un significado en cada uno de los lenguajes

Ejemplo de gramática del lenguaje español (no formal)

<oración>→<sujeto> <predicado>
<sujeto> → <sustantivo>
<predicado> →<verbo><adjetivo>
<sustantivo> →Juan
<sustantivo> →María
<verbo> → corre
<verbo> → sube
<adjetivo> → rápido
<adjetivo> → antes

<oración>→<sujeto> <predicado> (1)
<sujeto> → <sustantivo> (2)
<predicado> →<verbo><adjetivo> (3)
<sustantivo> →Juan (4)
<sustantivo> →María (5)
<verbo> → corre (6)
<verbo> → sube (7)
<adjetivo> → rápido (8)
<adjetivo> → antes (9)

Se generan cadenas como:
<oración> → <sujeto> <predicado> → <sustantivo> <predicado> → <sustantivo><verbo><adjetivo> → Juan <verbo><adjetivo> → Juan corre <adjetivo> → Juan corre rápido

<oración> → <sujeto> <predicado> → <sustantivo> <predicado> → <sustantivo><verbo><adjetivo> → Juan <verbo><adjetivo> → Juan sube <adjetivo> → Juan sube rápido

<oración>→<sujeto> <predicado> (1)
<sujeto> → <sustantivo> (2)
<predicado> →<verbo><adjetivo> (3)
<sustantivo> →Juan (4)
<sustantivo> →María (5)
<verbo> → corre (6)
<verbo> → sube (7)
<adjetivo> → rápido (8)
<adjetivo> → antes (9)

Se generan cadenas como:
<oración> → <sujeto> <predicado> → <sustantivo> <predicado> → <sustantivo><verbo><adjetivo> → María <verbo><adjetivo> → María corre <adjetivo> → María corre antes

<oración> → <sujeto> <predicado> → <sustantivo> <predicado> → <sustantivo><verbo><adjetivo> → María <verbo><adjetivo> → María corre <adjetivo> → María corre rápido

No es cadena del lenguaje por ejemplo: “corre Juan antes”

Lenguajes formales

Gramática son las reglas que combinan los símbolos del vocabulario para generar el lenguaje.

Las reglas de la gramática se valen de producciones donde se involucran símbolos terminales y no terminales

Símbolos terminales son los símbolos que forman las cadenas del lenguaje.

Símbolos no terminales son auxiliares para la formación de las cadenas

Formalmente, la gramática para un lenguaje formal se define en un cuádruple (N,T,P,s_0) donde:

N: conjunto finito de símbolos no terminales

T: conjunto finito de símbolos terminales

P: conjunto finito de producciones

$s_{0:}$ símbolo inicial

Ejemplos de cadenas de lenguajes formales

L(1)={a,b,ab,ba}

L(2)={a,b,ab,ba,aa,bb,aba,bab}

L(3)={aaa,bbb,abb,baa,aab,bba,aba,bab}

Lenguaje que representa la ocurrencia del lanzamiento de de una moneda dos veces

$L(G)=\{cc,cs,sc,ss\}$

$G(N,T,P,s_0)$ donde

$N= \{s_0\}$ $T=\{c,s\}$

$P = \{s_0 \rightarrow cc, s_0 \rightarrow cs, s_0 \rightarrow sc, s_0 \rightarrow ss\}$ mejoramos P

$N= \{s_0, A\}$ $T=\{c,s\}$

$P = \{s_0 \rightarrow AA, A \rightarrow c, A \rightarrow s\}$

Una Gramática para estructuras de expresiones G es un cuádruple (N, T, s_0, P), donde:

N es un *conjunto finito de símbolos no terminales o símbolos auxiliares*.

T es un *conjunto finito de símbolos terminales*.

s_0 es un *símbolo inicial* , $s_0 \in N$

P es un *sub conjunto finito* del producto $[(N \cup T)^* - T] \times (N \cup T)^*$ llamado conjunto de producciones.

Tenga presente que $N \cap T = \emptyset$

Algunos autores expresan la gramática G = (V, S, v_0, →) donde:

S es conjunto de símbolos terminales, también llamado T

v_0 es el símbolo inicial, también llamado s_0

→ es la relación determinada en el conjunto producción, que en este trabajo llamamos P.

