UNIVERSIDAD DE SAN BUENAVENTURA CALI

Guía
Modelos de Computación

Tema I: Lenguajes y Gramáticas

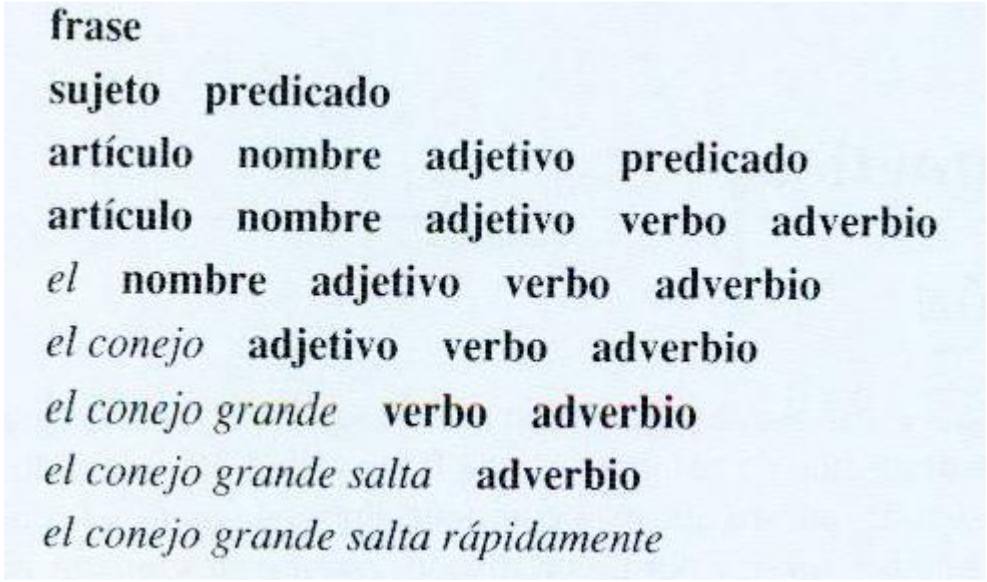
Introducción

La sintaxis de un lenguaje natural, esto es, la de los lenguajes hablados, como el inglés, el español, el alemán o el francés, es extremadamente complicada. De hecho, parece imposible especificar todas las reglas de sintaxis de un lenguaje natural. Las investigaciones en el área de la traducción automática de un lenguaje a otro ha dado lugar al concepto de lenguaje formal, que, a diferencia del lenguaje natural, esta especificado por un conjunto de reglas bien definidas. Las reglas sintácticas son importantes no solo en la filología, que es el estudio del lenguaje natural, sino también en e estudio de los lenguajes de programación. Describiremos las frases de un lenguaje formal utilizando la gramática. El uso de la gramática es de gran ayuda cuando se trata de resolver las dos clases de problemas que aparecen con mucha frecuencia en las aplicaciones a los lenguajes de programación: 1. ¿Cómo se puede determinar cuando una combinación de palabras es una frase valida en un lenguaje formal? 2. ¿Cómo se pueden generar las frases validas de un lenguaje formal? Antes de dar una definición formal de lo que es la gramática describiremos un ejemplo de gramática que genera un subconjunto del español. Definimos este subconjunto usando una lista de reglas que describe como construir una frase valida concretamente:

1. Una frase se compone de un sujeto seguido de un predicado;
2. Un sujeto se compone de un articulo seguido de un nombre seguido de un adjetivo, o
3. Un sujeto se compone de un artículo seguido de un nombre;
4. Un predicado se compone de un verbo seguido de un adverbio o
5. Un predicado se compone de un verbo
6. Un articulo es *un*, o
7. Un articulo es *el*;
8. Un adjetivo es *grande*, o
9. Un adjetivo es *hambriento*;
10. Un nombre es *conejo*, o
11. Un nombre es *matemático*;
12. Un verbo es *come*, o
13. Un verbo es *salta*;

14. Un adverbio es *rápidamente*, o
15. Un adverbio es *salvajemente*.

A partir de estas reglas, podemos formar frases realizando una serie de reemplazamientos hasta que no podamos aplicar ninguna regla más. Por ejemplo, para obtener una frase valida podemos realizar la siguiente secuencia de sustituciones:



También es sencillo comprobar que son enunciados validos *un matemático hambriento come salvajemente*, *un enorme matemático salta*, *el conejo come rápidamente*, etc. Así mismo, se puede comprobar que no es valida la frase *el rápidamente come matemático*.

