



Notas de Curso

INTRODUCCIÓN A LA TEORÍA DE LOS LENGUAJES FORMALES

SISTEMAS GENERATIVOS

Alberto A. Herrera Becerra

Noviembre 2022

Curso Universitario

ICAT - UNAM, Facultad de Ingeniería - UNAM

Financiamiento interno

INTRODUCCIÓN A LA TEORÍA DE LOS LENGUAJES FORMALES SISTEMAS GENERATIVOS

ALBERTO ARTURO HERRERA BECERRA

Grupo Académico de Modelado y Simulación de Procesos, Departamento
de Instrumentación Científica e Industrial

Resumen

En este documento se presentan las definiciones fundamentales de las gramáticas, su operación como sistemas (o dispositivos) generadores de palabras y su uso para definir y especificar lenguajes formales. Se introduce además el importante concepto de la Jerarquía de Chomsky y, con ello, la definición de las principales clases de familias de lenguajes que se estudian en la Teoría de los Lenguajes Formales. En el documento sólo se trata el tema de las gramáticas de estructura de frase (o gramáticas generativas), ya que es el modelo básico de gramática y el referente para definir otras clases de gramáticas. Las gramáticas de estructura de frase son un tema fundamental para diseñar un curso a nivel universitario acerca de la Teoría de las Computadoras, ya que con ellas se introducen los elementos matemáticos básicos para describir la operación de las computadoras digitales como procesadores lingüísticos. Es un tema que se ha vuelto primordial en la preparación de los modernos ingenieros en computación. Desafortunadamente, el tema es tratado de manera poco sistemática (y en muchos casos de manera muy superficial) en muchos de los más recientes libros introductorios sobre Lenguajes Formales y Autómatas. Este documento se presenta como una ayuda al profesor y al alumno para subsanar tales deficiencias. Los antecedentes matemáticos ya son conocidos para los estudiantes que han tomado un curso de Estructuras Discretas. En adición, el documento es una aportación al conjunto de notas técnicas del grupo de Modelado y Simulación de Procesos del ICAT. El enfoque del documento se distingue porque el tema se aborda desde el punto de vista de los sistemas formales, en especial de los sistemas de reescritura. Este enfoque permite una presentación sistemática y coherente del tema, permitiendo una descripción progresiva de los principales resultados.

Índice

Introducción.....	3
1 CONCEPTOS BÁSICOS.....	4
2 EJEMPLOS DE GRAMÁTICAS.....	9
3 CLASES DE GRAMÁTICAS.....	24
4 LA PALABRA VACÍA.....	29
Referencias Bibliográficas.....	32

Introducción

El concepto de gramática se originó en las investigaciones sobre los lenguajes naturales. La descripción de un lenguaje natural típicamente se denomina “gramática” y con ella se debería indicar cómo las oraciones de un lenguaje están compuestas de elementos lingüísticos más simples; cómo tales elementos forman unidades más grandes; y cómo estas unidades están relacionadas en el contexto de la oración. La teoría de los lenguajes formales se origina de la necesidad de proporcionar una base matemática y formal a tales descripciones. Noam Chomsky, el fundador de la teoría en la década de los 1950, visualizó algo más que un simple refinamiento de la teoría lingüística tradicional. Él estaba preocupado por realizar un examen más profundo de las bases de la teoría lingüística. Él buscó respuestas a interrogantes como “¿cuáles son los objetivos de la teoría lingüística?”, “¿qué condiciones debe satisfacer una gramática para que sea adecuada de acuerdo con los objetivos de la teoría?”, y “¿cuál es la forma general de la teoría lingüística?” Sin una base formal, preguntas como éstas, y otras similares, no podrían ser abordadas con suficiente precisión. Lo que Chomsky (y luego sus colaboradores y seguidores) encontraron es que un lenguaje formal puede servir como un modelo matemático para un lenguaje natural, mientras que una gramática formal puede actuar como un modelo para una teoría lingüística.

Desde un punto de vista matemático, una gramática es un sistema de reescritura de palabras que se usa para generar una secuencia de palabras a partir de una palabra inicial o semilla. Cada palabra de la secuencia (excepto la semilla) se obtiene de su antecesora por medio de reglas generativas explícitas y aplicables de manera mecánica, esto es, sin que se requiera de ingenio por parte de quien las aplica. Dado que las palabras son combinaciones de símbolos de un alfabeto (o vocabulario) previamente establecido, las reglas se basan en cambios de símbolos, los cuales se denominan etapas de reescritura. Así pues, se puede entender a una gramática como una operación sobre palabras, su argumento es la palabra semilla, su valor es la secuencia de palabras generadas. Debido a esto, una gramática puede ser visualizada como un dispositivo (abstracto) generador. Las gramáticas se utilizan principalmente como métodos rigurosos para definir, o especificar, lenguajes; de hecho, como ya se mencionó, las gramáticas fueron usadas inicialmente como modelos matemáticos de las características más

relevantes de los lenguajes naturales. Adicionalmente, las gramáticas se pueden entender como modelos matemáticos de procesos discretos, en especial de procesos de datos, considerando a las palabras como esquemas de datos.

En este documento se presentan las definiciones fundamentales de las gramáticas, su operación como dispositivos generadores de palabras y su uso para definir lenguajes formales. Se introduce además el importante concepto de la Jerarquía de Chomsky y, con ello, la definición de las principales clases de familias de lenguajes que se estudian en la Teoría de los Lenguajes Formales.

Antes de empezar con las definiciones es pertinente remarcar lo siguiente. Una gramática bien construida necesariamente especifica a un lenguaje; lo especifica palabra por palabra. Sin embargo, un cierto lenguaje puede ser especificado por gramáticas diferentes. La asociación entre gramáticas y lenguajes no es de tipo biyectivo. Las gramáticas pueden ser agrupadas en tipos de gramáticas. Todas las gramáticas de un cierto tipo especifican lenguajes de ese mismo tipo. El gran descubrimiento de Noam Chomsky es que ciertos tipos de gramáticas se agrupan jerárquicamente, existe un tipo supremo de gramática que contiene a los otros tipos como casos particulares, y existe un tipo ínfimo de gramática. Cada tipo de gramática superior contiene a los tipos inferiores. La gramática ínfima ya no contiene a otros tipos de gramáticas, pero es contenida por todos los otros tipos. Así pues, estos tipos de gramáticas organizadas jerárquicamente definen familias de lenguajes que también se organizan jerárquicamente.

1 CONCEPTOS BÁSICOS

Una *gramática* $G = (V_N, V_T, F, X_0)$ es un sistema de reescritura de palabras en el que V_N y V_T son alfabetos inconexos, $V_N \cap V_T = \emptyset$, en el que F es un conjunto finito de pares ordenados (α, β) de palabras sobre $V = V_N \cup V_T$, y en el que X_0 es una palabra unitaria formada por el símbolo no terminal X_0 . El conjunto $V = V_N \cup V_T$ se denomina el *vocabulario* de la gramática, siendo V_N el alfabeto de *símbolos no terminales* (o *símbolos variables*) y V_T el alfabeto de *símbolos terminales*; en particular, $X_0 \in V_T$ se denomina el *símbolo inicial* de la gramática. Los pares ordenados (α, β) de F se denominan *reglas de reescritura*, o *reglas de producción*, o simplemente *producciones*, y se representan $\alpha \rightarrow \beta$; se debe

cumplir que α sea una palabra no vacía y contener por lo menos un símbolo no terminal. Es común decir que α es la *cabeza* de la producción, mientras que β es su *cuerpo*.

Las gramáticas así definidas suelen denominarse *gramáticas generativas* o, con base en las definiciones de Chomsky, *gramáticas de estructura de frase*. La primera denominación es un tanto reiterativa, como ahora se muestra.

Una gramática G es un sistema generador $G = (RW, AX)$, formado por el sistema de reescritura $RW = (V, F)$ y por el conjunto (o esquema) de axiomas AX ; como ya se mencionó, $V = V_N \cup V_T$ es el vocabulario, F es el conjunto finito de reglas de producción, y $AX = \{X_0\}$. Conviene remarcar aquí que el único axioma de la gramática es una palabra unitaria X_0 , la *palabra inicial*. En la práctica suele decirse indistintamente que X_0 es la palabra o el símbolo inicial, considerando que el contexto de la discusión permitirá establecer el uso correcto.

Dado que en toda gramática G subyace un sistema de reescritura RW , toda gramática produce una secuencia (posiblemente ilimitada) de palabras; esto ocurre como se describe a continuación. Si $\alpha \rightarrow \beta$ es una producción de F y si γ, δ son palabras de V^* , entonces se establece la relación $\gamma\alpha\delta \Rightarrow_G \gamma\beta\delta$ y se dice que $\gamma\alpha\delta$ *deriva directamente* en (o se reescribe como) $\gamma\beta\delta$ en G . Si se sobreentiende el uso de la gramática G , entonces se escribe simplemente $\gamma\alpha\delta \Rightarrow \gamma\beta\delta$. También se dice que la palabra α se ha sustituido por la palabra β en el contexto formado por las palabras γ y δ . Es conveniente hacer notar que si $\alpha \rightarrow \beta \in F$, entonces necesariamente $\alpha \Rightarrow_G \beta$. Así pues, $\Rightarrow_G$ relaciona dos palabras exactamente cuando la segunda se obtiene de la primera por reescritura, esto es, por medio de la aplicación de una producción.

De lo anterior, es importante hacer notar que están involucradas dos relaciones binarias sobre V^* : la primera es la relación finita $\rightarrow$ de reglas de producción, la segunda es $\Rightarrow_G$ de derivación directa, donde $\Rightarrow_G$ se define en términos de $\rightarrow$. En la notación de relaciones se tiene que $(\gamma\alpha\delta, \gamma\beta\delta) \in \Rightarrow_G$ si y sólo si $(\alpha, \beta) \in \rightarrow$.