Todo esto no merece mayores comentarios, pero V es T $\cup$ N,

Así, definida una gramática G = (V, S, v_0, →) a partir de:

V = { v_0, W, a, b, c }   S = { a, b, c }   y

una relación en V*

que verifica: $v_0 \rightarrow a\ W$; $W \rightarrow b\ b\ W$; $W \rightarrow c$

la misma gramática $G = (N, T, s_0, P)$ donde:

$N = \{ s_0, W \}$ $T = \{ a, b, c \}$ y $P = \{ s_0 \rightarrow a\ W, W \rightarrow b\ b\ W, W \rightarrow c \}$

nominaremos

con letra minúscula a los símbolos terminales (T),

con letra mayúscula a los símbolos no terminales (N),

con s_0 el símbolo inicial (cuando no se especifique explícitamente otra cosa) y

con P las producciones que componen la gramática.

Los símbolos consecutivos pueden aparecer separados por un espacio, lo que no significa la presencia del espacio como símbolo, cuando esto suceda, el espacio aparecerá explícito como “_”

Ejemplo Con la gramática antes definida, hallar al menos dos cadenas terminales de L(G), y si es posible deducir una expresión general de L(G).

$N = \{ s_0, W \}$ $T = \{ a, b, c \}$ y $P = \{ s_0 \rightarrow a\ W,\ W \rightarrow b\ b\ W,\ W \rightarrow c \}$

Desarrollo:

$s_0 \Rightarrow a\ W \Rightarrow a\ c$

$s_0 \Rightarrow a\ W \Rightarrow a\ b\ b\ W \Rightarrow a\ b\ b\ c$

También puede obtenerse

$s_0 \Rightarrow a\ W \Rightarrow a\ b\ b\ W \Rightarrow a\ b\ b\ b\ b\ W \Rightarrow a\ b\ b\ b\ b\ c = a\ b^4\ c$

donde b^4 indica la repetición de b 4 veces

son cadenas del lenguaje L(G)

$$L(G) = \{ a\ b^{2n}\ c\ ;\ n \geq 0 \}$$

Este lenguaje está formado por todas las cadenas que comienzan con un símbolo “a”, terminan en un símbolo “c” y en el medio tienen un número par de símbolos “b” o no tienen ningún símbolo b.

Ejemplo Sea la gramática $G = (N, T, s_0, P)$ donde:

$N = \{s_0, W\}$ $T = \{ a, b, c \}$ $P = \{ s_0 \rightarrow a\, s_0\, b \,;\, s_0\, b \rightarrow b\, W \,;\, a\, b\, W \rightarrow c \}$

hallar al menos dos cadenas terminales de L(G), y si es posible deducir una expresión general de L(G).

Desarrollo:

$s_0 \Rightarrow a\, s_0\, b \Rightarrow a\, b\, W \Rightarrow c$

$s_0 \Rightarrow a\, s_0\, b \Rightarrow a\, a\, s_0\, b\, b \Rightarrow a\, a\, b\, W\, b \Rightarrow a\, c\, b$

$s_0 \Rightarrow a\, s_0\, b \Rightarrow a\, a\, s_0\, b\, b \Rightarrow a\, a\, a\, s_0\, b\, b\, b \Rightarrow a\, a\, a\, b\, W\, b\, b \Rightarrow a\, a\, c\, b\, b$

Este lenguaje está formado por todas las cadenas que comienzan con una o mas “a”, terminan con una o mas “b” y en el medio siempre tienen una “c”. También la cadena formada por un solo símbolo “c”.

Entonces: $L(G) = \{ a^n\, c\, b^n \,;\, n \geq 0 \}$

Ejemplo Construir una gramática G para el lenguaje L(G):

$$L(G) = \{ a\ a\ a\ a,\ a\ a\ b\ b,\ b\ b\ a\ a,\ b\ b\ b\ b \}$$

Desarrollo :

Reconocemos los símbolos no terminales contenidos en el lenguaje cuya gramática queremos construir y expresamos el conjunto

$$T = \{ a, b \}$$

nos valemos del símbolo inicial no terminal s_0 para poder derivar, y generamos la producción:

$$P = \{ s_0 \rightarrow a\ a\ a\ a,\ s_0 \rightarrow a\ a\ b\ b,\ s_0 \rightarrow b\ b\ a\ a,\ s_0 \rightarrow b\ b\ b\ b \}$$

con $N = \{ s_0 \}$ queda definida la gramática G (N, T, s_0, P)

podemos tener una gramática mas sencilla si agregamos al conjunto N, el símbolo no terminal A y definimos la producción:

$P = \{ s_0 \rightarrow A\ A\ ;\ A \rightarrow a\ a\ ;\ A \rightarrow b\ b \}$ $N = \{ s_0, A \}$ $T = \{ a, b \}$

Ejemplo Construir una gramática para el lenguaje $L(G) = \{ a^i b^{2i} \mid i \geq 1 \}$

Desarrollo:

$L(G) = \{ a\,b\,b,\ a\,a\,b\,b\,b\,b,\ a\,a\,a\,b\,b\,b\,b\,b\,b, \ldots \}$

Los símbolos terminales conforman el conjunto $T = \{ a, b \}$

las cadenas que conforman L(G) están compuestas por el doble de símbolos b que símbolos a ; proponemos una primera producción

$s_0 \rightarrow a\, s_0\, b\, b$

y una segunda producción que se aplicará para finalizar el ciclo de derivaciones

$s_0 \rightarrow a\, b\, b$

así el conjunto $P = \{ s_0 \rightarrow a\, s_0\, b\, b;\ s_0 \rightarrow a\, b\, b \}$

los símbolos no terminales $N = \{ s_0 \}$ y la gramática G será: $G = (N, T, s_0, P)$

Ejemplo Construir una gramática (G) para el lenguaje

$$L(G) = \{ a^i b^j \mid i, j \geq 1, i \neq j \}$$

Desarrollo:
Este lenguaje está formado por todas las cadenas que comienzan con una o mas “a”, terminan con una o mas “b” con la condición que la cantidad de “a” y de “b” son distintas.

Podemos pensarlo al L(G) como L1 $\cup$ L2 donde:

$L1 = \{ a^i b^j \mid i > j \}$ y $L2 = \{ a^i b^j \mid i < j \}$

reconocemos una producción en una gramática para L1 donde:

T = { a, b } y N = { A, B } como:

$A \to a\ A$; $A \to a\ B$; $B \to a\ B\ b$ $B \to a\ b$

reconocemos también una producción en una gramática para L2 donde:

$T = \{ a, b \}$ y $N = \{ C, D \}$ como:

$C \rightarrow C\ b;$ $C \rightarrow D\ b;$ $D \rightarrow a\ D\ b;$ $D \rightarrow a\ b$

es necesario una producción que dé inicio al “sistema”, así introducimos:

$s_0 \rightarrow A$ $s_0 \rightarrow C$

Considerando que $D \rightarrow a\ b$ y $B \rightarrow a\ b$; $D \rightarrow a\ D\ b$ y $B \rightarrow a\ B\ b$ son equivalentes, podemos reemplazar el símbolo no terminal D por el símbolo no terminal B.

$P = \{ s_0 \rightarrow A;\ s_0 \rightarrow C;\ A \rightarrow a\ A;\ A \rightarrow a\ B;\ B \rightarrow a\ B\ b;\ B \rightarrow a\ b;\ C \rightarrow C\ b;\ C \rightarrow B\ b \}$

$T = \{ a, b \}$ $N = \{ s_0, A, B, C \}$

Esta será la gramática $G = (N, T, s_0, P)$

JERARQUIA DE CHOMSKY

Gramática Tipo 3

Son producciones que tienen a la izquierda solo un símbolo NO terminal y a la derecha símbolos terminales, o a lo sumo un no terminal que se ubica a la derecha de la cadena.

$A \rightarrow a$ $\quad B \rightarrow bc$ $\quad C \rightarrow Ba$ $\quad B \rightarrow Ab$

Gramática Tipo 3 (Regulares)

Son una restricción de las gramáticas libres de contexto.

Pueden contener a la izquierda de sla producción solo un símbolo N, a la izquierda o a la derecha de la cadena

$A \rightarrow Ba$ o $A \rightarrow a$

$A \rightarrow Ba$ o $A \rightarrow abc$

JERARQUIA DE CHOMSKY

Gramática Tipo 2

Son producciones que tienen a la izquierda solo un símbolo NO terminal y a la derecha símbolos terminales y no terminales

$A \rightarrow aBBb$ $B \rightarrow Bbc$ $C \rightarrow BacB$ $B \rightarrow BA$

Gramáticas Tipo 2 – Libres de contexto

No contiene producciones del tipo $\varphi A \psi \rightarrow \varphi \omega \psi$

donde la cadena $\varphi \omega \psi$ aparecerá solo si al símbolo A le precede φ y le sucede ψ

Sus producciones son del tipo $A \rightarrow \omega$ donde

$A \in N$ y $\omega \in \{ N \cup T \}^*$

Las gramáticas de tipo 2 se usan frecuentemente en lenguajes artificiales, en especial en lenguajes de programación

JERARQUIA DE CHOMSKY

Gramática Tipo 1

Son producciones que tienen a la izquierda símbolos terminales y no terminales y a la derecha símbolos terminales y no terminales, con la restricción de ser producciones del tipo