Gramáticas Con Estructura De Frases

Antes de dar la definición formal de una gramática introducimos algunos términos.

Definición 1.
Un vocabulario (o alfabeto) V es un conjunto finito y no vació, cuyos elementos se llaman símbolos. Una palabra sobre V es una cadena finita de elementos de V. La palabra vacía o cadena vacía denotada por λ, es la cadena sin símbolos. El conjunto de todas las palabras sobre V se denota por V^*. Un lenguaje sobre V es un subconjunto de V^*.

Nótese que λ, la palabra vacía, es la cadena que no contiene símbolos, y es diferente del conjunto vació $\emptyset$. Por tanto, $|\lambda|$ es el conjunto que contiene exactamente una palabra, la palabra vacía. Los lenguajes se pueden especificar de varias formas. Una de ellas consiste en enumerar todas las palabras que conforman el lenguaje. Otra es dar algunos criterios que una palabra debe satisfacer para pertenecer al lenguaje. Se describe otro modo importante de especificar un lenguaje, a saber, utilizando una gramática proporciona un conjunto de símbolos de varios tipos y un conjunto de reglas para construir palabras. Concretamente, una gramática consta de un **vocabulario** o **alfabeto** V, que es el conjunto de símbolos usados para obtener los elementos de un lenguaje. Algunos de los elementos del vocabulario no se pueden reemplazar por otros símbolos. Estos elementos se llaman **terminales**, y los restantes elementos del vocabulario, aquellos que pueden sustituirse por otros símbolos, se llaman **no terminales**. El conjunto de símbolos terminales y no terminales se denotan, usualmente, por T y N, respectivamente. En el ejemplo dado en la introducción, el conjunto de terminales es $T = \{un, el, conejo, matemático, salta, come, rápidamente, salvajemente\}$, y el conjunto de no terminales es $N = \{frase, sujeto, predicado, adjetivo, articulo, nombre, verbo, adverbio\}$. En el alfabeto hay un elemento especial llamando **símbolo inicial**, denotado por S, que es un elemento del vocabulario por el que siempre comenzamos. En el ejemplo de la introducción, el símbolo inicial es frase. Se llama **producción** de la gramática a toda regla que especifica cuando se puede reemplazar una cadena de V, el conjunto de todas las cadenas finitas de elementos del vocabulario, por otra cadena. Se denota por $z_0 \rightarrow z_1$ la producción que establece que z_0 puede reemplazarse por z_1 en una cadena. En el ejemplo de la introducción se listaron las producciones de la gramática. La primera producción, escrita utilizando esta notación, es frase → sujeto predicado. Resumimos estas cuestiones en la siguiente definición.

Definición 2.
Una gramática con estructura de frases $G = (V, T, S, P)$ consiste en un vocabulario V, un subconjunto T de V formado por los elementos terminales, un símbolo inicial S de V-T y un conjunto P de producciones. El conjunto V-T se denota por N. Los elementos de N se llaman elementos no terminales. Toda producción de P debe contener al menos un elemento no terminal en su lado izquierdo.

Ejemplo 1:
Sea $G = (V, T, S, P)$, *donde* $V = \{a, b, A, B, S\}$, $T = \{a,b\}$, S es el símbolo inicial y $P = \{S \rightarrow Aba,\ A \rightarrow BB,\ B \rightarrow ab,\ AB \rightarrow b\}$. G es un ejemplo de gramática con estructura de frases.

Definición 3.
Sea $G = (V, T, S, P)$ una gramática con estructura de frases. Sean $w0 = lz0r$ (esto es, la concatenación l, $z0$ y r) y $w1 = lz1r$ cadenas sobre V. Si $z_0 \rightarrow z_1$ es una producción de G, decimos que w_1 se deriva directamente de w_0 (o que es directamente derivable), y escribimos $w_0 \rightarrow w_1$. Si $w_0, w_1, \ldots, w_n$ son cadenas sobre V tales que $w_0 \rightarrow w_1$, $w_1 \rightarrow w_2, \ldots, w_{n-1} \rightarrow w_n$, decimos que w_n es derivable, o se deriva, de w_0, y se denotara $w_0 \rightarrow w_n$. La secuencia de pasos utilizada para obtener w_n a partir de w_0 se llama derivación.