Es la relación binaria $\Rightarrow_G$ la que va a ser usada para especificar a un lenguaje formal. Para lograr esto es necesario que la relación, ante todo, sea transitiva. De inicio, la relación $\Rightarrow_G$ muy seguramente no es transitiva, pero la propiedad se puede garantizar procediendo como se indica a continuación. Supóngase que $\alpha_0, \alpha_1, \dots, \alpha_m$ son palabras de V^* tales que $\alpha_0 \Rightarrow_G \alpha_1, \alpha_1 \Rightarrow_G \alpha_2, \dots, \alpha_{m-1} \Rightarrow_G \alpha_m$, entonces se dice que α_0 *deriva* en

α_m en la gramática G . La secuencia de derivaciones directas anterior se puede denotar como $\alpha_0 \Rightarrow_G^m \alpha_m$, donde m es la longitud de la derivación. La relación $\alpha_0 \Rightarrow_G^m \alpha_m$ se denomina *derivación* de m etapas de reescritura o, simplemente, derivación. Efectivamente, $\Rightarrow_G^m$ es la potencia m de $\Rightarrow_G$; la transitividad se establece aplicando $\Rightarrow_G$ repetidas veces como se muestra a continuación:

$$\begin{aligned} \alpha_0 &\Rightarrow \alpha_1; \\ \alpha_1 &\Rightarrow \alpha_2; \alpha_0 \Rightarrow^2 \alpha_2; \\ \alpha_2 &\Rightarrow \alpha_3; \alpha_1 \Rightarrow^2 \alpha_3; \alpha_0 \Rightarrow^3 \alpha_3; \\ \alpha_3 &\Rightarrow \alpha_4; \alpha_2 \Rightarrow^2 \alpha_4; \alpha_1 \Rightarrow^3 \alpha_4; \alpha_0 \Rightarrow^4 \alpha_4; \\ &\text{Etcétera.} \end{aligned}$$

Con base en esto, se deduce lo siguiente:

$$\begin{aligned} (\alpha_0, \alpha_1), (\alpha_0, \alpha_2), (\alpha_0, \alpha_3), \dots, (\alpha_0, \alpha_m) &\in \Rightarrow^m. \\ (\alpha_1, \alpha_2), (\alpha_1, \alpha_3), (\alpha_1, \alpha_4), \dots, (\alpha_1, \alpha_m) &\in \Rightarrow^m. \\ &\text{Etcétera.} \end{aligned}$$

De la relación $\Rightarrow_G$ se puede construir su cerradura transitiva $\Rightarrow_G^+$ de la manera siguiente. Supóngase que se puede establecer la derivación $\alpha \Rightarrow_G^k \beta$ para algún $k \geq 1$, entonces se escribe $\alpha \Rightarrow_G^+ \beta$ y se dice que α *deriva en* β *de manera no trivial*. En esto, el valor exacto de k no es importante, sólo importa que exista tal número. Es conveniente recordar que $\Rightarrow_G^+$ es la relación más pequeña que contiene a $\Rightarrow_G$ y que es transitiva, esto es, si $(\alpha, \beta) \in \Rightarrow_G^+$, entonces α y β necesariamente están relacionadas transitivamente. Además, va a ser importante considerar que la relación de derivación también sea reflexiva en el sentido siguiente. Se dice que $\alpha \Rightarrow_G^0 \beta$ si y sólo si $\alpha = \beta$; literalmente, toda palabra se relaciona consigo misma por una derivación de cero transformaciones, si la palabra no sufre cambios. De esta manera, se escribe $\alpha \Rightarrow_G^* \beta$, y se dice que α *deriva en* β , si existe una derivación $\alpha \Rightarrow_G^k \beta$, para algún $k \geq 0$. La relación $\Rightarrow_G^*$ es la cerradura transitiva reflexiva de $\Rightarrow$.

Resumiendo lo anterior. Una gramática $G = (V_N, V_T, F, X_0)$ establece una relación binaria F , o $\rightarrow$, sobre el conjunto V^* , donde $V = V_N \cup V_T$. El sistema de reescritura que subyace a G permite definir la relación binaria $\Rightarrow$, también sobre V^* , a partir de la relación $\rightarrow$. Las cerraduras transitiva y transitiva reflexiva, $\Rightarrow_G^+$ y $\Rightarrow_G^*$ respectivamente, permiten establecer un conjunto de palabras sobre V^* , las palabras que están relacionadas por

reescritura según las reglas de producción de la gramática, o, dicho de otra manera, las palabras que se derivan unas de las otras. A partir de estas palabras relacionadas por reescritura, o derivación, se va a definir un lenguaje particular. Esto se hace como sigue.

Se dice que una gramática $G = (V_N, V_T, F, X_0)$ especifica (o define) a un lenguaje $L(G)$ sobre V_T , denominado el *lenguaje generado* por la gramática G , como el conjunto siguiente:

$$L(G) = \{x \in V_T^* \mid X_0 \Rightarrow^* x\}.$$

Esto es, el lenguaje generado está formado por todas las palabras de símbolos terminales (o palabras terminales) que se derivan a partir de la palabra inicial X_0 de la gramática. Dicho de otra manera, el lenguaje generado está formado por todas las palabras terminales generadas a partir del único axioma de la gramática. Este hecho impone restricciones muy importantes al conjunto F como se verá más adelante. En toda derivación de la forma $X_0 \Rightarrow^* \alpha$, con $\alpha \in V^*$, la palabra α se denominan *forma oracional* de la gramática; si sucede que $\alpha \in V_T^*$, entonces se denomina *oración* de la gramática. Esta forma de denominación es una reminiscencia del origen lingüístico de las gramáticas.

De lo anterior, debe resultar claro que una gramática generativa G es un método para definir o especificar a un lenguaje $L(G)$. Adicionalmente, la misma gramática G también permite resolver el problema de especificación: una palabra dada w de símbolos terminales pertenece a $L(G)$ si y sólo si existe una derivación $X_0 \Rightarrow^* w$ en G . Y dado que las acciones que se realizan sobre las formas oracionales de una derivación sólo involucran a sus símbolos y sus combinaciones, la gramática define a su lenguaje generado por su sintaxis. De hecho, suele decirse que la gramática es la sintaxis de su lenguaje generado.

Para completar las definiciones fundamentales, se introduce ahora un concepto dual al de gramática generativa que es importante en la construcción de la teoría de las gramáticas.

Una *gramática analítica* es un sistema reconocedor $G = (V_N, V_T, F, X_0)$, donde V_N , V_T y X_0 se definen exactamente igual que como se hizo en una gramática generativa, y aunque el conjunto F sigue siendo un conjunto finito de pares ordenados (α, β) de palabras sobre $V = V_N \cup V_T$, ahora β debe ser una palabra no vacía que contiene por lo menos un símbolo no terminal.

Similarmente a lo que sucede en las gramáticas generativas, en las gramáticas analíticas también subyace un sistema de reescritura; este sistema permite establecer las relaciones de derivación directa y de derivación en la gramática y con ellas se puede definir ahora un *lenguaje reconocido*, o *aceptado*, de la manera siguiente:

$$L(G) = \{x \in V_T^* \mid x \Rightarrow^* X_0\}.$$

Esto es, el lenguaje reconocido por una gramática analítica es el conjunto de todas las palabras terminales que derivan en la palabra inicial de la gramática; por así decirlo, son las palabras terminales que se pueden reducir a la palabra inicial, o axioma, de la gramática mediante reescritura. Así pues, una gramática generativa, como su nombre lo sugiere, establece un proceso generador de palabras, mientras que una gramática analítica establece un proceso de reconocimiento de palabras.

Dos gramáticas G y G' , independientemente de si son generativas o analíticas, se dice que son (débilmente) *equivalentes* si sucede que $L(G) = L(G')$. Este concepto se originó en la teoría lingüística, junto con el concepto relacionado de equivalencia fuerte, y ambos están asociados con la naturaleza y uso de los lenguajes naturales. En este documento se va a usar el primer concepto simplemente como se ha enunciado.

Una característica especial que surge de las definiciones de las dos clases de gramáticas es que las gramáticas aparecen en pares de gramáticas equivalentes. Este hecho se establece en el resultado siguiente.

Teorema 1.1

Sea $G = (V_N, V_T, F, X_0)$ una gramática, generativa o analítica. Sea ahora el conjunto F_1 de todas las producciones $\beta \rightarrow \alpha$ tales que $\alpha \rightarrow \beta$ es una producción de F . Entonces la gramática $G' = (V_N, V_T, F_1, X_0)$ es equivalente a G , esto es, $L(G) = L(G')$.

Prueba

Claramente, si G es una gramática generativa (analítica), entonces G' es una gramática analítica (generativa). Entonces, y sin pérdida de generalidad, se asume que G es una gramática generativa. Se propone que para toda palabra ω sobre $V = V_N \cup V_T$ se tiene que $X_0 \Rightarrow_G^* \omega$ si y sólo si $\omega \Rightarrow_{G'}^* X_0$. De hecho, se puede mostrar que, para todo número k , la secuencia finita:

$$\omega_0, \omega_1, \dots, \omega_k,$$

es una derivación respecto a G si y sólo si la secuencia finita:

$$\omega_k, \omega_{k-1}, \dots, \omega_0,$$

es una derivación respecto a G' . Debido a la definición del conjunto F_1 , claramente la aseveración anterior es verdadera para $k = 1$, esto es, si ω_0, ω_1 es una derivación respecto a G , entonces ω_1, ω_0 es una derivación respecto a G' . A partir de esto, se asume que la aseveración es verdadera para algún $k \geq 1$, esto es, $\omega_0, \dots, \omega_k$ se satisface respecto a G si y sólo si $\omega_k, \dots, \omega_0$ se satisface respecto a G' . Ahora bien, ω_k, ω_{k+1} se satisface en G si y sólo si ω_{k+1}, ω_k se satisface en G' . Combinando los dos últimos resultados se concluye que la aseveración se satisface para $k + 1$. Para finalizar se concluye que $X_0 \Rightarrow_G^* \omega$ si y sólo si $\omega \Rightarrow_{G'}^* X_0$ con lo que $L(G) = L(G')$, como se quería probar.

De lo anterior se establece que para una gramática G , generativa o analítica, la gramática G' se denomina la *dual*. Es conveniente remarcar que G y G' son equivalentes, pero no son iguales: si una genera (reconoce) a su lenguaje, entonces su dual reconoce (genera) al mismo lenguaje.

En lo que sigue del documento, la discusión sólo se referirá a las gramáticas generativas, por lo que el uso del término gramática hará referencia a una gramática de esa clase.

2 EJEMPLOS DE GRAMÁTICAS

En esta sección se presentan unos cuantos ejemplos de gramáticas. Los ejemplos tienen el interés de ilustrar que no siempre es fácil obtener el lenguaje generado por una gramática dada. Los ejemplos se basan en el uso de gramáticas en las que se pueden establecer derivaciones que permiten, de manera más o menos clara, inferir el lenguaje generado.

Ejemplo 1 Como primer ejemplo se va a mostrar que la gramática siguiente:

$$G_1 = (\{X_0, X_1, X_2\}, \{a, b, c\}, F_1, \{X_0\}),$$

donde el conjunto F_1 contiene las producciones siguientes:

- (1) $X_0 \rightarrow abc$, (2) $X_0 \rightarrow aX_1bc$,
- (3) $X_1b \rightarrow bX_1$, (4) $X_1c \rightarrow X_2bcc$,
- (5) $bX_2 \rightarrow X_2b$, (6) $aX_2 \rightarrow aaX_1$, (7) $aX_2 \rightarrow aa$.

genera el lenguaje $L(G_1) = \{a^i b^i c^i \mid i \geq 1\}$.

Antes de comenzar con el análisis de la gramática es útil introducir una convención de notación que permite una presentación compacta de las gramáticas. Para esto, se establece que los símbolos no terminales de V_N se representan con letras mayúsculas del español (principalmente del final del alfabeto), con o sin subíndices o superíndices; los símbolos terminales de V_T se representan con letras minúsculas (principalmente del inicio del alfabeto), también con o sin subíndices o superíndices, o con dígitos. Con esto, sólo es necesario listar las producciones de la gramática considerando que las primeras producciones que se listan corresponden a las reglas que involucran a la palabra inicial en su cabeza, y como este axioma es una palabra unitaria, sólo es necesario reconocer un símbolo no terminal. Como complemento de esta notación, las palabras terminales serán denotadas por letras minúsculas del final del español, mientras que las palabras de no terminales y terminales serán denotadas por letras griegas minúsculas.

Siguiendo la convención anterior, la gramática G_1 también se puede presentar como sigue:

- G_1 :
- (1) $X_0 \rightarrow abc$,
 - (2) $X_0 \rightarrow aX_1bc$,
 - (3) $X_1b \rightarrow bX_1$,
 - (4) $X_1c \rightarrow X_2bcc$,
 - (5) $bX_2 \rightarrow X_2b$,
 - (6) $aX_2 \rightarrow aaX_1$,
 - (7) $aX_2 \rightarrow aa$.

Analizando las producciones de G_1 se puede deducir lo siguiente. Toda derivación útil debe comenzar con la palabra inicial X_0 y, en esta gramática, tal palabra se puede reescribir de dos maneras diferentes:

- a) $X_0 \Rightarrow abc$, aplicando la regla (1) $X_0 \rightarrow abc$; o
- b) $X_0 \Rightarrow aX_1bc$, aplicando la regla (2) $X_0 \rightarrow aX_1bc$.

La posibilidad de reescribir una palabra de varias formas diferentes es una característica común que se observa en muchas gramáticas y es denominada *no determinismo*. Esto indica que el proceso de derivación en una gramática puede tener resultados diferentes a partir de una misma condición inicial. Los resultados alternativos se obtienen de la realización de una elección arbitraria en, por lo menos, una de las etapas de reescritura. Dicha elección implica la aplicación de una regla de producción, y sólo una regla, de varias alternativas posibles. La elección es arbitraria, no probabilística; una derivación no

es un proceso aleatorio gobernado por alguna función de distribución de probabilidad. Obviamente, la elección es realizada por la persona (o agente computacional) que está desarrollando la derivación. Así, los procesos generativos que ocurren en una gramática son, ante todo, procesos no deterministas: una forma oracional dada puede generar cero, una o más de una forma oracional nueva. Así, el conocimiento de las reglas de producción y de la condición inicial del proceso no necesariamente garantiza la predicción del resultado final.

Si se ha desarrollado el proceso (a) anterior, de inmediato se deduce que el mismo ya no puede continuar pues la palabra derivada, abc , es una palabra formada únicamente por símbolos terminales y, en consecuencia, ya no se puede reescribir. También se deduce que la palabra derivada pertenece al lenguaje generado por la gramática, es decir, resulta que $abc = a^1b^1c^1 \in L(G_1)$.

Por otro lado, si se ha desarrollado el proceso (b) anterior, se obtiene la forma oracional aX_1bc , la cual puede ser reescrita nuevamente ya que contiene al símbolo no terminal X_1 . La gramática G_1 contiene dos reglas que involucran a X_1 , las reglas (3) y (4), pero su posible aplicación depende del “contexto” en el que se encuentre X_1 . La regla (3) sólo puede ser aplicada si X_1 tiene como sucesor una b ; por su parte, la regla (4) sólo puede ser aplicada si X_1 tiene como sucesor una c . Entonces, debería ser claro que la forma oracional aX_1bc sólo puede ser reescrita aplicando la regla (3):

$$X_0 \xRightarrow{(2)} aX_1bc \xRightarrow{(3)} abX_1c.$$

(La notación $\xRightarrow{(i)}$ se usa para indicar que el paso de reescritura señalado se estableció por la aplicación de la regla (i).) La nueva forma oracional derivada abX_1c vuelve a contener al no terminal X_1 , por lo que puede ser reescrita nuevamente. Sin embargo, en este caso es claro que sólo se puede aplicar la regla (4), el no terminal X_1 tiene a su derecha el terminal c , obteniéndose lo siguiente:

$$X_0 \xRightarrow{(2)} aX_1bc \xRightarrow{(3)} abX_1c \xRightarrow{(4)} abX_2bcc.$$

Nótese que la nueva forma oracional abX_2bcc contiene al no terminal X_2 , el cual puede ser rescrito de tres formas diferentes. Una de ellas, dada por la regla (5), sólo se puede aplicar si a la izquierda de X_2 se encuentra una b . Los otros dos cambios, dados por las reglas (6) y (7), sólo se pueden aplicar si a la izquierda de X_2 se encuentra una a ; aquí

aparece de nueva cuenta la característica no determinista de la gramática. En particular, a la palabra abX_2bcc sólo se puede aplicar la regla (5), obteniéndose lo siguiente:

$$X_0 \xRightarrow{(2)} aX_1bc \xRightarrow{(3)} abX_1c \xRightarrow{(4)} abX_2bcc \xRightarrow{(5)} aX_2bbcc.$$

Ahora la forma oracional aX_2bbcc puede ser reescrita de dos formas diferentes, una aplicando la regla (6), la otra aplicando la regla (7). En este último caso se obtiene lo siguiente:

$$X_0 \xRightarrow{(2)} aX_1bc \xRightarrow{(3)} abX_1c \xRightarrow{(4)} abX_2bcc \xRightarrow{(5)} aX_2bbcc \xRightarrow{(7)} aabbcc = a^2b^2c^2.$$

En este caso la derivación termina pues se ha obtenido una palabra terminal. Hasta el momento se han encontrado, y especificado, dos palabras del lenguaje generado $L(G_1)$.

Por otro lado, si a la forma oracional aX_2bbcc se le aplica la regla (6), se obtiene la derivación siguiente:

$$X_0 \xRightarrow{(2)} aX_1bc \xRightarrow{(3)} abX_1c \xRightarrow{(4)} abX_2bcc \xRightarrow{(5)} aX_2bbcc \xRightarrow{(6)} aaX_1bbcc.$$

Nótese que en la última forma oracional aaX_1bbcc ha reaparecido el no terminal X_1 y, por lo tanto, parte del proceso ya descrito se puede repetir. A continuación, se describe la manera en la que se puede repetir el proceso y el efecto de tal repetición.

Regresando al proceso generativo inicial, se observa que el no terminal X_1 aparece por primera vez al aplicar la regla (2). Después sólo se puede aplicar la regla (3), con lo cual permanece el no terminal X_1 . Después sólo se puede aplicar la regla (4), con lo cual desaparece X_1 y aparece X_2 . Después sólo se puede aplicar la regla (5), permaneciendo X_2 . Para finalizar esta parte del proceso se aplica la regla (6), reapareciendo X_1 . Nótese que, al inicio del proceso, X_1 tiene una a a su izquierda y la secuencia bc a su derecha. Al final del proceso, X_1 tiene la secuencia $aa = a^2$ a su izquierda y la secuencia $bbcc = b^2c^2$ a su derecha. Esto es, se estableció la derivación siguiente:

$$a^1X_1b^1c^1 \Rightarrow^* a^2X_1b^2c^2.$$

La forma oracional $a^2X_1b^2c^2$ ahora puede ser reescrita de la manera siguiente:

$$\begin{aligned} a^2X_1b^2c^2 &\xRightarrow{(3)} a^2b^1X_1b^1c^2 \xRightarrow{(3)} a^2b^2X_1c^2 \\ &\xRightarrow{(4)} a^2b^2X_2bc^3 \\ &\xRightarrow{(5)} a^2b^1X_2b^2c^3 \xRightarrow{(5)} a^2X_2b^3c^3. \end{aligned}$$

A partir de la forma oracional $a^2X_2b^3c^3$ se tienen dos opciones. Una consiste en aplicar la regla (7) para obtener la derivación:

$$X_0 \Rightarrow^* a^2X_2b^3c^3 \xRightarrow{(7)} a^3b^3c^3.$$

Se obtiene una nueva palabra terminal. La segunda opción, aplicar la regla (6), es la que interesa, pues con ella se obtiene la derivación siguiente:

$$\begin{aligned} a^2X_1b^2c^2 &\xRightarrow{(3)} a^2b^1X_1b^1c^2 \xRightarrow{(3)} a^2b^2X_1c^2 \\ &\xRightarrow{(4)} a^2b^2X_2bc^3 \\ &\xRightarrow{(5)} a^2b^1X_2b^2c^3 \xRightarrow{(5)} a^2X_2b^3c^3. \\ &\xRightarrow{(6)} a^3X_1b^3c^3. \end{aligned}$$

Se regenera el no terminal X_1 y se puede empezar a deducir el patrón que siguen las derivaciones. Se parte de una palabra de la forma $a^nX_1b^nc^n$, con $n \geq 1$, y se obtiene la palabra de la forma $a^{n+1}X_1b^{n+1}c^{n+1}$, aplicando primero la regla (3) n veces, luego la regla (4) una sola vez, después la regla (5) nuevamente n veces, y se termina con la aplicación de la regla (6) una sola vez. Si al terminar la aplicación de las n repeticiones de la regla (5) se aplica la regla (7), entonces se obtiene la palabra terminal $a^{n+1}b^{n+1}c^{n+1}$.

Con base en la evidencia anterior, se puede establecer un razonamiento de tipo inductivo para demostrar que el lenguaje generado por la gramática G_1 efectivamente es:

$$L(G_1) = \{a^ib^ic^i \mid i \geq 1\}.$$

Se considera cualquier derivación D que comience con la palabra $a^iX_1b^ic^i$, con $i \geq 1$, y que termine con una palabra sobre el alfabeto terminal. Dichas derivaciones D deben comenzar con i aplicaciones de la producción (3) y terminar con una aplicación de la producción (4), llegando a la forma oracional $a^iX_2b^{i+1}c^{i+1}$. A partir de esta palabra se genera la palabra $a^{i+1}X_1b^{i+1}c^{i+1}$ o la palabra $a^{i+1}b^{i+1}c^{i+1}$. La segunda es la palabra buscada. La primera es del mismo tipo que con la que comienza D salvo que el exponente está incrementado en una unidad. Considerando el hecho de que toda derivación especificada por G_1 comienza con una aplicación de la primera o segunda producciones, se puede establecer la validez de nuestra aseveración inicial.

Nótese que el problema de especificación ha quedado resuelto. En efecto, sólo las palabras de la forma $a^ib^ic^i$, para $i \geq 1$, pueden ser generadas por la gramática y, en consecuencia, son las únicas palabras que pertenecen al lenguaje generado. Ninguna

palabra que no sea de la forma dicha puede ser generada por la gramática y, por lo tanto, esas palabras no pueden pertenecer al lenguaje generado.

Nótese además que todas las reglas de producción de la gramática involucran, a lo más, un solo símbolo no terminal, tanto en su cabeza como en su cuerpo. Nótese también que solo las reglas (1) y (7) eliminan a un símbolo no terminal; X_0 en el caso de la regla (1) y X_2 en el caso de la regla (7). Así, el no terminal X_0 puede desaparecer completamente, regla (1), o desaparecer para que aparezca X_1 , regla (2); el no terminal X_1 puede regenerarse, regla (3), o desaparecer para que aparezca X_2 , regla (4); y el no terminal X_2 puede regenerarse, regla (5), desaparecer para que vuelva a aparecer X_1 , regla (6), o desaparecer completamente, regla (7). Adicionalmente, las reglas de producción (3) a (7) pueden ser interpretadas de la manera siguiente. Todas ellas indican la posible reescritura de un símbolo no terminal, digamos X_i , sólo si se encuentra en un “contexto” apropiado. Esto es, el cambio sólo se puede realizar si X_i tiene a sus lados los símbolos apropiados. Esta situación fue denominada sensibilidad al contexto por Noam Chomsky.

Ejemplo 2 Como segundo ejemplo se considera a la gramática G_2 dada por:

- $$\begin{aligned} G_2: \quad (1) \quad & S \rightarrow aSBC, \\ (2) \quad & S \rightarrow aBC, \\ (3) \quad & CB \rightarrow BC, \\ (4) \quad & aB \rightarrow ab, \\ (5) \quad & bB \rightarrow bb, \\ (6) \quad & bC \rightarrow bc, \\ (7) \quad & cC \rightarrow cc. \end{aligned}$$

Un aspecto llamativo de la gramática G_2 es su regla (3), que involucra a dos no terminales en su cabeza.

La palabra de inicio, S en este caso, puede ser reescrita de dos maneras diferentes. Se considera primero la reescritura debido a la aplicación de la regla (2), obteniéndose:

$$S \xRightarrow{(2)} aBC.$$

Claramente el no terminal S ha desaparecido y aparecieron los no terminales B y C en este orden precisamente. La forma oracional aBC sólo puede ser reescrita de la manera siguiente:

$$S \xRightarrow{(2)} aBC \xRightarrow{(4)} abC.$$

En este caso, sólo el no terminal B fue reescrito, desapareciendo de hecho. Finalmente, la forma oracional abC sólo puede ser rescrita de la manera siguiente:

$$S \xRightarrow{(2)} aBC \xRightarrow{(4)} abC \xRightarrow{(6)} abc.$$

En este último paso desapareció el no terminal C , con lo que se obtuvo la palabra terminal abc . La desaparición de los no terminales B y C fue posible debido a que se encontraban en el “contexto” adecuado.

Así pues, toda derivación que comienza con la aplicación de la regla (2), debe proseguir necesariamente con la aplicación de las reglas (4) y (6), en este orden, para concluir en la obtención de la palabra terminal abc .

La segunda alternativa de reescritura de la palabra inicial S es mediante la aplicación de la regla (1), obteniéndose lo siguiente:

$$S \xRightarrow{(1)} aSBC.$$

Se ha regenerado el no terminal S , por lo que la forma oracional $aSBC$ puede ser reescrita de dos maneras diferentes. Si se aplica la regla (2) se obtiene la derivación siguiente:

$$S \xRightarrow{(1)} aSBC \xRightarrow{(2)} aaBCBC.$$

Finalmente desapareció el no terminal S , apareciendo en su lugar cuatro no terminales en la secuencia $BCBC$. Ciertamente, en la forma oracional $aaBCBC$ hay cuatro no terminales, pero sólo hay dos posibles derivaciones, una debida a la aplicación de la regla (3), la otra debida a la aplicación de la regla (4). Se analizan pues ambas posibilidades.

La primera derivación posible es:

$$a) \quad S \xRightarrow{(1)} aSBC \xRightarrow{(2)} aaBCBC \xRightarrow{(3)} aaBBCC.$$

Nótese que en este caso no se alteró el número de no terminales, pero sí se alteró el orden en el que aparecen. No es difícil determinar que la forma oracional $aaBBCC$ sólo puede ser reescrita de la manera siguiente:

$$aaBBCC \xRightarrow{(4)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbC \xRightarrow{(7)} aabbcc = a^2b^2c^2.$$

En cada paso de la derivación se elimina un no terminal, comenzando con las B es y terminando con las C es.

La segunda derivación posible es:

$$b) \quad S \xRightarrow{(1)} aSBC \xRightarrow{(2)} aaBCBC \xRightarrow{(4)} aabCBC.$$

En este caso desapareció una de las *B*es, la primera de ellas. La forma oracional *aabCBC* puede ser reescrita de dos formas diferentes; la primera de ellas es la siguiente:

$$aabCBC \xRightarrow{(3)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbcC \xRightarrow{(7)} aabbcc.$$

En el primer paso no se alteró el número de no terminales, sólo se reordenaron. En los pasos siguientes se fueron eliminando progresivamente los no terminales hasta obtener una palabra terminal. La segunda derivación posible es la siguiente:

$$aabCBC \xRightarrow{(6)} aabcBC.$$

En este caso, después de eliminar el primer no terminal, se obtiene una forma oracional que ya no puede ser reescrita. Sin embargo, esa forma oracional no es terminal.

Así pues, a partir de la forma oracional *aaBCBC* se pueden obtener las derivaciones siguientes:

$$(A) \quad aaBCBC \xRightarrow{(3)} aaBBCC \xRightarrow{(4)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbcC \xRightarrow{(7)} aabbcc = a^2b^2c^2.$$

$$(B) \quad aaBCBC \xRightarrow{(4)} aabCBC \xRightarrow{(3)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbcC \xRightarrow{(7)} aabbcc = a^2b^2c^2.$$

$$(C) \quad aaBCBC \xRightarrow{(4)} aabCBC \xRightarrow{(6)} aabcBC.$$

Nótese que las dos primeras derivaciones terminan en la misma palabra terminal; esta situación se conoce como *ambigüedad*, y se dice que la gramática G_2 es ambigua. En las derivaciones (A) y (B) se usan exactamente las mismas reglas de producción, pero no en el mismo orden. La tercera derivación no termina en una palabra terminal, y aunque forma parte de la operación de la gramática, no contribuye a la formación del lenguaje generado. No todas las derivaciones de una gramática son “exitosas” en el sentido de que terminan en una palabra terminal.

Las derivaciones (A), (B) y (C) anteriores se pueden describir de manera compacta como sigue:

$$aaBCBC \begin{cases} \xRightarrow{(3)} aaBBCC \xRightarrow{(4)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbcC \xRightarrow{(7)} aabbcc \\ \xRightarrow{(4)} aabCBC \xRightarrow{(3)} aabBCC \xRightarrow{(5)} aabbCC \xRightarrow{(6)} aabbcC \xRightarrow{(7)} aabbcc \\ \xRightarrow{(4)} aabCBC \xRightarrow{(6)} aabcBC \end{cases}$$

Con esto se resalta el carácter no determinista de la gramática.

Con los resultados obtenidos hasta este momento, se puede deducir la característica siguiente de las derivaciones exitosas en G_2 :

$$S \xRightarrow{(2)} aBC \Rightarrow^* abc = a^1 b^1 c^1.$$

$$S \xRightarrow{(1)} aSBC \xRightarrow{(2)} aaBCBC \Rightarrow^* aabbcc = a^2 b^2 c^2.$$

$$S \xRightarrow{(1)} aSBC \xRightarrow{(1)} aaSBCBC \xRightarrow{(2)} aaaBCBCBC \Rightarrow^* aaabbbccc = a^3 b^3 c^3.$$

Y así por el estilo.

Lo anterior sugiere que la parte inicial de la derivación se puede generalizar de la manera siguiente:

$$S \xRightarrow{(1)}^n a^n S(BC)^n \xRightarrow{(2)} a^{n+1} (BC)^{n+1}, \text{ con } n \geq 0.$$

La forma oracional $a^{n+1} (BC)^{n+1}$ puede ser reescrita de dos maneras diferentes. La primera de ellas procede como sigue:

$$a^{n+1} (BC)^{n+1} \xRightarrow{(3)}^m a^{n+1} B^{n+1} C^{n+1}, \text{ donde } m = \begin{cases} n, & n = 0, 1, \\ 2n - 1, & n \text{ es par,} \\ 2n, & n \text{ es impar.} \end{cases}$$

$$\begin{aligned} a^{n+1} B^{n+1} C^{n+1} &\xRightarrow{(4)} a^{n+1} b B^n C^{n+1} \\ &\xRightarrow{(5)}^n a^{n+1} b^{n+1} C^{n+1} \\ &\xRightarrow{(6)} a^{n+1} b^{n+1} c C^n \\ &\xRightarrow{(7)}^n a^{n+1} b^{n+1} c^{n+1}. \end{aligned}$$

La segunda forma de reescritura procede como sigue:

$$a^{n+1} (BC)^{n+1} \xRightarrow{(4)} a^{n+1} b (CB)^n C.$$

$$a^{n+1} b (CB)^n C \xRightarrow{(3)}^m a^{n+1} b B^n C^{n+1}, \text{ donde } m = \begin{cases} n, & n = 0, 1, \\ 2n - 1, & n \text{ es par,} \\ 2n, & n \text{ es impar.} \end{cases}$$

$$\begin{aligned} a^{n+1} b B^n C^{n+1} &\xRightarrow{(5)}^n a^{n+1} b^{n+1} C^{n+1} \\ &\xRightarrow{(6)} a^{n+1} b^{n+1} c C^n \\ &\xRightarrow{(7)}^n a^{n+1} b^{n+1} c^{n+1}. \end{aligned}$$

Claramente se manifiesta una característica de ambigüedad en la gramática: ambas derivaciones utilizan las mismas reglas de producción, el mismo número de veces, pero las derivaciones se distinguen por el orden en el que se aplican las producciones.

La evidencia anterior sugiere que las únicas palabras que genera la gramática son de la forma $a^i b^i c^i$, con $i \geq 1$. Con base en esto se puede establecer el razonamiento siguiente.

En cualquier derivación que comience con la palabra inicial S , hasta que no se use la producción (2) no se puede usar ninguna de las producciones (4) a (7); esto se debe a que todas estas producciones requieren que un símbolo terminal se encuentre inmediatamente a la izquierda del no terminal que se intenta reescribir. Hasta que se usa la producción (2), todas las formas oracionales derivadas consisten en un cierto número de aes , seguidas de una S , a la cual le siguen una cierta cantidad de Bes y Ces .

Después de que se ha aplicado la producción (2), la forma oracional generada está formada por una secuencia de n aes , seguida de una secuencia de n factores BC ; la forma oracional es una secuencia que comienza con todos los símbolos terminales y termina con todos los símbolos no terminales. En esta palabra ya no aparece el no terminal S , por lo que las producciones (1) o (2) ya no pueden ser utilizadas. Es más, a partir de este momento sólo se pueden aplicar las producciones (3) a (7), y dado que ninguna de ellas regenera a S , las producciones (1) o (2) ya no se pueden volver a aplicar. Ahora bien, la producción (3) sólo involucra a los no terminales B y C , y su efecto es invertir el orden en el que aparecen. Las producciones restantes, (4) a (7), involucra a un terminal seguido de un no terminal, y su efecto es convertir las Bes en bes y las Ces en ces . Las reglas (4) y (6) sólo se aplican en lo que se puede llamar una “frontera”: al pasar de la última a a la primera B , regla (4), y al pasar de la última b a la primera C , la regla (6).

La forma oracional obtenida después de aplicar por única vez la regla (2) tiene la forma general $a^n(BC)^n$ y sólo puede ser reescrita aplicando las reglas (3) o (4); esta última regla sólo puede ser aplicada una vez, ya que sólo hay una “frontera” aB . La regla (4) puede ser aplicada inmediatamente después de haber aplicado la regla (2), o después de haber aplicado la regla (3) una, varias veces, o todas las veces posibles. La aplicación repetida de la regla (3), provoca que todas las Bes se desplacen hacia la izquierda, desplazando con ello a las Ces hacia la derecha. Procediendo así se obtiene la forma oracional $a^n b B^{n-1} C^n$. Esta forma oracional sólo puede ser rescrita usando las producciones (5) a (7), y su efecto final es convertir todas las Bes en bes y todas las Ces en ces .

Supóngase ahora que una C se convierte a una c antes de que todas las Bes hayan sido convertidas a bes ; para que esto ocurra, previamente algunas Bes ya se convirtieron a bes . En este caso, la forma oracional que se obtiene tiene la forma $a^n b^i c \alpha$, donde $i < n$ y α es una secuencia de Bes y Ces , pero ya no de todas las Ces . A esta última forma oracional sólo se le pueden aplicar las producciones (3) o (7). En particular, la producción (7) sólo se puede aplicar a una secuencia terminal no terminal, mientras que la producción (3) se aplica cuando se encuentran dos no terminales consecutivos. La producción (3) reordena a los no terminales, pero no elimina ninguno de ellos; en particular, no remueve ninguna B . Por su parte, la producción (7) convierte las Ces en ces , pero no altera ninguna de las Bes . Eventualmente se debe encontrar una B que ya no puede ser rescrita. Así pues, ya no se puede obtener una palabra terminal.

Se debe concluir entonces que todas las Bes deben ser convertidas a bes antes de convertir las Ces a ces . Dada una forma oracional formada por una secuencia inicial de n aes seguida de una secuencia con n Bes y n Ces , en algún orden, sólo se puede obtener la palabra terminal $a^n b^n c^n$. En consecuencia, $L(G_2) = \{a^i b^i c^i \mid i \geq 1\}$.

Se debe resaltar que en la demostración se consideraron tanto las derivaciones exitosas, esto es, las que comienzan en la palabra de inicio y terminan con una palabra terminal, como las derivaciones no exitosas, o las derivaciones que terminan en una forma oracional que todavía contiene símbolos no terminales y que ya no puede ser rescrita.

Gramáticas equivalentes

Debe resultar claro que las gramáticas G_1 del Ejemplo 1 y G_2 del Ejemplo 2 son equivalentes, esto es, $L(G_1) = L(G_2)$. Nótese que la equivalencia se establece en términos de los lenguajes generados, no en términos de las producciones ni en términos de las derivaciones que se pueden obtener en cada gramática. Las gramáticas G_1 y G_2 son sistemas de rescritura diferentes, pero que tienen un mismo efecto, generan el mismo lenguaje.

Desde el punto de vista del lenguaje se puede afirmar lo siguiente. Al lenguaje $\{a^i b^i c^i \mid i \geq 1\}$ se le pueden imponer, por lo menos, dos estructuras gramaticales (o sintaxis) diferentes, por ejemplo, la establecida por la gramática G_1 , o la establecida por la gramática G_2 . Por el momento sólo se puede señalar que ambas sintaxis son válidas,

pero no se puede decir nada adicional sin tener más información acerca del lenguaje. Adicionalmente, una de las sintaxis es ambigua, pero la otra no. Esto indica que el lenguaje no es inherentemente ambiguo. Se dice que un lenguaje es *inherentemente ambiguo* si todas las sintaxis que se le pueden imponer, o todas las gramáticas que lo generan, son ambiguas. Esta característica tiene consecuencias importantes para los métodos de procesamiento de lenguajes, por ejemplo, en los que se usan en la teoría de la traducción y la ingeniería de compiladores.

Ejemplo 3 Como tercer ejemplo se considera a la gramática G_3 dada por:

$$\begin{aligned} G_3: \quad & X_0 \rightarrow X_1 X_2 X_3, \\ & X_1 X_2 \rightarrow i X_1 Y_i, Y_i j \rightarrow j Y_i, Y_i X_3 \rightarrow X_2 i X_3, i X_2 \rightarrow X_2 i, \\ & X_1 X_2 \rightarrow \varepsilon, X_3 \rightarrow \varepsilon. \end{aligned}$$

En este caso, i, j corresponden a los símbolos 0, 1 de manera independiente. Los símbolos terminales de la gramática son 0 y 1, mientras que los símbolos no terminales son X_0, X_1, X_2, X_3, Y_0 y Y_1 .

A continuación, se muestran algunas de las derivaciones que se pueden obtener con esta gramática.

a) $X_0 \Rightarrow X_1 X_2 X_3 \Rightarrow \varepsilon X_3 = X_3 \Rightarrow \varepsilon.$

La secuencia de producciones empleadas es: $X_0 \rightarrow X_1 X_2 X_3, X_1 X_2 \rightarrow \varepsilon, X_3 \rightarrow \varepsilon.$

b) $X_0 \Rightarrow X_1 X_2 X_3 \Rightarrow 0 X_1 Y_0 X_3 \Rightarrow 0 X_1 X_2 0 X_3 \Rightarrow 0 \varepsilon 0 X_3 = 00 X_3 \Rightarrow 00 \varepsilon = 00.$

La secuencia de producciones empleadas es:

$$\begin{aligned} & X_0 \rightarrow X_1 X_2 X_3, \\ & X_1 X_2 \rightarrow 0 X_1 Y_0, Y_0 X_3 \rightarrow X_2 0 X_3, \\ & X_1 X_2 \rightarrow \varepsilon, X_3 \rightarrow \varepsilon. \end{aligned}$$

Nótese el inicio y la terminación de las derivaciones.

c) $X_0 \Rightarrow X_1 X_2 X_3 \Rightarrow 0 X_1 Y_0 X_3 \Rightarrow 0 X_1 X_2 0 X_3 \Rightarrow 01 X_1 Y_1 0 X_3 \Rightarrow 01 X_1 0 Y_1 X_3 \Rightarrow$
 $01 X_1 0 X_2 1 X_3 \Rightarrow 01 X_1 X_2 01 X_3 \Rightarrow 01 \varepsilon 01 X_3 = 0101 X_3 \Rightarrow 0101 \varepsilon = 0101.$

La secuencia de producciones empleadas es:

$$\begin{aligned} & X_0 \rightarrow X_1 X_2 X_3, \\ & X_1 X_2 \rightarrow 0 X_1 Y_0, Y_0 X_3 \rightarrow X_2 0 X_3, \\ & X_1 X_2 \rightarrow 1 X_1 Y_1, Y_1 0 \rightarrow 0 Y_1, Y_1 X_3 \rightarrow X_2 1 X_3, 0 X_2 \rightarrow X_2 0, \\ & X_1 X_2 \rightarrow \varepsilon, X_3 \rightarrow \varepsilon. \end{aligned}$$

Nótese el rol que tienen los factores X_1X_2 y X_3 ; en especial nótese que no es conveniente eliminar X_3 sino hasta el final de la derivación. De manera similar, nótese el rol de los factores Y_0 y Y_1 en relación con los símbolos terminales que aparecen.

Se demuestra ahora que el lenguaje generado es $L(G_2) = \{xx \mid x \in V_T^*\}$, empleando nuevamente un razonamiento inductivo. En dicho razonamiento es conveniente ordenar las producciones en los tres subconjuntos siguientes:

- (1) $X_0 \rightarrow X_1X_2X_3.$
- (2) $X_1X_2 \rightarrow iX_1Y_i, Y_ij \rightarrow jY_i, Y_iX_3 \rightarrow X_2iX_3, iX_2 \rightarrow X_2i.$
- (3) $X_1X_2 \rightarrow \varepsilon, X_3 \rightarrow \varepsilon.$

El razonamiento procede como sigue.

Considérense las derivaciones D que parten de la palabra $xX_1X_2xX_3$, donde x está formada únicamente con símbolos terminales, y que terminan con una palabra terminal. (Nótese que al aplicar por primera vez la producción (1) se tiene que $x = \varepsilon$.) Si ahora se utiliza alguna de las producciones del subconjunto (3), entonces la otra necesariamente debe ser usada y se termina con la palabra xx , una palabra terminal. Por otro lado, si se aplica la primera producción del subconjunto (2), entonces no queda más alternativa que seguir la derivación hacia las palabras $xiX_1xY_iX_3$ y $xiX_1xX_2iX_3$. De aquí se puede llegar a $xiX_1X_2xiX_3$ o incluso a xiX_1X_2xi . La primera opción tiene la misma forma general que el inicio de D y, en consecuencia, la derivación continúa como ya se mencionó. La segunda opción sólo puede continuar a $xixi$, la cual es una palabra terminal. Con esto se puede concluir que la suposición es válida.

Ejemplo 4 Considérese ahora la gramática G_4 cuyas reglas de producción son las siguientes:

$$\begin{aligned} G_4: \quad S &\rightarrow aB \mid bA \\ A &\rightarrow bAA \mid a \mid aS \\ B &\rightarrow aBB \mid b \mid bS. \end{aligned}$$

En esta descripción, la barra vertical se usa como una disyunción excluyente, se usa el cuerpo de la izquierda o se usa el de la derecha, pero no ambos. Se debe destacar que en esta gramática todas las producciones implican la reescritura de un solo no terminal.

Toda derivación que parta de la palabra inicial S genera una de dos formas oracionales, aB o bA . Si la derivación finaliza en una palabra terminal, entonces estas palabras,

o comienzan con una a , o comienzan con una b . El resto de la palabra terminal depende de las derivaciones que se puedan establecer a partir de los factores A o B . A continuación, se analizan estas posibles derivaciones.

A partir del no terminal A se puede generar el factor bAA , el cual tiene el efecto de añadir una b más y dos A es a la forma oracional; o se puede generar el factor a , eliminando una A ; o se puede generar el factor aS , añadiendo una a y regenerando el no terminal S . Así, a partir del no terminal A se genera una palabra que tiene una a más que las bes , sin no terminales o con un único no terminal que aparece al final de la palabra. El efecto total de estos cambios es obtener, por un lado, una forma oracional terminal formada con la misma cantidad de a es que de bes ; o, por otro lado, una forma oracional formada por una secuencia inicial con la misma cantidad de a es que de bes , en algún orden, que finaliza con el no terminal S . Esto es, se pueden tener las derivaciones siguientes:

$$S \Rightarrow bA \Rightarrow^* w_i, \text{ o}$$

$$S \Rightarrow bA \Rightarrow^* w_j S, \text{ donde } w_i, w_j \text{ son palabras formadas por la misma cantidad de } a\text{es que de } bes, \text{ en algún orden.}$$

Una situación similar se presenta a partir del no terminal B , salvo que ahora se obtienen palabras con una b más que las a es. Al final, nuevamente se obtienen formas oracionales formadas, ya sea por una misma cantidad de a es que de bes , en algún orden, y sin no terminales, o por una misma cantidad de a es que de bes , también en algún orden, que terminan con el no terminal S . Se tienen entonces las derivaciones siguientes:

$$S \Rightarrow aB \Rightarrow^* w_i, \text{ o}$$

$$S \Rightarrow aB \Rightarrow^* w_j S, \text{ donde } w_i, w_j \text{ son palabras formadas por la misma cantidad de } a\text{es que de } bes, \text{ en algún orden.}$$

La evidencia anterior sugiere que las palabras terminales que se pueden obtener a partir de la palabra inicial S están formadas por la concatenación de uno o más factores, cada uno formada por una palabra con la misma cantidad de a es que de bes . De esta manera, el lenguaje generado por la gramática G_4 está formado por todas las palabras no vacías que tienen el mismo número de a es que de bes , en cualquier orden. Esto es, se tiene que:

$$L(G_4) = \{w \in \{a, b\}^* \mid |w|_a = |w|_b\}.$$

Esta aseveración se puede demostrar empleando un razonamiento inductivo, el cual se describe a continuación.

La hipótesis inductiva se establece de la manera siguiente. Para cualquier palabra w formada únicamente por símbolos terminales sucede lo siguiente:

1. $S \Rightarrow^* w$ si y sólo si w consiste en un número igual de a es que de b es.
2. $A \Rightarrow^* w$ si y sólo si w tiene una a más que las b es.
3. $B \Rightarrow^* w$ si y sólo si w tiene una b más que las a es.

No es difícil mostrar que la hipótesis es válida para el caso base $|w| = 1$. En efecto, sucede que $A \Rightarrow^* a$, $B \Rightarrow^* b$ y ninguna palabra de longitud unitaria se puede derivar a partir de S . Además, ninguna otra palabra de longitud unitaria se puede derivar a partir de A o de B .

Supóngase ahora que la hipótesis inductiva es válida para toda palabra w de longitud $k - 1$ o menor. Se muestra ahora que la hipótesis es válida para $|w| = k$. Para empezar, si se asume que $S \Rightarrow^* w$, entonces la derivación debe comenzar con la producción $S \rightarrow aB$ o con la producción $S \rightarrow bA$. En el primer caso, la palabra w debe ser de la forma aw_1 , donde $|w_1| = k - 1$, y donde $B \Rightarrow^* w_1$. Por la hipótesis de inducción, la palabra w_1 debe contener una b más que las a es y, en consecuencia, la palabra w debe contener el mismo número de a es que de b es. Un razonamiento similar se puede establecer si la derivación comienza con la segunda producción.

Se prueba ahora la parte conversas, esto es, si $|w| = k$ y w consiste del mismo número de a es que de b es, entonces debe suceder que $S \Rightarrow^* w$. Para empezar, w debe comenzar con una a o con una b . Supóngase que ocurre lo primero, por lo que $w = aw_1$, con $|w_1| = k - 1$ y w_1 debe tener una b más que las a es. La hipótesis de inducción permite establecer que $B \Rightarrow^* w_1$. Con todo esto se puede establecer entonces la derivación $S \Rightarrow aB \Rightarrow^* aw_1 = w$. Un argumento similar se puede establecer si se asume que la palabra w comienza con una b .

Falta probar las dos hipótesis restantes, pero sus pruebas proceden por razonamientos similares al ya expuesto, por lo que éstos ya no se escriben.

En la siguiente sección se introduce una clasificación especial de las gramáticas (generativas) que va a ser fundamental para definir a las familias de lenguajes gramaticales, o lenguajes generados por gramáticas generativas.

3 CLASES DE GRAMÁTICAS

Las gramáticas son clasificadas imponiendo restricciones a la forma de sus reglas de reescritura. Esta clasificación va a dar origen a una jerarquía de gramáticas, y, en consecuencia, a una jerarquía de familias de lenguajes generados por gramáticas.

En 1959, Noam Chomsky diseñó un esquema para la clasificación de las gramáticas que sigue siendo usado en la actualidad, aunque con algunos pequeños cambios, como se mencionará un poco más adelante. Dicho esquema de Chomsky está basado en tres condiciones restrictivas sobre la forma de las producciones, cada una más restrictiva que la anterior. Tales condiciones son las siguientes.

Sea $G = (V_N, V_T, F, X_0)$ una gramática (generativa) tal y como se definió en la Sección 1.

Primera condición límite. Toda producción $\alpha \rightarrow \beta$ de F satisface la restricción $|\alpha| \leq |\beta|$; se dice que estas producciones no disminuyen la longitud de la palabra generada. En la literatura técnica, las reglas anteriores suelen denominarse reglas que incrementan la longitud, aunque esta denominación es un tanto imprecisa.

Segunda condición límite. Toda producción $\alpha \rightarrow \beta$ de F satisface las condiciones (1) que α es una palabra unitaria, formada necesariamente con un solo símbolo no terminal, y (2) que β es una palabra no vacía sobre el vocabulario. Esto es, las producciones son de la forma $X \rightarrow \beta$.

Tercera condición límite. Toda producción $\alpha \rightarrow \beta$ de F satisface las condiciones (1) que α nuevamente es una palabra unitaria, formada necesariamente con un solo símbolo no terminal, y (2) que β tiene la forma a o la forma aY , donde a es una palabra unitaria terminal y Y es una palabra también unitaria pero no terminal. Esto es, las producciones son de la forma $X \rightarrow a$ o de la forma $X \rightarrow aY$.

Con base en estas condiciones limitantes, las gramáticas se clasifican de la manera siguiente.

- Las *gramáticas de tipo 0* (tipo cero) son las gramáticas que no tienen las restricciones anteriores. Simplemente son las gramáticas (generativas) tal y como se definieron en la Sección 1. Nótese que son gramáticas que deben poseer por lo menos una producción que no incrementa la longitud. Estas gramáticas también se conocen como gramáticas sin restricciones.

- Las *gramáticas de tipo 1* (tipo uno) son las gramáticas que están restringidas por la primera condición límite. Estas gramáticas también son denominadas gramáticas *sensibles al contexto* por razones que se expondrán un poco más adelante. Estas gramáticas obviamente constituyen una subclase de la clase de las gramáticas de tipo 0 y, de hecho, se demuestra que constituyen un subconjunto propio: todas las gramáticas de tipo 1 también son gramáticas de tipo 0, pero existen gramáticas de tipo 0 que no son de tipo 1. En este último caso, debe ser claro que se hace referencia a gramáticas de tipo 0 que contienen por lo menos una producción que no incrementa la longitud.
- Las *gramáticas de tipo 2* (tipo dos) son las gramáticas limitadas por la segunda condición límite. Estas gramáticas se conocen como gramáticas *independientes del contexto*. Dado que sus producciones son de la forma $X \rightarrow \beta$, donde $|X| = 1$ y $|\beta| \geq 1$, claramente estas gramáticas también son sensibles al contexto. Lo inverso, sin embargo, no es cierto, se demuestra que existen gramáticas sensibles al contexto que no son independientes del contexto.
- Las *gramáticas de tipo 3* (tipo tres) son las que satisfacen la tercera condición límite. Estas gramáticas también son conocidas como *gramáticas regulares* o *gramáticas de estado finito* (sobre todo en el contexto de las investigaciones lingüísticas); las razones de estos nombres se presentan en otro documento. Dado que las producciones de estas gramáticas tienen la forma $X \rightarrow aY$ o la forma $X \rightarrow a$, claramente también son gramáticas independientes del contexto; y nuevamente lo converso no es cierto, se demuestra que existen gramáticas independientes del contexto que no son regulares.

Por razones obvias, los tres últimos tipos de gramáticas se conocen como gramáticas con restricciones. Así pues, se ha establecido la siguiente clasificación para las gramáticas generativas.

$$\text{Gramáticas} \begin{cases} \text{Sin Restricciones} \{ \text{Tipo 0} \\ \text{Con Restricciones} \begin{cases} \text{Tipo 1} \\ \text{Tipo 2} \\ \text{Tipo 3} \end{cases} \end{cases}$$

Por extensión, las familias de lenguajes generados por gramáticas de tipo 0, 1, 2 y 3 se conocen como las familias de lenguajes de tipo 0, 1, 2 y 3 respectivamente. Las familias de lenguajes recién definidas se denotan $\mathcal{L}_i$ para $i = 0, 1, 2, 3$. De la definición se puede establecer que los lenguajes de tipo 3 también son de los tres tipos restantes; los

lenguajes de tipo 2 también son de los dos tipos restantes, y los lenguajes de tipo 1 también son lenguajes de tipo 0. Uno de los resultados más importantes de la teoría de los lenguajes formales es que existen lenguajes de un tipo superior que no son del tipo inferior, por decir, existen lenguajes de tipo 2 que no son de tipo 3. Esto significa que existen lenguajes que pueden ser generados por gramáticas de tipo 2, pero que no pueden ser generados por gramáticas de tipo 3. Este importante resultado se presentará en otro documento, ya que para su demostración se requiere del uso de los autómatas. En este mismo documento también se presentarán algunas de las propiedades más importantes de los diferentes tipos de gramáticas y, en consecuencia, de sus familias de lenguajes generados.

Cabe señalar que Noam Chomsky originalmente definió a las gramáticas sensibles al contexto como las gramáticas cuyas producciones son de la forma $\omega_1 X \omega_2 \rightarrow \omega_1 \beta \omega_2$, donde ω_1, ω_2 son palabras sobre V^* , $X \in V_N^1$ y $\beta (\neq \varepsilon) \in V^*$. Con base en esta definición, dichas producciones pueden ser interpretadas de la manera siguiente. La palabra X puede ser reemplazada con (o reescrita como) β sólo si se encuentra entre las palabras ω_1 y ω_2 , o en el contexto que forman ω_1 y ω_2 . De aquí que se diga que la gramática es sensible al contexto. A diferencia de esto, la forma de las producciones de las gramáticas de tipo 2 simplemente indica que la palabra X puede ser reemplazada por la palabra β , sin que importe el entorno en el que se encuentra X . Por esto se dice que estas gramáticas son independientes del contexto.

La aparente discrepancia en las definiciones en realidad no existe, como se muestra en el resultado siguiente.

Teorema 3.1

La clase de lenguajes generados por gramáticas que contienen exclusivamente producciones sensibles al contexto es la clase de los lenguajes de tipo 1.

Prueba

Sea G_1 una gramática de tipo 1 y sea G_{sc} una gramática que sólo contiene producciones sensibles al contexto. Claramente una gramática como G_{sc} es una gramática de tipo 1 ya que todas sus producciones incrementan la longitud. Ahora se muestra que lo converso también es cierto, esto es, toda gramática de tipo 1 es sensible al contexto.

Considérese que la gramática G_1 de tipo 1 es de la forma $G_1 = (V_N, V_T, F, X_0)$. Se muestra ahora que existe una gramática equivalente $G_a = (V_{N_a}, V_T, F_a, X_0)$ en la que todas las producciones $\alpha \rightarrow \beta$ de F_a tienen la “forma normal” siguiente: ya sea que α y β sean palabras exclusivamente de no terminales, o que α y β tienen la forma A y a , respectivamente, esto es, las producciones son de la forma $A \rightarrow a$. La gramática G_a se construye a partir de G_1 como sigue. Considérese que $V_{N_a} = V_N \cup \{X_a \mid a \in V_T\}$, esto es, V_{N_a} consiste en todos los símbolos de V_N más los símbolos X_a para cada a de V_T . Con este alfabeto, todas las producciones de F_a se forman a partir de las producciones de F cambiando cualquier aparición de un símbolo terminal a por el no terminal X_a y añadiendo la producción $X_a \rightarrow a$ para todo a de V_T . Por construcción se tiene que $L(G_1) = L(G_a)$.

A partir de G_a se construye ahora una gramática equivalente G_b cuyas producciones son del tipo sensible al contexto. Sea $\alpha \rightarrow \beta$ una producción de F_a , donde α tiene la forma $A_1 A_2 \dots A_m$, mientras que β tiene la forma $B_1 B_2 \dots B_n$, con $n \geq m$. Para cada producción en esta “forma normal” se puede construir un conjunto de producciones sensibles al contexto en F_b que tienen la forma siguiente:

$$\begin{aligned}
A_1 A_2 A_3 \dots A_{m-1} A_m &\rightarrow A'_1 A_2 A_3 \dots A_{m-1} A_m, \\
A'_1 A_2 A_3 \dots A_{m-1} A_m &\rightarrow A'_1 A'_2 A_3 \dots A_{m-1} A_m, \\
A'_1 A'_2 A_3 \dots A_{m-1} A_m &\rightarrow A'_1 A'_2 A'_3 \dots A_{m-1} A_m, \\
&\vdots \\
A'_1 A'_2 A'_3 \dots A_{m-1} A_m &\rightarrow A'_1 A'_2 A'_3 \dots A'_{m-1} A_m, \\
A'_1 A'_2 A'_3 \dots A'_{m-1} A_m &\rightarrow A'_1 A'_2 A'_3 \dots A'_{m-1} A'_m. \\
A'_1 &\rightarrow B_1, \\
A'_2 &\rightarrow B_2, \\
A'_3 &\rightarrow B_3, \\
&\vdots \\
A'_{m-1} &\rightarrow B_{m-1}, \\
A'_m &\rightarrow B_m B_{m+1} \dots B_n.
\end{aligned}$$

El primer conjunto de producciones sensibles al contexto (las que involucran a los no terminales A_1 a A_m) remplazan la palabra $\alpha = A_1 A_2 \dots A_m$ por la palabra $A'_1 A'_2 \dots A'_m$ de nuevos símbolos no terminales. Esta última palabra puede ser remplazada, a su vez, por la palabra $B_1 B_2 \dots B_n$ empleando el segundo conjunto de producciones sensibles al contexto (las que involucran a los no terminales A'_1 a A'_m), recordando que $n \geq m$. Si ahora se construye un nuevo conjunto de símbolos no terminales V_{N_b} que contenga a todos los

símbolos de V_{N_a} más todos los nuevos símbolos A'_i , para $1 \leq i \leq m$, y con este conjunto más el conjunto de producciones anteriores se construye una nueva gramática G_b , resulta que, por construcción, esta nueva gramática es equivalente a G_a y, en consecuencia, también es equivalente a G_1 . Pero ahora se tiene que G_b sólo contiene producciones sensibles al contexto.

Para ejemplificar la construcción empleada en la demostración del último teorema, se va a usar a la gramática G_1 ejemplificada en la sección anterior, ya que es una gramática de tipo 1. Por comodidad, se repiten las producciones que forman a esta gramática.

$$\begin{aligned} G_1: \quad & X_0 \rightarrow abc, X_0 \rightarrow aX_1bc, \\ & X_1b \rightarrow bX_1, X_1c \rightarrow X_2bcc, \\ & bX_2 \rightarrow X_2b, aX_2 \rightarrow aaX_1, aX_2 \rightarrow aa. \end{aligned}$$

Para construir la gramática G_a es necesario considerar los nuevos símbolos no terminales X_a, X_b y X_c , así como las nuevas producciones $X_a \rightarrow a$, $X_b \rightarrow b$ y $X_c \rightarrow c$. Con estos componentes, la gramática G_a equivalente está formada por las producciones siguientes:

$$\begin{aligned} G_a: \quad & X_0 \rightarrow X_aX_bX_c, \\ & X_0 \rightarrow X_aX_1X_bX_c, \\ & X_1X_b \rightarrow X_bX_1, \\ & X_1X_c \rightarrow X_2X_bX_cX_c, \\ & X_bX_2 \rightarrow X_2X_b, \\ & X_aX_2 \rightarrow X_aX_aX_1, \\ & X_aX_2 \rightarrow X_aX_a, \\ & X_a \rightarrow a, \\ & X_b \rightarrow b, \\ & X_c \rightarrow c. \end{aligned}$$

Es claro que la gramática G_a sigue siendo una gramática de tipo 1, pero no todas sus producciones son de tipo sensible al contexto. También es claro que en esta nueva gramática hay dos tipos de conjuntos de producciones. En el primero de ellos están involucradas únicamente producciones formadas por símbolos no terminales. En el segundo conjunto ya se encuentran producciones que contienen ambos tipos de símbolos.

A partir de la gramática G_a se puede construir una nueva gramática equivalente formada exclusivamente por producciones sensibles al contexto. Esta construcción procede de la manera siguiente.

Producción de G_a	Producciones de G_b
---------------------	-----------------------

$$\begin{array}{ll}
X_1X_b \rightarrow X_bX_1 & X_1X_b \rightarrow X'_1X_b, \\
& X'_1X_b \rightarrow X'_1X'_b, \\
& X'_1 \rightarrow X_b, \\
& X'_b \rightarrow X_1. \\
X_bX_2 \rightarrow X_2X_b & X_bX_2 \rightarrow X'_bX_2, \\
& X'_bX_2 \rightarrow X'_bX'_2, \\
& X'_b \rightarrow X_2, \\
& X'_2 \rightarrow X_b.
\end{array}$$

Claramente, todas las producciones de la gramática G_b son de la forma sensible al contexto. También es muy claro que la gramática G_b tiene más producciones. Esto generalmente se repite en muchos métodos de construcción de gramáticas equivalentes. Normalmente, los métodos de construcción que se proponen son efectivos, generan el resultado deseado, pero no necesariamente son eficaces, las gramáticas construidas pueden ser muy engorrosas. En este momento interesa sólo el efecto, se muestra que la gramática deseada realmente existe.

4 LA PALABRA VACÍA

Según la definición de las familias de lenguajes, las familias generadas por gramáticas con restricciones están formadas por lenguajes que no pueden contener a la palabra vacía ε ; esto se debe a que producciones de la forma $\alpha \rightarrow \varepsilon$ no incrementan la longitud. Esto no necesariamente es una limitación ya que si un lenguaje L es generado por alguna gramática con restricciones G , esto es $L = L(G)$, y se requiere que L contenga la palabra vacía, entonces simplemente se le añade, $L \cup \{\varepsilon\}$. Sin embargo, se puede realizar una pequeña modificación a la definición de las gramáticas restringidas para permitir la inclusión de gramáticas que puedan generar a la palabra vacía. Esta modificación se describe a continuación.

La idea en la que se basa la modificación considera lo siguiente. Supóngase que la gramática en cuestión tiene una sola producción de la forma $X_0 \rightarrow \varepsilon$, donde X_0 es la palabra axioma (o de inicio), y que esta palabra axioma ya no aparece en el cuerpo de ninguna otra de las producciones de la gramática. Bajo estos supuestos, es claro que dicha producción sólo puede ser usada al inicio de una sola derivación, $X_0 \Rightarrow \varepsilon$,

precisamente la que genera a la palabra vacía. Esta forma de constitución de las gramáticas es general y se puede establecer considerando primero el resultado siguiente.

Lema 4.1

Si $G = (V_N, V_T, F, X_0)$ es una gramática restringida, entonces existe otra gramática restringida G' , del mismo tipo y que genera el mismo lenguaje que G , pero con la diferencia de que la palabra axioma (o inicial) de G' no aparece en el cuerpo de ninguna producción de G' .

Prueba

Sea X un símbolo que no se encuentra ni en V_N ni en V_T . Considérese ahora la nueva gramática $G' = (V_N \cup \{X\}, V_T, F_1, X)$. En este caso, el conjunto F_1 consiste en todas las producciones de F más todas las producciones de la forma $X \rightarrow \alpha$, donde $X_0 \rightarrow \alpha$ es una producción de F . Debido a que X no es un símbolo ni de V_N ni de V_T , entonces no aparece en el cuerpo de ninguna producción de F . Se va a mostrar que $L(G) = L(G')$.

Supóngase que $X_0 \Rightarrow_G^* \omega$ y que la primera producción que se usó fue $X_0 \rightarrow \alpha$. Entonces se tiene que $X_0 \Rightarrow_G \alpha \Rightarrow_G^* \omega$. Ahora bien, por la definición de F_1 se tiene que $X \rightarrow \alpha$ está en F_1 y, en consecuencia, se tiene que $X \Rightarrow_{G'} \alpha$. Y puesto que F_1 contiene todas las producciones de F , resulta que $\alpha \Rightarrow_{G'}^* \omega$. Así pues, se tiene que $X \Rightarrow_{G'}^* \omega$, y esto prueba que $L(G) \subseteq L(G')$.

Supóngase ahora que $X \Rightarrow_{G'}^* \omega$ y que la primera producción usada fue $X \rightarrow \alpha$ para alguna palabra α . De esto se deduce que $X_0 \rightarrow \alpha$ es una producción de F y con ello se tiene que $X_0 \Rightarrow_G \alpha$. A partir de la suposición inicial se tiene que $\alpha \Rightarrow_{G'}^* \omega$, donde α no contiene a X . Además, como X no aparece en el cuerpo de ninguna producción de F_1 , ninguna palabra en la derivación $\alpha \Rightarrow_{G'}^* \omega$ puede involucrar a X . Debido a esto, la derivación anterior también es una derivación que sucede en G , esto es, $\alpha \Rightarrow_G^* \omega$. Así pues, se concluye que $X_0 \Rightarrow_G^* \omega$ y con ello que $L(G') \subseteq L(G)$. Finalmente se tiene que $L(G) = L(G')$.

Con base en lo anterior, se establece ahora el resultado que anticipamos.

Teorema 4.2

Si L es un lenguaje generado por alguna gramática restringida, entonces $L \cup \{\varepsilon\}$ y $L - \{\varepsilon\}$ también son lenguajes del mismo tipo.

Prueba

Dada una gramática restringida G , mediante el lema anterior se puede construir una gramática equivalente G' cuya palabra axioma (o de inicio) no aparece en el cuerpo de ninguna de sus producciones. Sea pues $G' = (V_N, V_T, F, X_0)$.

Se define ahora a la gramática $G_1 = (V_N, V_T, F_1, X_0)$ en la que su conjunto F_1 contiene todas las producciones de F junto con la producción $X_0 \rightarrow \varepsilon$. Nótese que X_0 no aparece en el cuerpo de ninguna producción de F_1 , por lo que la producción $X_0 \rightarrow \varepsilon$ no puede ser usada más que como la primera y única producción en una derivación. Cualquier derivación de G_1 que no involucre a $X_0 \rightarrow \varepsilon$ también es una derivación de G' , por lo que $L(G') = L(G) \cup \{\varepsilon\}$.

Por otro lado, si la gramática $G' = (V_N, V_T, F, X_0)$ genera al lenguaje L que contiene a la palabra vacía ε , entonces el conjunto F debe contener a la producción $X_0 \rightarrow \varepsilon$, y además la palabra X_0 no aparece en el cuerpo de ninguna producción de F . Entonces se puede formar la gramática $G_1 = (V_N, V_T, F_1, X_0)$ en la que el conjunto F_1 contienen todas las producciones de F , a excepción de $X_0 \rightarrow \varepsilon$. Debido a que la producción $X_0 \rightarrow \varepsilon$ no puede ser usada en la derivación de ninguna palabra no vacía, G_1 sólo puede generar palabras no vacías y se concluye que $L(G_1) = L(G') - \{\varepsilon\}$.

Este resultado permite extender la clasificación de los tipos de gramáticas restringidas a gramáticas cuyo conjunto de producciones incluye a la producción $X_0 \rightarrow \varepsilon$, siempre y cuando la palabra X_0 no aparezca en el cuerpo de ninguna producción.

Para ejemplificar el método de construcción usado en la demostración de los resultados anteriores se va a hacer uso de la gramática G_2 analizada párrafos atrás. Recordemos que las producciones de esta gramática son:

- | | |
|-------------------------|------------------------|
| 1. $S \rightarrow aSBC$ | 5. $bB \rightarrow bb$ |
| 2. $S \rightarrow aBC$ | 6. $bC \rightarrow bc$ |
| 3. $CB \rightarrow BC$ | 7. $cC \rightarrow cc$ |
| 4. $aB \rightarrow ab$ | |

Se puede construir una gramática $G'_2 = (\{S_1, S, B, C\}, \{a, b\}, F_1, S_1)$, en la que el símbolo no terminal S_1 claramente no aparece en la gramática G_2 original, y cuya lista de

producciones incluye a todas las producciones anteriormente listadas más las dos producciones siguientes:

$$S_1 \rightarrow aSBC \text{ y } S_1 \rightarrow aBC.$$

De esta manera, ninguna de las producciones de G'_2 presenta a la palabra axioma en su cuerpo. Se cumplen entonces las condiciones del Lema anterior y se concluye que la nueva gramática G'_2 es equivalente a la gramática G_2 original.

Adicionalmente, a partir de la gramática G'_2 se puede construir una nueva gramática G''_2 que además de generar las mismas palabras del lenguaje $L(G'_2)$ también genera a la palabra vacía ε . Esta nueva gramática es $G''_2 = (\{S_1, S, B, C\}, \{a, b\}, F'_{11}, S_1)$, cuyo conjunto de producciones está constituido por todas las producciones de G'_2 más la producción $S_1 \rightarrow \varepsilon$. De esta manera se cumplen las condiciones del teorema anterior y se puede concluir que $L(G''_2) = L(G'_2) \cup \{\varepsilon\}$.

Nótese que las gramáticas equivalentes G_2 y G'_2 son de tipo 1, y se considera que la gramática G''_2 sigue siendo de tipo 1.

Como ya se había mencionado previamente, en otro documento se presentarán las propiedades más importantes de los diferentes tipos de gramáticas y, en consecuencia, de sus familias de lenguajes generados, así como una demostración del llamado teorema de jerarquía.

Referencias Bibliográficas

Hopcroft, J. E. and Ullman, J. D. (1969) **Formal Languages and Their Relation to Automata**. Reading, Massachusetts, Addison-Wesley Publishing Company.

Levelt, W. J. M. (1974) **Formal Grammars in Linguistics and Psycholinguistics. Volume I, An Introduction to the Theory of Formal Languages and Automata**. The Hague – Paris, Mouton & Co. N. V.

Parkes, A. P. (2008) **A Concise Introduction to Languages and Machines**. Undergraduate Topics in Computer Science, London Limited, Springer-Verlag.

Salomaa, A. (1973) **Formal Languages**. New York and London, Academic Press.