$\alpha \rightarrow \beta$ donde $long(\alpha) \leq long(\beta)$

$aA \rightarrow aBBb$ $AB \rightarrow Bbc$ $CaB \rightarrow BacB$

Gramáticas de Tipo 1
(Sensibles al contexto)

Contienen producciones del tipo $\varphi A \psi \rightarrow \varphi \omega \psi \quad \omega \neq \lambda$

Sus producciones $\alpha \rightarrow \beta$ verifican $|\alpha| \leq |\beta|$

En consecuencia $\omega_1 \Rightarrow \omega_2 \Rightarrow \ldots \Rightarrow \omega_n$

Con longitud $|\omega_1| \leq |\omega_2| \leq \ldots \leq |\omega_n|$

La longitud de las cadenas no decrecen

JERARQUIA DE CHOMSKY

Gramática Tipo 0

Son producciones de cualquier tipo

$BBaA \rightarrow aB$ $\quad BaaB \rightarrow bc$ $\quad CB \rightarrow a$ $\quad aBa \rightarrow BA$

Los tipos de gramática en la jerarquía de Chomsky son abarcativos, es decir:

$$\text{Tipo 3} \subset \text{Tipo 2} \subset \text{Tipo 1} \subset \text{Tipo 0}$$

Gramáticas Tipo 0 (sin restricciones)

Por ser muy generales pierden fuerza descriptiva-
Sus cadenas no son interesantes como lenguajes artificiales
Son interesantes para las reglas de deducción en lógica.

Su importancia se centra en que: $\varphi A \psi \rightarrow \varphi \lambda \psi$ entonces

$long(\omega_i) < long(\omega_{i+1})$ la cadena de deducción se contrajo

Producciones que contraen son del tipo $aC \rightarrow b$

Tipo	Formato de Producciones		Comentarios	
3	$B \to aA$ $A \to a$ $s_0 \to \lambda$	Lineales por izquierda	Gramáticas regulares	Gramática que no contrae
	$B \to Ab$ $A \to a$ $s_0 \to \lambda$	Lineales por derecha		
2	$A \to \omega$	$\omega \neq \lambda$	Gramáticas Libres del contexto	
1	$\varphi A \psi \to \varphi \omega \psi$	$\omega \neq \lambda$	Gramáticas sensibles al contexto	
0	$\varphi A \psi \to \varphi \omega \psi$		Gramáticas sin restricción	Gramática que contrae

NOTACION BNF (FORMA DE BACKUS-NAUR)

Una alternativa que poseen las gramáticas de tipo 2 (y tipo 3), para desplegar sus producciones es la notación BNF.

1) Se juntan todas las producciones que tienen a un mismo no terminal del lado izquierdo (por ejemplo A), en la misma línea, colocando el no terminal referido (A) del lado izquierdo y del lado derecho todos los terminales y no terminales de la producción, en serie, separados solamente por una barra vertical | (se lee "ó") que no aporta ningún otro significado que el hecho de separar símbolos.

2) El símbolo → se reemplaza por el símbolo : := que se lee "puede ser".

3) Cada símbolo no terminal, cuando aparezca, será encerrado entre paréntesis agudos < >. Esto tiene la ventaja que los símbolos no terminales, pueden tener espacios dentro de ellos, por ejemplo: palabra1 palabra2 se considera un solo símbolo y no dos. Se usa el espacio como una "letra o símbolo" conveniente.

Gramática para generar los números enteros

N = {s_0, signo, entero, cifra} T = {1,2,3,4,5,6,7,8,9,0}

s_0 ::= <signo><entero>

<signo> ::= + | -

<entero> ::= <entero> <cifra>|<cifra>

<cifra>::= 1| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 |

Generamos el -1024

s_0 ::= <signo><entero>::= - <entero><cifra>::= - <entero> <cifra> <cifra> ::=

- <entero> <cifra> <cifra> <cifra> ::= - <cifra> <cifra> <cifra> <cifra> ::=

  - 1 <cifra> <cifra> <cifra>::= - 1 0 <cifra> <cifra> ::= - 1 0 2 <cifra>

  ::= - 1 0 2 4

Pasar las siguientes gramáticas a la notación BNF

$N = \{ s_0, W \}$ $T = \{ a, b, c \}$ y $P = \{ s_0 \rightarrow a\ W, W \rightarrow b\ b\ W, W \rightarrow c \}$

$N = \{ s_0, A, B, C \}$ $T = \{ a, b \}$

$P = \{ s_0 \rightarrow A; s_0 \rightarrow C; A \rightarrow a\ A; A \rightarrow a\ B; B \rightarrow a\ B\ b; B \rightarrow a\ b; C \rightarrow C\ b; C \rightarrow B\ b \}$