Ejemplo 2:
Sea $G = (V, T, S, P)$, *donde* $V = \{a, b, A, B, S\}$, $T = \{a,b\}$, S es el símbolo inicial y $P = \{S \rightarrow Aba,\ A \rightarrow BB,\ B \rightarrow ab,\ AB \rightarrow b\}$.La cadena Aaba se deriva directamente de ABa, puesto que $B \rightarrow ab$ es una producción de dicha gramática. La cadena abababa se deriva de ABa, puesto que ABa$\rightarrow$ *Aaba*$\rightarrow$ *BBaba*$\rightarrow$ *Bababa*$\rightarrow$ *abababa,* donde se han utilizado las producciones $B \rightarrow ab$, $A \rightarrow BB$, $B \rightarrow ab$ *y* $B \rightarrow ab$ sucesivamente.

Definición 4.
Sea $G= (V, T, S, P)$ una gramática con estructura de frases. El lenguaje generado por G (o el lenguaje de G), denotado por $L (G)$, es el conjunto de todas las cadenas de terminales que se derivan del estado inicial S. En otras palabras, $L(G) = \{w \in T \mid S \rightarrow w\}$

Ejemplo 3:

Sea G la gramática con vocabulario $V = \{S, A, a, b\}$. Conjunto de terminales $T = \{a,b\}$, símbolo inicial S y producciones
P = {S $\rightarrow aA$, $S \rightarrow b$, $A \rightarrow aa\}$. ¿Cuál es $L (G)$, el lenguaje generado por esta gramática?
Solución: A partir del estado inicial S, se puede derivar aA utilizando la producción $S \rightarrow aA$. También se puede usar la producción $S \rightarrow b$ para derivar b. De aA mediante la producción $A \rightarrow aa$ se deriva aaa. Puesto que no puede derivarse ninguna otra palabra utilizando las producciones, se tiene que $L (G) = \{b, aaa\}$.

Ejemplo 4:

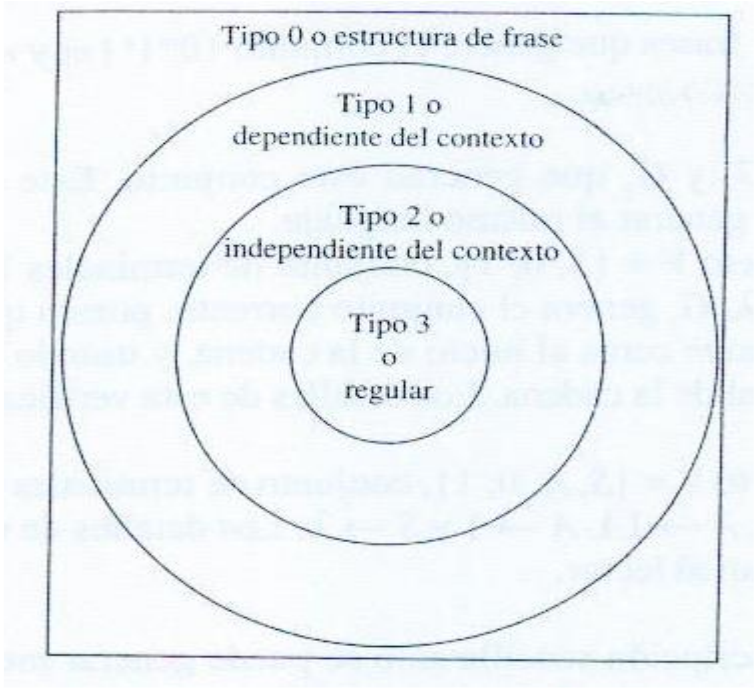
Sea G la gramática con vocabulario $V= \{S, 0, 1\}$, conjunto de terminales $T= \{0, 1\}$, símbolo inicial S y producciones $P= \{S \rightarrow 11S, S \rightarrow 0\}$.
¿Cuál es $L (G)$, el lenguaje generado por esta gramática?
Solución: A partir de S, se puede derivar 0 utilizando $S \rightarrow 0$ o bien $11S$ si empleamos $S \rightarrow 11S$. A partir de $11S$, se puede derivar bien 110 o bien $1111S$. De $1111S$ se puede derivar 11110 y $111111S$. En cualquier paso de una derivación. Se pueden añadir dos unos al final de la cadena o terminar la derivación añadiendo un 0 al final de la cadena. Conjeturamos que $L (G) = \{0, 110, 11110, 1111110, \ldots\}$, el conjunto de todas las cadenas que comienzan con un numero par de unos y terminan con un 0.

Tipos De Gramática Con Estructura De Frases

Las gramáticas con estructura de frases se pueden clasificar de acuerdo con el tipo de producción que utilicen. Una gramática de tipo 0 no impone ninguna restricción a sus producciones. Una gramática de tipo 1 puede tener producciones de la forma $w_1 \rightarrow w_2$, donde la longitud de w_2 es mayor que la de w_3, o de la forma $w_1 \rightarrow \lambda$. Una gramática de tipo 2 solo puede tener producciones de la forma $w_1 \rightarrow w_2$, donde w_1 es un único símbolo no terminal. Una gramática de tipo 3 solo puede tener producciones de la forma $w_1 \rightarrow w_2$, con $w_1=A$, y bien $w_2=aB$ o bien $w_2=a$, siendo A y B símbolos no terminales y a un símbolo terminal, o con $w_1=S$ y $w_2=\lambda$.

De estas definiciones se sigue que toda gramática de tipo 3 es de tipo 2, toda gramática de tipo 2 es de tipo 1 y toda gramática de tipo 1 es de tipo 0. Las gramáticas de tipo 2 también se llaman gramáticas libres de contexto o independientes del contexto, puesto que un símbolo no Terminal que este en el lado izquierdo de una producción puede ser reemplazado en una cadena siempre que aparezca independientemente de lo que figure en la cadena. Un lenguaje generado por una gramática de tipo 2 se llama lenguaje libre de contexto o independiente del contexto. Cuando se tiene una producción de la forma $lw_1r \rightarrow lw_2r$ (pero no de la forma $w_1 \rightarrow w_2$), la gramática se denomina de tipo 1, sensible al contexto o dependiente del contexto, puesto, que w_1 puede ser reemplazado por w_2 solo cuando esta entre las cadenas l y r. Las gramáticas de tipo 3 se llaman también regulares. Un lenguaje generado por una gramática regular se llama regular.

Tipos de gramáticas	
Tipo	Restricciones en las producciones $w_1 \rightarrow w_2$
0	Sin restricciones
1	$l(w_1) < l(w_2)$, o $w_2 = \lambda$
2	w_1 = A, siendo A un símbolo no Terminal
3	w_1= A y w_2 = aB o w_2 = a, siendo A ∈ N, B ∈ N y a ∈ T, o S → λ



Ejemplo 5:

Construye una gramática con estructura de frases que genere el conjunto *{0n 1n| n= 0, 1, 2,…}.*

Solución: Se pueden utilizar dos producciones para generar todas las cadenas que consisten en una cadena de ceros seguida de una cadena con el mismo número de unos, incluyendo la palabra vacía. La primera palabra crece formando cadenas cada vez mas largas del lenguaje, mediante la concatenación de un 0 al principio de la cadena y un 1 al final. La segunda producción reemplaza *S* por la cadena vacía. La solución es la gramática *G = (V,T,S,P),* donde *V = {0, 1, S}, T = {0, 1}, S* es el símbolo inicial y las producciones son $S \rightarrow 0S1$, $S \rightarrow \lambda$.

Es un lenguaje libre o independiente del contexto (tipo 2) puesto que las producciones en esta gramática son $S \rightarrow 0S1$ y $S \rightarrow \lambda$.

Ejemplo 6:

Obtener una gramática con estructura de frases que genere el conjunto $\{0^m 1^n \mid m$ y n son enteros no negativos$\}$.

Solución: La gramática G consta del alfabeto $V= \{S, 0, 1\}$, conjunto de terminales $T= \{0,1\}$ y las producciones $S \rightarrow 0S$, $S \rightarrow S1$ y $S \rightarrow \lambda$. G genera el conjunto correcto, puesto que utilizando la primera producción m veces, se colocan m ceros al inicio de la cadena, y usando la segunda producción n veces, tenemos n unos al final de la cadena. Este es un lenguaje regular o de tipo 3, puesto que puede generarse mediante una gramática regular.