INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

TESIS

IMPLEMENTACIÓN DE UN ANALIZADOR SINTÁCTICO DEL IDIOMA ESPAÑOL PARA UNA INTERFAZ DE LENGUAJE NATURAL A BASES DE DATOS

Para obtener el grado de:

Maestro en Ciencias de la Computación

Presenta:

I.S.T.I. Oscar Manuel Mellado Camacho

Director de tesis:

Dr. Rodolfo A. Pazos Rangel

Codirector de tesis:

Dr. José Antonio Martínez Flores

Ciudad Madero, Tamaulipas Mayo 2014

Cd. Madero, Tamps; a 24 de Febrero de 2014.

OFICIO No.: U5.051/14
AREA: DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

ING. OSCAR MANUEL MELLADO CAMACHO
NO. DE CONTROL G12072007
P R E S E N T E

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su examen de grado de Maestría en Ciencias de la Computación, el cual está integrado por los siguientes catedráticos:

PRESIDENTE :	DRA. GUADALUPE CASTILLA VALDEZ
SECRETARIO :	DRA. CLAUDIA GUADALUPE GÓMEZ SANTILLÁN
VOCAL :	DR. RODOLFO ABRAHAM PAZOS RANGEL
SUPLENTE	M.C. JOSÉ APOLINAR RAMÍREZ SALDIVAR
DIRECTOR DE TESIS :	DR. RODOLFO ABRAHAM PAZOS RANGEL
CO-DIRECTOR:	DR. JOSÉ ANTONIO MARTÍNEZ FLORES

Se acordó autorizar la impresión de su tesis titulada:

"IMPLEMENTACIÓN DE UN ANALIZADOR SINTÁCTICO DEL IDIOMA ESPAÑOL PARA UNA INTERFAZ DE LENGUAJE NATURAL A BASES DE DATOS"

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con Usted el logro de esta meta.

Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

A T E N T A M E N T E
"Por mi patria y por mi bien"

Ma. Yolanda Chávez Cinco

M. P. MARIA YOLANDA CHÁVEZ CINCO
JEFA DE LA DIVISIÓN

S. E. P.
DIVISION DE ESTUDIOS
DE POSGRADO E
INVESTIGACION
I T C M

c.c.p.- Archivo
Minuta

MYCHC 'NLCO ' jar

DECLARACIÓN DE ORIGINALIDAD

Declaro y prometo que este documento de tesis es producto de mi trabajo original y que no infringe los derechos de terceros, tales como derechos de publicación, derechos de autor, patentes y similares.

Además, declaro que las citas textuales que he incluido las cuales aparecen entre comillas y en los resúmenes que he realizado de publicaciones agenas, indico explícitamente los datos de los autores y publicaciones.

Además, en caso de infracción a los derechos de terceros derivados de este documento de tesis, acepto la responsabilidad de la infracción y relevo de esta a mi director y codirector de tesis, así como al Instituto Tecnológico de Ciudad Madero y sus autoridades.

I.S.T.I. Oscar Manuel Mellado Camacho

DEDICATORIA

A mi madre Ma. Elena Camacho González por la confianza y apoyo brindado en las decisiones que he tomado a largo de estos años.

A mi tía Leticia Mellado Mar por el cariño y motivación brindada que me ha inspirado en alcanzar las metas que me he propuesto.

AGRADECIMIENTOS

A los miembros de mi comité tutorial de la tesis: Dr. Rodolfo Abraham Pazos Rangel, Dr. José Antonio Martínez Flores, Dra. Claudia Guadalupe Gómez Santillán, Dra. Guadalupe Castilla Valdez, Dra. María Lucila Morales Rodríguez y M.C. José Apolinar Ramírez Saldivar. Por sus críticas constructivas, observaciones y sugerencias, que fueron de ayuda en el desarrollo de esta tesis.

En especial al Dr. Rodolfo Abraham Pazos Rangel por todo el apoyo brindado en este proyecto de tesis. Por encaminarme y fomentarme a ser una persona proactiva a lo largo del desarrollo de este proyecto.

Al Dr. Jesús David Terán Villanueva por sus enseñanzas y ayuda, en especial por alentarme a realizar mis estudios en este posgrado.

A mis compañeros de generación Miguel Ángel Ramiro Zúñiga y Andrés Bautista Alvarado.

A todas las personas que conocí en el Instituto Tecnológico de Ciudad Madero a lo largo de estos dos años, en especial a Eduardo Rodríguez del Angel, Yazmin Gómez Rojas, Fanny Gabriela Maldonado Nava, Rafael Ortega Cortez, Javier Alberto Rangel González.

RESUMEN

El problema del análisis sintáctico consiste en que, dada la entrada de una oración en español, el analizador sintáctico debe ser capaz de etiquetar cada una de las palabras que forman parte de dicha oración, dando como resultado las estructuras sintácticas apropiadas para que un analizador semántico tenga recursos suficientes para efectuar el análisis.

Particularmente el español es uno de los lenguajes con mayor complejidad debido al alto grado de libertad para formar oraciones; por lo tanto, se propone el desarrollo de este analizador sintáctico, el cual, apegándose a las reglas gramaticales de la Real Academia Española, busca arrojar resultados confiables en base a una nueva implementación de la gramática categorial para el análisis sintáctico.

Este proyecto de tesis se enfoca en el diseño e implementación no sólo del analizador sintáctico, sino también en un conjunto de características gramaticales del lenguaje español, las cuales son aprovechadas con el fin de mejorar los resultados esperados.

CONTENIDO

ÍNDICE DE TABLAS

Capítulo 2

Capítulo 4

Capítulo 5

ANEXO C

ÍNDICE DE FIGURAS

Capítulo 4

Capítulo 5

Capítulo 6

ANEXO A

ANEXO B

CAPÍTULO 1: INTRODUCCIÓN

En estas últimas décadas el uso de la computadora ha tenido un gran crecimiento en los países primermundistas, y se han invertido enormes cantidades de recursos para producir poderosas herramientas que procesen grandes volúmenes de información de manera rápida mientras proporcionan soporte a sus lenguas nativas; sin embargo, éste no ha sido el caso para el idioma español.

El análisis de una oración en el idioma español ha resultado ser muy complejo debido a la flexibilidad del mismo. Para expresar una misma idea se pueden utilizar una gran variedad de combinaciones de palabras. Según los expertos, transformar los conocimientos lingüísticos generales a una lingüística computacional es el aspecto más importante a lograr para realizar un avance significativo en este campo. Debido a esto, el desarrollo de un analizador sintáctico que sirva de complemento a una interfaz de lenguaje natural es un nicho de oportunidad. Con la implementación de un analizador sintáctico, se pretende que una interfaz de lenguaje natural sea capaz de mejorar la traducción de consultas de lenguaje natural a lenguaje de consulta estructurado (SQL).

Este trabajo de tesis muestra cómo nuevos métodos de análisis gramatical, en conjunto con un compendio de reglas gramaticales obtenidas del estudio de la lengua española, son capaces de obtener resultados confiables superiores al 95% para los grupos de ejemplares sujetos a experimentación.

Para ello, en este documento de tesis se muestra cómo fueron diseñadas las reglas sintácticas, cómo fueron diseñadas e implementadas las estructuras de datos usadas, el funcionamiento de los algoritmos que permiten obtener dichos resultados para los diferentes grupos de ejemplares.

1.1 Objetivo general

Diseñar e implementar un analizador sintáctico del idioma español para una interfaz de lenguaje natural a bases de datos.

1.2 Objetivos particulares

- Decidir la gramática usada para el desarrollo del analizador sintáctico.
- Generar un modelo dinámico de inclusión de reglas, donde se permita incrementar el número de reglas sintácticas con las que el analizador trabaje.
- El analizador debe procesar oraciones con al menos dos verbos.
- El analizador debe enfocarse a procesar las oraciones que se encuentran en las consultas a bases de datos.

1.3 Justificación

El problema más frecuente al que se enfrentan los usuarios al realizar búsquedas en bases de datos relacionales es la necesidad de conocer el lenguaje SQL para poder extraer la información. La solución para esta problemática es la creación de una interfaz de lenguaje natural orientada a consultas de bases de datos relacionales.

Los problemas más frecuentes a los que se enfrentan los desarrolladores de una interfaz de lenguaje natural son [1]:

- Alcanzar independencia de dominio.

- Uso de palabras o frases de diferentes categorías sintácticas.
- Elipsis semántica y ambigüedades a nivel léxico, sintáctico y semántico.

Para traducir una oración en lenguaje natural a una sentencia en SQL, es necesario analizarla sintácticamente para obtener información detallada, la cual genera información útil para interpretar el significado de la oración.

El análisis sintáctico apropiado da como resultado un árbol sintáctico legible y útil para la interfaz de lenguaje natural, el cual incrementa las probabilidades de obtener un mejor resultado. Por lo anterior, surge la necesidad de implementar un analizador sintáctico que sea capaz de identificar la o las estructuras sintácticas que están inmersas en una consulta para su procesamiento en una interfaz de lenguaje natural.

1.4 Antecedentes del proyecto

“Analizador sintáctico de oraciones en español usando el método de dependencias” por José Antonio Cervantes Alvarez en 2005, en donde se implementó un analizador sintáctico flexible y dinámico, el cual permitía que el sistema pudiera incrementar su conocimiento lingüístico realizando mínimas inclusiones de datos. Gracias a que el sistema es flexible y dinámico, cualquier usuario experto en procesamiento de lenguaje natural puede agregar o modificar la información sintáctica con la que actualmente cuenta sin la necesidad de editar el código fuente (Java).

En este trabajo una de las mejoras más significativas es el diseño de la base de datos del conocimiento lingüístico, dicha mejora permite que un verbo pueda

tener los mismos patrones que otro verbo y además permite que los patrones de un mismo verbo sean diferentes, dependiendo del tiempo en el que se encuentre conjugado.

Otro aspecto que contempla el diseño de la base de datos es la existencia de verbos con polisemia, los cuales pueden manejar patrones distintos dependiendo del contexto en el que se usen. Sin embargo, a pesar de las características implementadas en este proyecto no se pudo eliminar la ambigüedad sintáctica que se presentó en algunas oraciones de los casos de prueba [2].

“Modelo Sistemáticamente Enriquecido de Bases de Datos para su Explotación por Interfaces de Lenguaje Natural” por Marco Antonio Aguirre Lam (en proceso), el cual tiene como fin desarrollar una interfaz de lenguaje natural encargada de traducir consultas de lenguaje natural a SQL. A pesar de ser la tercera versión, dicha interfaz no cuenta con un analizador sintáctico, por lo que este proyecto de tesis busca mejorar su desempeño con la eventual inclusión del analizador sintáctico.

La complejidad de las consultas que se manejan en dicho proyecto es tal que se tienen dificultades para alcanzar resultados exitosos por encima del 90% [1].

Por otro lado, el diccionario de datos usado en el proyecto de Aguirre, además de incluir la información estructural de la base de datos, se ve enriquecido con suficiente información semántica para facilitar el proceso de traducción de lenguaje natural a SQL [1].

Uno de los problemas que busca resolver la interfaz de lenguaje natural que presenta Aguirre con la incorporación de un analizador sintáctico, es eliminar la

ambigüedad que se deriva por el uso de palabras que pueden tomar el papel de verbos y sustantivos dependiendo del contexto en el que se usen.

1.5 Alcances y limitaciones

Entre los alcances de este proyecto se encuentran los siguientes:

- El análisis de las oraciones debe ser en lenguaje natural, determinando si la oración posee alguna de las posibles formas gramaticales permitidas para la formación de oraciones en español. El analizador debe trabajar con oraciones con al menos dos verbos.
- El analizador sintáctico debe ser dinámico, se puede incrementar o modificar las reglas gramaticales, es decir, las reglas deben estar en un módulo independiente del código.

Las limitaciones de este proyecto son las siguientes:

- El analizador sintáctico debe trabajar sólo con una parte del español, debido a la diversidad de palabras que lo conforman.
- Dentro de las pruebas a realizar, el analizador debe enfocarse a procesar las oraciones que se encuentran en las consultas a bases de datos. Debido a que el objetivo de esta tesis es crear un analizador sintáctico para mejorar el desempeño de una interfaz de lenguaje natural a bases de datos.
- Este analizador no tiene la finalidad de funcionar como un corrector de ortografía; si se recibe una palabra incorrecta, el analizador podrá tener salidas no deseadas.
- El analizador sintáctico sólo debe trabajar con el idioma español de México.
- No se busca eliminar por completo la ambigüedad sintáctica.
- No se pretende superar el porcentaje de acierto con respecto a los mejores analizadores sintácticos.

- No se busca que el analizador sintáctico sea capaz de eliminar la anáfora gramatical en las oraciones.
- No se busca que el analizador sintáctico sea capaz de eliminar la elipsis sintáctica.
- No existe el propósito de que el analizador sintáctico sea integrado a una interfaz de lenguaje natural como parte de este proyecto de tesis.

CAPÍTULO 2: ESTADO DEL ARTE

El análisis sintáctico es una parte del proceso de traducción que clarifica la comprensión de las consultas. Los analizadores sintácticos probados para algunos lenguajes tales como el inglés, portugués, francés, danés y noruego han alcanzado altos niveles de precisión y robustez [3].

Desafortunadamente debido a que el idioma español tiene una amplia variedad de reglas gramaticales, el desempeño de los analizadores sintácticos es menor, aproximadamente entre el 70% y 90% de precisión y en el mejor de los casos hasta un 96% dependiendo de los casos de prueba [4].

En la Tabla 2.1 se resumen algunos de los trabajos relacionados que se han realizado:

Tabla 2.1 Resumen de analizadores sintácticos

Trabajo	Características
Galicia, 2000 [5]	• Compendio de reglas sintácticas separado del código fuente. • Uso de la gramática MTT. • Cuenta con aproximadamente 150 reglas gramaticales. • Método de desambiguación por medio de pesos. • Trabajó con oraciones reales además del compendio LEXESP usado para las pruebas. • 53/100 oraciones resueltas con un promedio de colocación del 25 %.
Galicia, 2002 [6]	• Se presenta la aplicación de la teoría Significado-Texto para el análisis sintáctico del español. • Estudio basado en el enfoque de dependencias. • Inclusión no sólo del valor valencial sintáctico del verbo, sino también de su relación semántica. • Se descubrieron patrones combinatorios en la semántica gracias al diccionario usado.
Losada, 2003 [7]	• Elabora una tesis que desarrolla una solución para la automatización del análisis sintáctico. • Aportación de métodos de desambiguación funcional y estructural.

	• DeFuSE (Desambiguador Funcional de Sentencias en Español) se enfoca a la desambiguación que requiere una ILN en niveles superiores. • AMoSinE (Analizador Morfosintáctico del Español) desarrolla estructuras sintácticas en base a reglas gramaticales.
Cervantes, 2005 [2]	• Compendio de reglas sintácticas separado del código fuente. • Uso de gramática de dependencias. • Cuenta con aproximadamente 70 reglas gramaticales. • Conocimiento lingüístico de verbos gracias al diseño de la base de datos. • Enfocado a oraciones interrogativas e imperativas. • Capacidad analítica de oraciones simples (1 verbo).
Bengoetxa, 2007 [8]	• Modificación de un analizador sintáctico determinista propuesto por Nivre en el 2007. • De 4 posibles opciones realizables, determina la correcta por medio de aprendizaje automático para generar un único análisis sintáctico. • Se compara contra LAS (Labeled Attachment Score) que cuenta con un 74.41% de acierto y consigue superarlo por 2.53%.
Carrera, 2008 [9]	• Realiza estudios sobre gramáticas para mejorar las aplicaciones de procesamiento de lenguaje natural. • Hace uso del entorno FreeLing para evaluar resultados. • Toma en consideración las gramáticas de los lenguajes: castellano, catalán, inglés y vasco. • Encuentra que, a pesar de haber similitudes gramaticales, los resultados no son iguales para todas las gramáticas.
Comelles, 2010 [10]	• Realiza estudios contrastativos entre analizadores sintácticos con enfoques de constituyentes y de dependencias. • Los analizadores sintácticos con enfoque de constituyentes (como Charniak y Collins) muestran resultados de entre el 80% y 89 %. • Los analizadores sintácticos con enfoque de dependencias (como Stanford, DeSR, RASP, MINIPAR, MALT) muestran resultados de entre el 45% y 67%.
UPC, 2012 [11]	• No distingue ambigüedades entre palabras que puedan ser usadas como verbos y sustantivos. • Cuenta con un lexicón de más de 100,000 palabras.
TUSIR, 2012 [12]	• Proyecto enfocado al desarrollo de técnicas para resolver problemas que presentan las interfaces de lenguaje natural. • Desarrolla técnicas de análisis y extracción de información. • Se enfoca en estructuras sintácticas y semánticas para resolver problemas.
3LB, 2012 [13]	• Proyecto enfocado a la elaboración de corpus para español, catalán y euskera. • Cuentan con anotación sintáctica y semántica. • Corpus con una lexicón de 100,000 palabras.

Teniendo en cuenta las características de los trabajos antes mencionados, los cuales muestran el nivel de desarrollo que tiene el análisis sintáctico para el lenguaje español, se eligieron aquéllos que comparten características similares a este proyecto con el fin de compararlos.

En la Tabla 2.2 se muestra una comparación de los aspectos relevantes siendo los más destacados:

- El número de reglas gramaticales empleadas

  Las reglas gramaticales son el elemento que le permiten a un analizador sintáctico encontrar las estructuras gramaticales de una oración determinada.

- La cantidad de árboles sintácticos generados

  El número de estructuras sintácticas que puede poseer una oración.

- La realización de pruebas negativas

  Las pruebas realizadas en todos los trabajos de la Tabla 2.2 tratan de encontrar una o varias estructuras sintácticas para la oración proporcionada, la cual está gramaticalmente escrita. Las pruebas negativas se enfocan a agregar, quitar o modificar alguna de las palabras de la oración para que ésta sea gramaticalmente incorrecta. Al realizar una entrada gramaticalmente incorrecta, se espera que el analizador sintáctico no muestre una estructura sintáctica.

Tabla 2.2 Contraste del estado del arte

Trabajo	Gramática empleada	No. de reglas gramaticales	Manejo de oraciones compuestas	Árboles sintácticos generados	Pruebas negativas
Galicia, 2000 [5]	MTT	150	Sí	1	No
Cervantes, 2005 [2]	Dependencias	70	No	Todos*	No
En este trabajo	Gramática categorial	**59**	Sí	**Todos****	**Sí**

*En el trabajo realizado por Cervantes, la cantidad de estructuras sintácticas está determinada por el número de categorías gramaticales que puede tener cada palabra de la oración.

**En este trabajo, la cantidad de árboles sintácticos está determinada, no solamente por el número de categorías gramaticales que puede tener cada palabra de la oración, sino también por la forma de reducir cada una de las configuraciones gramaticales de la oración.

CAPÍTULO 3: TERMINOLOGÍA

Se decidió titular a este capítulo terminología, ya que en él están contenidos ciertos conceptos e información de relevancia para el entendimiento del desarrollo de este proyecto de tesis.

Si se desea conocer acerca de las generalidades del lenguaje, el lenguaje natural, información sobre los analizadores sintácticos, sus enfoques, así como los diferentes tipos de gramáticas que hay, dicha información puede ser encontrada en el ANEXO A: Marco teórico.

Si se desea conocer acerca de las peculiaridades del lenguaje español y entender más acerca de elementos como: categorías gramaticales, la palabra, las frases, los sintagmas y las oraciones, dicha información puede ser encontrada en el ANEXO B: Fundamentos de gramática.

A continuación se presentan los conceptos que son de suma relevancia para capítulos venideros y el desarrollo del proyecto.

3.1 Categoría gramatical

Una categoría gramatical se encuentra definida por la Real Academia Española (RAE) [14] como cada una de las diferentes clases de oficio que desempeña una palabra dentro de una oración.

3.2 Etiqueta gramatical

Con el fin de representar cada una de las categorías gramaticales, así como las estructuras gramaticales como el sintagma, se optó por definir etiquetas gramaticales. Estas etiquetas son formas abreviadas de representar dichas estructuras. Su notación es la siguiente:

- Para las categorías gramaticales se define una etiqueta gramatical haciendo uso de las tres primeras letras de la misma; v. gr., sustantivo – sus, verbo – ver, artículo – art, etc. Dichas etiquetas se escriben en letras minúsculas.

- Para las estructuras gramaticales se define una etiqueta gramatical constituida por mayúsculas y minúsculas; v. gr., frase verbal – FV, sujeto – S, sintagma adjetival – SAdj, cadena vacía – CV. El número de letras empleadas en estas etiquetas puede variar.

Como se aprecia, las etiquetas gramaticales tratan de ser lo más cortas posibles, además que se busca que la etiqueta sea fácil de entender.

Las etiquetas gramaticales usadas en este proyecto se explicarán más a detalle en la Tabla 4.1 y la Tabla 4.2 del CAPÍTULO 4: Analizador sintáctico.

3.3 Identificador numérico

Para trabajar eficientemente con las etiquetas gramaticales, se optó por transformarlas a un identificador numérico, el cual es un número entero cuya función es la de representar una etiqueta gramatical.

Cada una de las etiquetas gramaticales usadas, fue transformada a un identificador numérico. Estos identificadores toman el lugar de las etiquetas gramaticales tanto en los componentes léxicos de una oración como en los elementos que conforman una regla gramatical.

Estos aspectos serán explicados más a detalle en la Tabla 4.3 del CAPÍTULO 4: Analizador sintáctico.

3.4 Gramática

Existen dos formas de ver la gramática: computacional y lingüísticamente hablando. El punto de vista lingüístico define a la gramática como la ciencia que estudia los elementos de una lengua y sus combinaciones. El enfoque computacional define a la gramática como un conjunto finito de reglas que son descritas por una secuencia de símbolos.

Enfoquémonos en el punto de vista computacional por un momento; entonces, Chomsky [15] [16] define algebraicamente a una gramática de la siguiente manera:

$$G = \{N, T, S, P\}$$

donde:

- *N* es un conjunto de elementos no terminales.
- *T* es un conjunto de elementos terminales.
- *S* es un símbolo inicial.
- *P* es un conjunto de reglas de producción.

Ahora, llevemos esto a la gramática vista desde un punto lingüístico. Dado que tenemos un conjunto de elementos terminales representados por las categorías gramaticales, un conjunto de elementos no terminales representados por las estructuras gramaticales como las frases y los sintagmas, un símbolo inicial que está representado por el símbolo: α, el cual puede tomar cualquier valor de la condición de salida y un conjunto de reglas de producción las cuales son todas aquellas reglas gramaticales que hay en el español, contamos con todos los elementos para implementar una gramática [16].

3.5 Consultas

Una consulta está definida por la RAE [17] como: parecer o dictamen que por escrito o de palabra se pide o se da acerca de algo. En base a esta definición, llamamos consulta a toda oración que está contenida en los archivos de prueba, dichas oraciones están estructuradas de manera interrogativa o imperativa y sirven para obtener información de alguna base de datos.

Las consultas que se usarán a lo largo de este proyecto de tesis son consultas para bases de datos en lenguaje natural.

3.6 Variaciones

Una variación está definida por la RAE [17] como: cada uno de los subconjuntos del mismo número de elementos de un conjunto dado, que difieren entre sí por algún elemento o por el orden de éstos.

Dado que se busca obtener todas las posibles estructuras sintácticas que posea una oración, se producen configuraciones de n elementos, siendo n el número de palabras que posee una oración.

Ahora bien, cada una de estas configuraciones difiere con respecto a otra por el valor de un elemento. Esta diferencia se da debido a que una misma palabra puede poseer más de una categoría gramatical; v. gr., la palabra *vuelo* puede ser considerada un verbo ya que es la conjugación del verbo *volar*, y puede ser considerada un sustantivo ya que es el trayecto que recorre un avión de un punto a otro sin realizar escalas.

3.7 Condición de salida esperada

La condición de salida esperada, representada por el símbolo: α, es el resultado que se pretende arroje el analizador sintáctico. Dicha condición consiste de una configuración de etiquetas gramaticales que determina si una variación es sintácticamente correcta o no.

La condición de salida está determinada por los siguientes valores:

1. **SNom** – En este caso la oración no contiene una frase verbal
2. **SVer** – En este caso la oración tiene un sujeto implícito
3. **SNom SVer** – En este caso la oración tiene una estructura básica donde aparecen un sujeto y una frase verbal
4. **SVer SNom** – En este caso la oración tiene los mismos elementos que la estructura anterior sólo que el sujeto aparece después de la frase verbal.

Las cuatro posibles condiciones de salida que se explican en este apartado tienen una estrecha relación con las etiquetas gramaticales de las estructuras que se explican en el capítulo siguiente.

CAPÍTULO 4: ANALIZADOR SINTÁCTICO

En el capítulo anterior, se mencionaron una serie de elementos que son de relevancia para el entendimiento del desarrollo del analizador sintáctico; sin embargo, se debe aclarar que el enfoque de análisis usado en el desarrollo de este analizador sintáctico está basado en una serie de características de la *Gramática Categorial (GC)*.

La idea central de la GC es que una concepción enriquecida de categorías gramaticales puede eliminar la necesidad de muchas de las construcciones que se encuentran en otras teorías gramaticales. Tomando en cuenta esta idea, se realizó un profundo estudio de la gramática española con lo que se logró idear un modelo más genérico de reglas de producción, con el fin de poder analizar un mayor número de consultas usando un menor número de reglas gramaticales.

En este capítulo se explicarán a detalle la estructura y los componentes del analizador sintáctico. Se cubrirán elementos como: las reglas gramaticales de producción, las estructuras de datos usadas para procesar la información y los algoritmos que en conjunto con las reglas gramaticales de producción permiten el análisis sintáctico de las consultas a bases de datos.

4.1 Reglas gramaticales

Habiendo realizado un estudio de la gramática española basado en los conocimientos de la Real Academia Española (RAE), se ideó una forma de construir las reglas gramaticales haciendo uso sólo de estructuras reconocidas por la RAE. Dichas estructuras comprenden las nueve categorías gramaticales conocidas, así como ciertos tipos de frases como lo son los sintagmas y los complementos.

Se optó asignar etiquetas a dichas estructuras gramaticales, las cuales serían usadas como símbolos terminales y no terminales para realizar el análisis sintáctico, dichas etiquetas se encuentran definidas en la Tabla 4.1 y Tabla 4.2.

Tabla 4.1 Etiquetas gramaticales de los símbolos terminales

Símbolo	Significado
art	Artículo
sus	Sustantivo
adj	Adjetivo
pro	Pronombre
ver	Verbo
adv	Adverbio
pre	Preposición
con	Conjunción
int	Interjección
CV	Cadena vacía

Tabla 4.2 Etiquetas gramaticales de los símbolos no terminales

Símbolo	Significado
O	Oración
S	Sujeto
FV	Frase verbal
CD	Complemento directo
CI	Complemento indirecto
CC	Complemento circunstancial
SNom	Sintagma nominal
SAdj	Sintagma adjetival
SAdv	Sintagma adverbial
Spre	Sintagma preposicional
SVer	Sintagma verbal

4.1.1 Construcción de las reglas gramaticales

Una vez realizado el etiquetado de las categorías a usar, se procedió a realizar la construcción de las reglas de producción tomando en cuenta que la GC pretende ser genérica en el uso de sus estructuras.

Para complementar dicha información, se revisó un extracto de la gramática española [18], el cual fue analizado minuciosamente para verificar la correspondencia de información presentada por la RAE [14] [17]. Dichos documentos aportaron información de relevancia para el diseño de las reglas gramaticales de este proyecto.

Se observó que para construir estructuras gramaticales en la gramática española como los sintagmas, se requieren de una a tres categorías gramaticales en

la mayoría de los casos y que en casos donde se requirieran más elementos, dichos elementos son sintagmas construidos a partir de una base de una a tres categorías gramaticales.

Habiendo descubierto esta propiedad, todas las reglas propuestas en este proyecto están construidas de manera tal que:

- tienen un máximo de tres elementos (símbolos terminales o no terminales),
- todos los elementos se reducen a un símbolo no terminal,
- pretenden ser genéricas para analizar un mayor número de consultas, y
- se puede incrementar el número de reglas según sea conveniente.

A continuación se enlistan las reglas creadas para el analizador sintáctico.

4.1.2 Listado de reglas gramaticales

1. SNom = art sus adj
2. SNom = art sus SNom
3. SNom = art sus SAdj
4. SNom = art sus SPre
5. SNom = art sus
6. SNom = sus adj
7. SNom = sus con
8. SNom = sus SNom
9. SNom = sus SAdj
10. SNom = sus SPre
11. SNom = sus
12. SNom = adj sus
13. SNom = adj SNom
14. SAdj = adj SPre
15. SNom = pro sus
16. SNom = pro SNom
17. SNom = pro SAdj
18. SNom = pro SPre
19. SNom = pro
20. SVer = ver art sus
21. SVer = ver adj sus
22. SVer = ver adv

23. SVer = ver SNom
24. SVer = ver SPre
25. SVer = ver
26. SAdj = adv adj SPre
27. SAdv = adv sus
28. SAdj = adv adj
29. SVer = adv ver
30. SAdv = adv SNom
31. SAdv = adv SPre
32. SAdv = adv
33. SPre = pre art sus
34. SPre = pre sus adj
35. SPre = pre adj sus
36. SPre = pre sus
37. SPre = pre adj
38. SPre = pre SNom
39. SPre = pre SAdj
40. SPre = pre SVer
41. SPre = pre SAdv
42. SPre = pre SPre
43. SNom = con art sus
44. SNom = con sus
45. SVer = con ver adv
46. SVer = con ver
47. SVer = con adv ver
48. SPre = con pre
49. SNom = con SNom
50. SNom = SNom SAdj
51. SNom = SNom SPre
52. SAdj = SAdj SNom
53. SVer = SVer SNom SAdj
54. SVer = SVer SNom
55. SVer = SVer SPre
56. SAdv = SAdv SNom
57. SAdv = SAdv SPre
58. SPre = SPre SNom
59. SVer = SPre SVer

4.2 Estructuras de datos

Las oraciones que maneja el analizador sintáctico propuesto, son consultas a bases de datos en lenguaje natural. Las oraciones empleadas en las consultas son de tipo interrogativas o imperativas. Dichas oraciones se encuentran estructuradas con un formato que permite analizar sus elementos léxicos a detalle.

A continuación mostramos la consulta *“Lista el número de pasajeros de cada vuelo*” en la Figura 4.1, la cual muestra los diferentes elementos que constituyen cada una de las consultas usadas en este proyecto.

Cada una de las consultas se encuentra delimitada por el símbolo @ que indica el inicio de una consulta y el símbolo @@ que indica el final de una consulta.

@

8			
Lista el número de pasajeros de cada vuelo			
Lista	ver	sus	adj
el	art		
número	sus		
de	pre		
pasajeros	sus		
de	pre		
cada	pro	adj	
vuelo	ver	sus	

@@

	Número de palabras en la oración
	Oración
	Elementos léxicos
	Categorías gramaticales

Figura 4.2 Estructura de una consulta

En la sección 4.1.2 Listado de reglas gramaticales, se encuentran todas las reglas gramaticales usadas en este proyecto; sin embargo, para trabajar con ellas, se optó por cambiar la mayoría de las etiquetas gramaticales usadas por un identificador numérico como se muestra en la Tabla 4.3.

Tabla 4.3 Identificadores numéricos de las etiquetas gramaticales

Símbolo	Significado	Identificador
CV	Cadena vacía	0
art	Artículo	1
sus	Sustantivo	2
adj	Adjetivo	3
pro	Pronombre	4
ver	Verbo	5
adv	Adverbio	6
pre	Preposición	7
con	Conjunción	8
int	Interjección	9
SNom	Sintagma nominal	10
SAdj	Sintagma adjetival	11
SVer	Sintagma verbal	12
SAdv	Sintagma adverbial	13
SPre	Sintagma preposicional	14
CC	Complemento circunstancial	15
CD	Complemento directo	16
CI	Complemento indirecto	17

Cada uno de los símbolos terminales y no terminales fue remplazado por un identificador numérico, y a su vez todas las reglas gramaticales fueron transformadas en números. Las reglas transformadas a su forma numérica fueron almacenadas en la estructura que se muestra en la Figura 4.3.

Regla gramatical		Índice	Longitud	Reducción	Elementos		
SNom art sus adj	→	0	3	10	1	2	3
SNom art sus SNom	→	1	3	10	1	2	10
SNom art sus SAdj	→	2	3	10	1	2	11
SNom art sus SPre	→	3	3	10	1	2	14
SNom art sus	→	4	2	10	1	2	
⋮		⋮	⋮	⋮	⋮	⋮	⋮
SVer ver art sus	→	19	3	12	5	1	2
SVer ver adj sus	→	20	3	12	5	3	2
SVer ver adv	→	21	2	12	5	6	
SVer ver SNom	→	22	2	12	5	10	
SVer ver SPre	→	23	2	12	5	14	
SVer ver	→	24	1	12	5		
⋮		⋮	⋮	⋮	⋮	⋮	⋮

Figura 4.3 Codificación de las reglas gramaticales

Una vez transformadas las reglas gramaticales de la sección 4.1.2 Listado de reglas gramaticales como muestra la Figura 4.3, se optó por almacenar para cada una de las reglas el primer elemento de la columna *Elementos* y su índice (ver Figura 4.3), con el fin de saber la ubicación de las reglas gramaticales que reducen a un determinado elemento.

El resultado es una estructura que muestra el rango de índices para cada una de las etiquetas gramaticales numeradas como se muestra en la Figura 4.4. Dicha estructura será de utilidad posteriormente al usar el algoritmo de reducción explicado en la sección 4.3.3 Algoritmo de reducción.

Etiqueta gramatical	Etiqueta codificada	Índice inicial	Índice final
Cadena vacía	0	-1	-1
Artículo	1	0	4
Sustantivo	2	5	10
Adjetivo	3	11	13
Pronombre	4	14	18
Verbo	5	19	24
Adverbio	6	25	31
⋮	⋮	⋮	⋮

Figura 4.4 Almacenamiento de índices de búsqueda

Para trabajar con las consultas a bases de datos, se creó una estructura capaz de almacenar todos los componentes léxicos que la conforman. Dicha estructura almacena los siguientes datos como:

- Palabra: Elemento léxico
- Expresión regular: Secuencia de caracteres que fungen como patrón de búsqueda para encontrar patrones en cadenas de texto.
- Categorías gramaticales: Cada una de las categorías gramaticales que posee cada una de las palabras de la oración en su forma numérica.
- Número de categorías gramaticales: El número de categorías gramaticales de cada palabra.

Tomando como ejemplo la consulta de la Figura 4.5 *“Lista el número de pasajeros de cada vuelo”* y analizando cada uno de sus elementos se realiza el siguiente proceso, el cual da como resultado la Tabla 4.3.

El número de palabras nos indica el número de registros que se crearán para la estructura. Se almacena cada una de las palabras, se transforman todas sus categorías gramaticales a sus identificadores numéricos y se almacena el número de categorías gramaticales que tiene cada palabra.

Al final, para cada una de las palabras se utiliza una expresión regular, la cual categoriza cada una de las palabras. Dicho etiquetado no tiene relevancia alguna para el análisis sintáctico; sin embargo, se pretende que esta información adicional sea benéfica para el análisis semántico.

Tabla 4.4 Estructura de la consulta "Lista el número de pasajeros de cada vuelo"

Palabra	Expresion regular	Categorías	No. categorías
Lista	palabra	[5][2][3]	3
el	palabra	[1]	1
número	palabra	[2]	1
de	palabra	[7]	1
pasajeros	palabra	[2]	1
de	palabra	[7]	1
cada	palabra	[4][3]	2
vuelo	palabra	[5][2]	2

Se cuenta con expresiones regulares, las cuales son cadenas de texto con un formato específico capaces de encontrar patrones en un texto [19]. Las etiquetas propuestas son:

- Nombre propio: Todas aquellas cadenas de texto que inician con mayúscula (la primer palabra de una oración, no está contemplada); v. gr., *Pedro, Islandia, México,* etc.
- Siglas: Todas aquellas cadenas de texto que están conformadas por sólo letras mayúsculas; v. gr., *ONU, IFE, ITCM,* etc.

- Números: Todas aquellas cadenas de texto que están conformadas por sólo dígitos; v. gr., *25, 1500, 700,* etc.
- Fechas: Todas aquellas cadenas de texto con el formato; dd/mm/aaaa, siendo los caracteres que las forman números separados por diagonales; v. gr., *02/11/1913, 04/08/2010,* etc.
- Palabras: Todas aquellas cadenas de texto que están conformadas por sólo letras; v. gr., *pasajero, vuelo, isla, cantidad,* etc.
- