2. LENGUAJES NATURALES Y LENGUAJES FORMALES

2.1 INTRODUCCIÓN

Existen dos tipos básicos y reconocidos de lenguajes: los lenguajes naturales y los lenguajes formales. El origen y desarrollo de los primeros, como pueden ser el castellano, el inglés o el francés, es natural, es decir, sin el control de ninguna teoría. Las teorías de lenguajes naturales y las gramáticas, fueron establecidas a priori, esto es, después de que el lenguaje había ya madurado. Por otro lado, los lenguajes formales como las matemáticas y la lógica, fueron desarrollados generalmente a través del establecimiento de una teoría, la cual le da las bases para dichos lenguajes.

2.2 DEFINICIÓN DE LENGUAJE

Las lenguas son sistemas más o menos complejos, que asocian contenidos de pensamiento y significación a manifestaciones simbólicas tanto orales como escritas. Aunque en sentido estricto, el lenguaje sería la capacidad humana para comunicarse mediante lenguas, se suele usar para denotar los mecanismos de comunicación no humanos (el lenguaje de las abejas o el de los delfines), o los creados por los hombres con fines específicos (los lenguajes de programación, los lenguajes de la lógica, los lenguajes de la aritmética...).

Nosotros, vamos a definir el lenguaje como un conjunto de palabras. Cada lenguaje está compuesto por secuencias de símbolos tomados de alguna colección finita. En el caso de cualquier lengua natural (castellano, inglés, francés...), la colección finita es el conjunto de las letras del alfabeto junto con los símbolos que se usan para construir palabras (tales como el guión, el apóstrofe en el caso del inglés...). De forma similar, la representación de enteros, son secuencias de caracteres del conjunto de los dígitos {0,1,2,3,4,5,6,7,8,9}.

Un conjunto no vacío y finito de símbolos se conoce como alfabeto. Si $\sum$ es un alfabeto, y $\sigma \in \sum$ denota que σ es un símbolo de $\sum$. Por tanto, si $\sum$ = {0,1,2,3,4,5,6,7,8,9}, podemos decir que $0 \in \sum$. Obsérvese, que puesto que un alfabeto es simplemente un conjunto finito no vacío, dados $\sum 1$ y $\sum 2$ alfabetos, se tiene que

$\Sigma 1 \cup \Sigma 2$ también lo es. Es más, $\Sigma 1 \cap \Sigma 2$, $\Sigma 1 - \Sigma 2$ y $\Sigma 2 - \Sigma 1$, también son alfabetos.

Una secuencia finita de símbolos de un determinado alfabeto, se conoce como palabra sobre dicho alfabeto. Nuestra experiencia, nos lleva a identificar el término palabra con las palabras de cualquier lenguaje natural, por esta razón, a menudo se usa el término cadena en lugar de palabra, con el fin de evitar esta idea preconcebida. Se tratarán igual los términos cadena y palabra.

Cada símbolo de un alfabeto, es una cadena sobre dicho alfabeto. La cadena vacía, es una palabra sobre cualquier alfabeto. La palabra vacía, es una secuencia vacía de símbolos, tomados de cualquiera que sea el alfabeto en cuestión.

Los lenguajes, pueden ser bastante grandes, como lo es el caso de todas las palabras "correctas" que se pueden formar en castellano sobre el alfabeto castellano. Dado que un lenguaje es un conjunto de cadenas, se puede tener el lenguaje compuesto por ninguna cadena, el lenguaje vacío. Éste, no es el mismo lenguaje que el que consta de la cadena vacía.

2.3 LENGUAJES NATURALES vs LENGUAJES FORMALES

Como se ha explicado en el apartado anterior, en un lenguaje, se tiene que los elementos más simples, son los símbolos llamados letras que constituyen un alfabeto Σ, que es un conjunto finito de símbolos $\{\sigma_1, \sigma_2, ..., \sigma_n\}$. Con la concatenación de las letras, formaremos palabras que determinan un conjunto Σ^*. El conjunto de palabras que tengan un significado, constituirán el diccionario del lenguaje (por ejemplo el Webster, diccionario del inglés). A partir de lo anterior, tendremos que un lenguaje se considera como un conjunto de oraciones, que usualmente es infinito y, se forman con palabras del diccionario. En este punto, podemos distinguir entre dos clases de lenguajes; los lenguajes naturales como el castellano o el inglés, y los lenguajes formales como las matemáticas y la lógica.

El lenguaje castellano, de un modo extensivo, puede ser definido como el conjunto (teóricamente infinito) de todas las oraciones en castellano. Como la mayoría de los lenguajes de interés, son recursivos en mayor o menor medida (a partir de una oración, existen procedimientos que permiten formar otras mayores y más complejas), debemos encontrar propiedades o conjuntos de propiedades, que las definan unívocamente (definición intensiva).

Dada la oración castellana: “el coche es gris”, es posible construir otras como:

- “mi amigo dice que el coche es gris”
- “si mi amigo dice que el coche es gris, es que el coche es gris”

- “si me contaron que mi amigo dice: ”el coche es gris”, mi amigo dice que el coche es gris”

Como es obvio, resultaría absurdo intentar escribir todas las posibles combinaciones de palabras que hay en el lenguaje castellano.

Las oraciones, son consistentes en forma natural con la experiencia práctica humana, que se organiza automáticamente al tiempo que organiza el lenguaje en sí mismo. Una oración en castellano, es una secuencia finita de palabras del castellano, donde sabemos que el conjunto de esas palabras es finito. Sin embargo, no todas las combinaciones de palabras son permitidas, es necesario que esas combinaciones sean correctas (con respecto a una sintaxis) y tengan sentido (con respecto a la semántica). Esa sintaxis y esa semántica constituyen un orden en la teoría del lenguaje castellano: aquel que permite la definición de todas las oraciones en castellano y así, del lenguaje castellano. Por ejemplo, dado el conjunto de palabras pertenecientes al diccionario del castellano: {el, hombre, tomó, compró, balón}, habrá frases que se puedan formar con dicho conjunto que sean correctas con respecto a una sintaxis y a una semántica, como:

- “el hombre tomó el balón”
- “el hombre compró el balón”

otras que serán correctas sintácticamente, pero no semánticamente:

- “el balón compró el hombre”, o
- “el balón tomó el hombre”

y otras que no sean ni sintáctica ni semánticamente correctas, como por ejemplo:

- “tomó compró balón el”
- “el tomó hombre balón el”

De la particularización anterior, se desprende que en un lenguaje natural, como el castellano, la formación de las oraciones precedió a la formalización del lenguaje por medio de una teoría o una gramática. Por esta razón, un lenguaje es llamado natural, porque es no artificial o no construido. El calificativo “natural”, se opone al de “formal”, el cual determina un lenguaje que es construido estableciendo una teoría y, por ende, se le llamaría artificial. Un lenguaje formal como la lógica, consiste de un conjunto de oraciones, llamadas fórmulas o expresiones bien formadas. La calificación de “lenguaje artificial”, se refiere al hecho de que se forma por medio de reglas de formación. El calificativo “formal”, se refiere específicamente al hecho de que las oraciones de estos lenguajes, consisten de una lista de símbolos sujetos a diversas interpretaciones. Por otro lado, en los lenguajes naturales, las palabras en una oración poseen un significado y tienen su significante. Esto quiere decir, que independientemente del significado de cada palabra, debemos tener en cuenta el sentido correcto que éstas adquieren, según el contexto en el que se expresen en un momento dado. Una de las metas en computación, es poder especificar rigurosamente estos significados, por los métodos de interpretación de los sistemas formales. Estos métodos en cuestión, constituyen las semánticas del lenguaje formal.

Resumidamente, tenemos que los lenguajes naturales y los formales, difieren significativamente uno de otro por su origen y por su área de aplicación. Vamos a intentar identificar las propiedades más importantes de estos dos tipos de lenguaje, con el fin de responder a una de las preguntas que se suele hacer a la hora de ponerse a programar un analizador sintáctico automático, que se aplicará sobre un texto escrito en cualquier lengua natural, como puede ser el castellano: ¿Hasta qué punto, pueden los lenguajes naturales ser representados (traducidos) por medio de lenguajes formales?

2.3.1 Propiedades de los lenguajes naturales

El lenguaje es la función que expresa pensamientos y comunicaciones entre la gente. Esta función, es llevada a cabo por medio de señales vocales (voz) y, posiblemente, por signos escritos (escritura), que conforman el lenguaje natural. Con respecto a nuestro mundo, el lenguaje nos permite designar las cosas actuales (y razonar a cerca de ellas) y crear significados.

Contrariamente a lo que ciertas teorías lingüísticas formales harían a uno creer, el lenguaje natural, no fue fundamentado sobre una verdad racional a priori, pero fue desarrollado y organizado, a partir de la experiencia humana, en el mismo proceso en que esta experiencia humana, fue organizada. En su forma actual, los lenguajes naturales, tienen un gran poder expresivo y pueden ser utilizados para analizar situaciones altamente complejas y razonar muy sutilmente. La riqueza de su componente semántico, y su cerrada relación con los aspectos prácticos de los contextos en los cuales son usados, da a los lenguajes naturales, su gran poder expresivo y su valor como una herramienta para razonamiento sutil.

Así como la formalización del componente semántico de un lenguaje natural, es decir, el constituyente del lenguaje por el cual las oraciones tienen o adquieren su significado, es bastante complicado, por otra parte, la sintaxis de un lenguaje natural, puede ser modelada fácilmente por un lenguaje formal similar a los utilizados en las matemáticas o en la lógica.

Otra propiedad única de los lenguajes naturales, es la polisémica, es decir, la posibilidad de que una palabra en una oración, tenga diversos significados, diversos valores. Por ejemplo, la palabra “pair” en el inglés, puede ser considerada primero como un sustantivo, y es usada entonces, en estructuras de frases como:

- “arrange in pairs”
- “the happy pair”

sin embargo, puede también ser interpretada como un verbo transitivo en frases como por ejemplo:

- “two vases that pair”
- “to pair off with someone”

El carácter polisemico de un lenguaje, tiende a incrementar la riqueza de su componente semántico, más aún, este hecho no hace la formalización difícil, sino

imposible. El carácter polisemico de los lenguajes, es considerada una propiedad adquirida recientemente, las formas primarias de los lenguajes naturales habrían sido similares a los lenguajes formales, y la polisemántica sería el resultado de un enriquecimiento progresivo. En suma, los lenguajes naturales se caracterizan por las siguientes propiedades:

- Desarrollados por enriquecimiento progresivo, antes de cualquier intento de formación de una teoría.
- La importancia de su carácter expresivo, es debida grandemente a la riqueza de el componente semántico.
- Dificultad o imposibilidad de una formalización completa.

2.3.2 Propiedades de los lenguajes formales

La definición de una teoría de un lenguaje formal dado, precedió a su definición intensiva, como hemos llamado antes al establecimiento de una serie de propiedades o fórmulas, que definan unívocamente las oraciones correctas que componen un lenguaje natural.

El proceso de generación y desarrollo de un lenguaje formal, es inverso al de los lenguajes naturales, consecuentemente, las palabras y las oraciones de un lenguaje formal, son perfectamente definidas: una palabra mantiene el mismo significado prescindiendo del contexto en el que se encuentre. Como resultado de este proceso, obtendremos las llamadas gramáticas libres del contexto. En adición, el significado de símbolos es determinado exclusivamente por la sintaxis, sin referencia a ningún contenido semántico. Una función y una fórmula, puede designar cualquier cosa, solamente los operadores y relaciones que nos permiten escribir una fórmula como por ejemplo la igualdad, desigualdad, pertenencia, no pertenencia, conectivos lógicos, etc., y operadores algebraicos +, *, etc., tienen significados especiales.

Los lenguajes formales son, por todo esto, necesariamente exentos de cualquier componente semántico fuera de sus operadores y relaciones, y es gracias a esta ausencia de significado especial, que los lenguajes formales pueden ser usados para modelar una teoría de la mecánica, de la ingeniería electrónica, etc., en la lingüística u otra naturaleza, la cual asume el estatus del componente semántico de tal lenguaje. Esto equivale a decir, que durante la concepción de lenguajes formales, toda la ambigüedad anteriormente expuesta respecto a la semántica de una palabra, es anulada, es como si esta reducción al significado único debe manifestarse por sí mismo, como la eliminación del “mundo de significados” en el proceso de construir las fórmulas, al tiempo que se toca el nivel abstracto de estas construcciones. Es solamente, por medio de un paso adicional, que el significado es asignado a las fórmulas. Este paso, nos permite la posibilidad de asignar un criterio falso/cierto a cada fórmula.

El mundo de significados que es el componente semántico, solo existe en la teoría que uno intenta expresar a través del lenguaje formal. Por ejemplo, un componente semántico normalmente asociado con el lenguaje formal de una teoría cónica, es el movimiento de los cuerpos celestes, así mismo, sistemas lineales de todas las órdenes, son posibles componentes semánticos de teoría de matrices.

No podemos evitar mencionar, la importancia de los números en lenguajes formales. En un sistema numérico, así como en un sistema de cálculo, los números siempre tienen el potencial de referir un cierto "contenido", el cual pertenecerá entonces al componente semántico del lenguaje: los objetos posibles cuando son contables o medibles. La asociación de un significado con un número o con un cálculo, no siempre es obvio, sin embargo, es útil recordar, que en física, cuando se completa un cálculo y se busca una interpretación del mismo, solamente se mantienen los números positivos de los resultados, ya que las soluciones negativas o imaginarias a las ecuaciones que se supone describen la realidad, son la mayoría de las veces rechazadas, porque no corresponden con la "realidad física". En resumen, los lenguajes formales, se caracterizan con las siguientes propiedades:

- Se desarrollan a partir de una teoría establecida.
- Tienen un componente semántico mínimo.
- Posibilidad de incrementar el componente semántico de acuerdo con la teoría a formalizar.
- La sintaxis produce oraciones no ambiguas, en lo que respecta al significado de sus palabras.
- Completa formalización, y por esto, el potencial de la construcción computacional.

2.3.3 Conclusiones

Son, por tanto, los formalismos los que pueden abrirnos la puerta del tratamiento informático cómodo y generalizado: dada una secuencia perteneciente a un idioma como el nuestro y un conjunto de propiedades que debe satisfacer, sólo habremos de aplicarlas para verificar su gramaticalidad. Bien es cierto, que será necesario imponer una serie de restricciones a los formalismos, si queremos que exista un algoritmo o una secuencia de pasos, que nos garantice la finalización exitosa o no, de la operación en un tiempo finito. Del mismo modo, se podría llegar a la traducción entre lenguajes formales, a la generación más o menos intencionada y automática de oraciones correctas...

Para llevar a cabo el objetivo de formalizar los lenguajes naturales, habrá que definir una gramática. Esta gramática será un conjunto de reglas que definirán si una secuencia arbitraria de símbolos es correcta. Decimos entonces, que una frase correcta pertenecerá al lenguaje.

2.4 DEFINICIÓN DE UNA GRAMÁTICA

El paradigma formal más famoso y rápidamente desarrollado para la caracterización de lenguajes, es el derivado del concepto de gramática generativa de Noam Chomsky. En Chomsky[56], este celebérrimo lingüista norteamericano que intentó formalizar los lenguajes naturales, partiendo de que un lenguaje L es un

subconjunto de todas las secuencias (finitas o no), que podemos formar mediante la concatenación de los elementos de un alfabeto $\sum$, define la gramática mediante la cuaterna siguiente:

- $\sum$: vocabulario finito de símbolos Terminales. Éstos son los símbolos que realmente aparecen en una frase. Nunca aparecerán en el lado izquierdo de una producción (lo cual definiremos más adelante dentro de esta cuaterna). Los símbolos terminales deben ser símbolos válidos del lenguaje.
- N : conjunto finito de símbolos No Terminales, los cuales son metasímbolos que deben ser definidos por otras producciones (o reglas gramaticales), es decir, que también aparecen en el lado izquierdo de las mismas. Los símbolos No Terminales se pueden definir como variables sintácticas.
- S : un símbolo No Terminal básico (axiomático, según la definición de Chomsky). Será el símbolo principal o axioma que describirá oraciones enteras (y no subcadenas, como describen los símbolos No Terminales) de un lenguaje natural.
- P : conjunto, también finito, de reglas que nos dicen cómo se pueden generar las oraciones, cómo partiendo del axioma, podemos llegar a la oración terminal. Este conjunto será un simple subconjunto de:

$$P \subseteq (N \cup \sum)^* N (N \cup \sum)^* \times (N \cup \sum)^*$$

lo cual, expresado en una notación más clásica en lingüística sería:

$$\alpha A \beta = \gamma$$

donde:

$$A \in N$$
$$\alpha, \beta, \gamma \in (N \cup \sum)^*$$

siendo posible que tanto α como β sean iguales a la cadena nula.

El lenguaje L así definido, se obtendría aplicando el siguiente procedimiento no algorítmico (no garantiza tiempo finito para una gramática genérica):

- S es una forma oracional.
- Si $\alpha\beta\gamma$ es una forma oracional, y $\beta = \delta$ pertenece a P, $\alpha\delta\gamma$ también será forma oracional.
- Una forma oracional compuesta únicamente por símbolos Terminales, constituirá una oración del lenguaje.

Como ejemplo, veamos que el enunciado en castellano: “El hombre compró el libro”, puede derivarse, basándonos en una pequeña gramática, mediante la siguiente secuencia de producciones:

S → SintagmaNominal SintagmaVerbal
→ Artículo Nombre SintagmaVerbal
→ El Nombre SintagmaVerbal

→El hombre SintagmaVerbal
→El hombre Verbo SintagmaNominal
→El hombre compró SintagmaNominal
→El hombre compró Artículo Nombre
→El hombre compró el Nombre
→El hombre compró el libro

Entonces, si abreviamos la frase "El hombre compró el libro" ≡ Ehcel, tenemos que: Ehcel* puede producir todas las combinaciones. Además, Ehcel será aceptado por el lenguaje definido por el conjunto de producciones anterior, Ehcel ∈ L, ya que Ehcel ∈ $\sum$*.

Ahora, podemos definir entonces un lenguaje L como el conjunto de todas las cadenas de símbolos Terminales que pueden derivarse del símbolo inicial o axioma S:

$$L = \{ \sigma \mid S \text{ es una secuencia de } \sigma \text{ y } \sigma \in \sum^* \}$$

siendo σ una cadena de símbolos Terminales.

2.4.1 Representación de gramáticas

Para la representación de una gramática utilizaremos la BNF. La BNF (Backus Normal Form, Backus-Naur Form, en homenaje a Backus, su creador y a Naur, su continuador), es un metalenguaje muy utilizado para definir la estructura sintáctica de lenguajes de programación (lenguajes formales). La forma de Backus-Naur fue creada para definir la escritura sintáctica del lenguaje de programación ALGOL60. Las notaciones BNF, reducen el número de reglas necesarias. Para ello, utilizan los siguientes metasímbolos:

- La barra disyuntiva '|': unifica en una, dos reglas con el mismo símbolo No Terminal a la izquierda del igual. Las reglas A = aA y A = bB, se van a convertir en la regla A = aA | bB.
- El paréntesis de opcionalidad '(...)': dos reglas iguales, salvo una expresión inserta, equivalen a la mayor de ellas con la expresión inserta entre paréntesis. Las reglas A = aB y A = a se convierten en A = a(B).
- El signo más de recursividad '+': adjunto a una expresión, equivale a las reglas A = ...A y A = A... Por ejemplo: A = aA y A = a se transformará en A = a+.
- El asterisco '*': equivale a una expresión con más y entre paréntesis. Así, A* es lo mismo que (A+).
- Los corchetes '[]': para alterar la prioridad en la interpretación de los metasímbolos. Como '+' y '*' tienen más prioridad que la barra '|', son expresiones diferentes: A | B+ y [A | B]+.

Traduciendo esta notación infija a prefija, se gana en facilidad de procesamiento pero no en facilidad de escritura (hay que escribir más).

NotaciónBNF → Expresión
Expresión → Término | Expresión
Expresión → Término
Término1 → Paréntesis
Término1 → Término *
Término1 → Término +
Término1 → Término
Término → Símbolo
Término → Corchete
Corchete → [Expresión]
Paréntesis → (Expresión)

Cuadro 2.1. Gramática de la notación BNF.

2.5 JERARQUÍA DE CHOMSKY

A fin de precisar más qué tipo de gramática es capaz de generar un lenguaje lo más parecido posible a los naturales, Chomsky clasificó las gramáticas y lenguajes dentro de cuatro familias jerárquicamente ordenadas como modelos potenciales del lenguaje natural.

Esta clasificación, conocida como jerarquía de Chomsky, se establece aumentando las restricciones sobre la forma de las producciones. Así pues, tenemos:

Gramática sin restricciones	Tipo 0
Gramáticas sensitivas al contexto	Tipo 1
Gramáticas libres de contexto	Tipo 2
Gramáticas regulares	Tipo 3

Las restricciones colocadas en las reglas, aumenta con el número de la gramática.

2.5.1 Tipo 0

Las gramáticas de tipo 0, son gramáticas sin restricciones, es decir, no hay restricciones ni para el lado izquierdo, ni para el lado derecho de las producciones.

Su potencia es la de una máquina de Turing, y sus reglas son del tipo:

$$\alpha = \beta$$

No existe algoritmo que en tiempo finito nos diga si una cadena obedece o no las reglas de una gramática de reescritura tan generalizada.

Cuando Chomsky formuló sus objeciones a las gramáticas de estructura de sintagma (tipo 2), propuso la utilización de reglas de tipo 0 para el reordenamiento, elisión, etc., de elementos. La no existencia de algoritmo de parsing, mostraba que su potencia superaba en mucho a las lenguas naturales, y el formato de las reglas de reescritura se limitó mucho (los clásicos movimientos de sintagmas por tematización, interrogación...).

Una gramática sin restricciones, es una cuaterna de la forma (V, $\sum$, P, S), donde V es un conjunto de variables o no terminales, $\sum$ (el alfabeto) es un conjunto finito de símbolos terminales, P es un conjunto finito de reglas, y S es un elemento de V llamado el símbolo inicial o axioma de la gramática. Una producción de una gramática de este tipo, tiene la forma que ya hemos visto anteriormente ($\alpha = \beta$), donde $\alpha \in (V \cup \sum)^+$ y donde $\beta \in (V \cup \sum)^*$. Los conjuntos V y $\sum$ son disjuntos.

Escribir un analizador sintáctico para una gramática de tipo 0, sería una tarea muy ardua.

2.5.2 Tipo 1

Contiene reglas que se ajustan a:

$$\beta A \gamma = \beta \delta \gamma$$

Este tipo de producciones, implica que las sustituciones sólo pueden efectuarse en cierto contexto (el símbolo 'A' se podrá sustituir por 'δ' si y sólo si, está precedido por 'β', y le sigue 'γ'), esto es, son gramáticas sensibles al contexto, por tanto, pueden hacer que un sintagma sea sistemáticamente igual a otro. Obsérvese en el cuadro número 2.2 lo complejo de las reglas necesarias.

La complejidad de su parsing es exponencial con la longitud de la cadena de entrada (lo cual es inaceptable con fines de reconocimiento).

Este tipo de gramáticas son las de menor éxito en toda la jerarquía.

S → a S B C
S → a b C
b B → b b
b C → b c
c C → c c
C B → C D
C D → E D
E D → E C

Cuadro 2.2. Gramática sensible al contexto.

2.5.3 Tipo 2

Describen los llamados lenguajes de contexto libre, es decir, en ellos se pueden insertar proposiciones dentro de proposiciones, independientemente del contexto de la oración. De acuerdo con Chomsky, una gramática y su lenguaje correspondiente son independientes del contexto, si y sólo si pueden definirse con un conjunto de producciones independientes del contexto. Las gramáticas de contexto libre (o independientes del contexto), son muy importantes en la teoría de los lenguajes de programación, ya que los lenguajes que definen, tienen en general, una estructura muy sencilla. Las técnicas de análisis sintáctico, suelen basarse en gramáticas de contexto libre.

El formato de las producciones según Chomsky será:

$$A = \alpha$$

donde α es una cadena, vacía o no, de símbolos terminales o no terminales.

Equivale en cuanto a potencia descriptiva, al autómata con pila o Pushdown, y nos permite describir adecuadamente las relaciones intra e inter-sintagmáticas de la lengua natural:

- Concordancia sujeto-verbo.
- Concordancia sujeto-atributo.
- Inserción de proposiciones en posibles centrales.
- Etc.

La representación gráfica de un análisis de contexto libre es el típico árbol sintáctico, con los símbolos no terminales en los nodos intermedios, y los símbolos terminales en los nodos finales. Se muestra un ejemplo en la ilustración número 2.1.

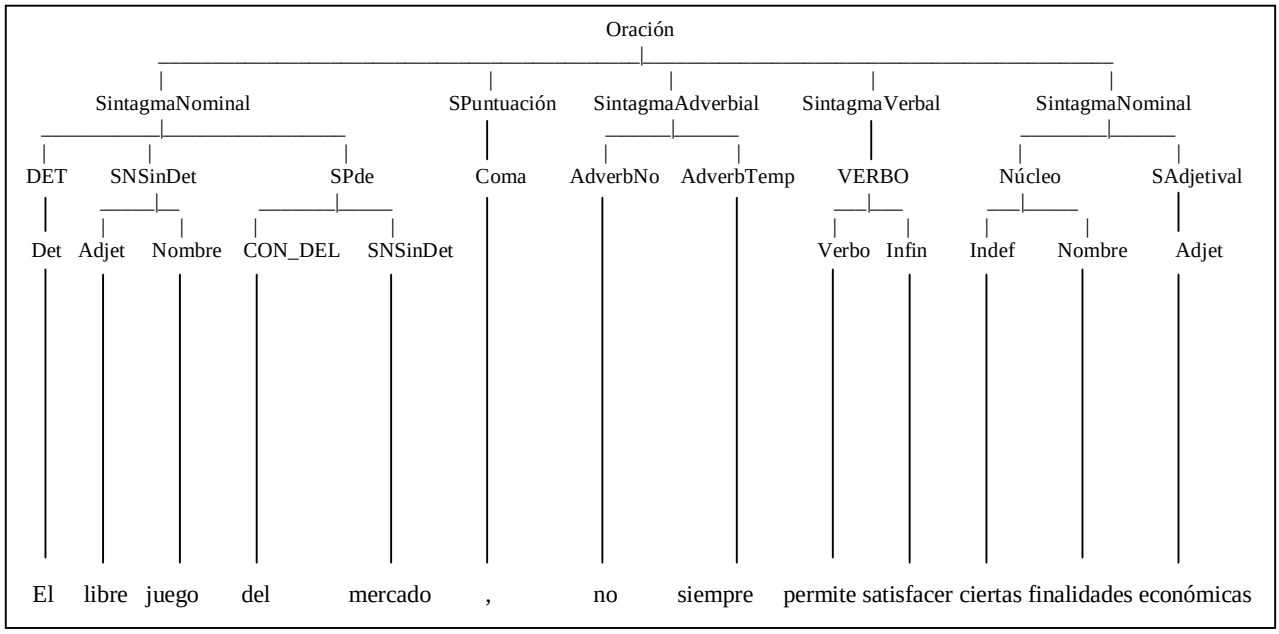

Ilustración 2.1. Ejemplo de árbol sintáctico de una frase en castellano.

2.5.3.1 Limitaciones de los lenguajes de contexto libre

Se ha debatido mucho sobre la posibilidad o imposibilidad de la descripción de lenguajes naturales por medio de gramáticas de contexto libre, basándose desde ejemplos sencillos aunque muy discutibles, hasta referencias a estructuras presentes en lenguas como el holandés o el bámbara (Perrault[84], Edinburg[89]).

Hay quien objeta que la multiplicación de categorías sintácticas necesaria para contemplar las concordancias, es claramente antinatural, que sería mejor que los símbolos llevaran asociados rasgos como el género, el número... Parece poco natural, la creación de categorías sintácticas como: Sintagma_Nominal_Masculino_Plural_Propio_Geográfico, Sintagma_Adjetival_Largo_Con_Sintagma_Adverbial_Antepuesto...; o incluso de categorías léxicas como: sustantivo_masculino_plural_propio_geográfico, adjetivo_femenino_singular_numeral...; sin embargo, nuestro intento como ingenieros, es modelar la lengua externamente de un modo sencillo. La multiplicación de reglas y símbolos, supondría una gramática poco elegante como la del cuadro número 2.3.

Los ejemplos del alemán de Suiza, donde existen construcciones del tipo:

$$a^n b^m c^n d^m$$

o de la lengua Bámbara de Senegal, que contiene sintagmas de estructura:

$$w\text{-}o\text{-}w$$

son dependientes de contexto, pero no afectan al castellano, el cual es nuestro objetivo.

Un ejemplo de gramática que incluyera la construcción expuesta anteriormente en el ejemplo del alemán de Suiza, es la siguiente:

La gramática está definida por: $\sum$ = {a, b, c}, N = {S, A, B, C, D, E}, con las siguientes producciones o reglas:

S = aAB, S = aB, A = aAC, A = aC, B = Dc,
D = b, CD = CE, CE = DE, DE = DC, Cc = Dcc.

Esta gramática será sensible al contexto. Por ejemplo, la composición CE = DE, indica que es posible reemplazar C por D, si C está seguido de E, así mismo, la composición Cc = Dcc, dice que se puede reemplazar C por Dc, si C está seguido de c.

Es posible derivar DC de CD ya que:

CD = CE = De = DC

Un arreglo del tipo del alemán de Suiza, como puede ser por ejemplo $a^3b^3c^3$, pertenecerá al lenguaje definido por esta gramática sensible al contexto, pues:

S=aAB=aaACB=aaaCCDc=aaaDCCc=aaaDCDcc=aaaDDCcc=aaaDDDccc=aaabbbccc.

Proposición = SNSujetoMasculinoPlural SintagmaVerbalPlural

SNSujetoMasculinoPlural = (DeterminanteMasculinoPlural)
(SintagmaAdjetivalPrevioMasculinoPlural) NombreMasculinoPlural
(SintagmaAdjetivalPrevioMasculinoPlural) (SintagmaPreposicional)
(OracionDeRelativo)

SintagmaAdjetivalPrevioMasculinoPlural =
(SintagmaAdverbialPrevioDeGrado)* AdjetivoPrevioMasculinoSingular

Cuadro 2.3. Ejemplo de reglas de una gramática con concordancias.

Siguiendo con la discusión acerca de la posibilidad de tratar lenguajes naturales mediante gramáticas de contexto libre, se ha argumentado, que en oraciones como:

"Daniel y Natalia vendieron un piso y un yate respectivamente"

hay un fenómeno que no es elegante describir mediante gramáticas de contexto libre, ya que requerirían categorías similares a las del ejemplo del cuadro número 2.2, como por ejemplo SintagmaNominalCon2Elementos... Sin embargo, la gramaticalidad parece más ligada a la semántica de "respectivamente", más fácil de verificar por el componente semántico:

"Sus dos hermanos heredaron el piso y el yate respectivamente"

En castellano, el ordenamiento libre de sintagmas dará lugar a constituyentes discontinuos, cuyos árboles sintácticos deberían tener ramas cruzadas. Dichos árboles, no son posibles si usamos gramática de contexto libre, pero cabe la posibilidad de realizar un análisis sintagmático (más que sintáctico) laso, que luego la semántica reordenará, ligará... (Colás[99]). Los hipérbaton, por ejemplo, suelen requerir del oyente, una cierta transformación, es decir, un posproceso que no todos los hablantes son capaces de realizar con corrección, esto depende mucho de su cultura, especialmente de la literaria si nos enfrentamos a enunciados de corte poético, por ejemplo, en inglés, son muy comunes las oraciones de relativo con la preposición al final en lugar de al principio. Un ejemplo de este caso sería:

"Who do you want to speak with?"

La partícula "with" (preposición) introduce a la frase, pero esta frase está por delante de la mencionada partícula. Estos casos serán bastante difíciles de tratar correctamente.

2.5.4 Tipo 3

Es el descriptivamente más débil. Sus reglas poseen el siguiente formato:

$$A = aB$$

donde B puede existir o no.

Equivalen en cuanto a poder descriptivo, a los autómatas finitos deterministas y no deterministas. Dado un lenguaje regular, que es aquel que es posible caracterizar usando una gramática de tipo 3, siempre será posible hallar su autómata equivalente. Cada símbolo no terminal es un estado, cada regla A = aB, una rama que conecta los estados A y B por medio del símbolo terminal a, y cada regla A = a, nos dice que A se une al estado final del símbolo a.

Dada la sencillez estructural de estos lenguajes, la notación chomskiana resulta pesada, por lo que utilizaremos una más compacta para la expresión de lenguajes regulares como es la BNF, ya explicada anteriormente.

Un ejemplo de gramática regular para el sintagma nominal del castellano, se muestra en el cuadro número 2.4, donde se ha permitido la utilización de símbolos auxiliares no recursivos, con lo cual, nunca podremos exceder los límites de los lenguajes regulares.

2.5.4.1 Limitaciones de los lenguajes regulares

Un autómata categorial cuidadosamente diseñado y compilado, puede reconocer un amplio subconjunto de un lenguaje natural como puede ser el castellano. En

Subirats[91], se encuentra un texto periodístico completo, aceptado por un autómata similar al nuestro, con frases tan espectaculares como simples:

"La importante reducción en la remuneración de los bonos del Tesoro hasta el 14,505 por ciento en tasa interna bruta de rentabilidad supone el inicio de un descenso progresivo de los tipos de los títulos público y adelantará al otoño la reducción de los tipos de interés".

Esta adecuación observacional, no oculta la incapacidad de los lenguajes regulares para, con sencillez y elegancia, decirnos cuál es la estructura de este sintagma nominal (a qué nucleo nominal o adjetival están complementando cada uno de los sintagmas preposicionales, cómo se encadenan éstos entre sí...), ni explicar ambigüedades estructurales que pueden aparecer en ciertas oraciones. Esta capacidad de reconocer cadenas, pero no dar correctamente su estructura, supone un caso de capacidad generativa débil y no fuerte (Edimburg[89], cap.1).

Sintagma_Nominal = (Determinante) (Numeral) (Sintagma_Adjetival)
Nucleo_Nominal (Sintagma_Adjetival) Sintagma_Preposicional*

Determinante = PreArticulo [Articulo | Posesivo | Demostrativo]
Numeral = Cardinal | Ordinal
Nucleo_Nominal = Sustantivo | Nombre_Propio | Infinitivo

Sintagma_Adjetival = Adverbio* [Adjetivo | Participio]+
([, Adverbio* [Adjetivo | Participio]]* [y | o] Adverbio* [Adjetivo | Participio])

Sintagma_Preposicional = Preposicion (Determinante) (Numeral)
(Sintagma_Adjetival) Nucleo_Nominal (Sintagma_Adjetival)

Cuadro 2.4. Posible gramática del sintagma nominal en castellano.

En Edimburg[89], cap.2, se incluye un ejemplo de Gazdar y Pullam contra el carácter regular del inglés:

"A white male (whom a white male)^n hired^n hired another white male"

que traducido al castellano, queda aproximadamente:

"Un hombre blanco (a quien un hombre blanco)^n alquiló^n alquiló un hombre blanco"

Aunque el ejemplo es extremo (una gramática que falle en esa frase, podría no ser rechazada en la práctica), pretende mostrar que la inserción generalizada y recursiva de proposiciones en posición relativa y no absoluta, con sus parejas de sujetos y predicados, constituye un fenómeno lingüístico no elegantemente regular. Cuando todos escuchamos la oración:

"El coche que el chófer que Pedro contrató condujo ayer parece rápido"

Mentalmente asignamos sujeto a los distintos verbos, percibiendo la estructura de niveles que supone la inserción de una proposición dentro de otra. Cada verbo, no tiene al sustantivo anterior más cercano como sujeto, sino al anterior más cercano que no tenga un verbo más cercano aún. La incapacidad para aceptar un elevado grado de recursividad de inserción, parece más ligada con nuestras capacidades para el lenguaje hablado (inferiores a las que tenemos al leer o escribir), que al posible carácter regular del lenguaje natural.

Trucos como la adición de una pila para la subordinación inserta, o la utilización de gramáticas ambiguas y desdoblamiento del análisis al pasar por un estado de ambigüedad, o el etiquetado de los sintagmas al llegar a determinados estados, complican en exceso el sencillo funcionamiento de los autómatas, y son formas poco recomendables de convertir el lenguaje aceptado en algo más que regular.

Desde el punto de vista de algunos autores, las gramáticas de tipo 2 y 3 son las más importantes. Mientras las gramáticas libres del contexto, definen la sintaxis de las declaraciones: las proposiciones, las expresiones, etc. (es decir, por hacer un símil con los lenguajes de programación, la estructura de un programa), las gramáticas regulares, definen en cambio, la sintaxis de los identificadores, número, cadenas y otros símbolos básicos del lenguaje. Por ello, es común encontrar gramáticas libres del contexto en el análisis sintáctico, a la vez que las gramáticas regulares se emplean como la base del análisis léxico.

Para resumir el apartado que se acaba de exponer, tenemos la siguiente tabla.

Gramáticas	**Lenguajes**	**Máquinas**
Tipo 0, Gramáticas estructura-frase Gramáticas sin restricciones	Lenguajes recursivamente enumerables	Máquinas de Turing, Máquinas de Turing no deterministas
Tipo 1, Gramáticas sensitivas al contexto Gramáticas monótonas	Lenguajes sensitivos al contexto	Autómata linear-bounded
Tipo2, Gramáticas libres de contexto	Lenguajes libres de contexto	Autómata Pushdown
Tipo3, Gramáticas regulares Gramáticas lineales hacia la izquierda Gramáticas lineales hacia la derecha	Lenguajes regulares	Autómata finito determinista, Autómata finito no determinista

Tabla 2.1. Jerarquía de Chomsky.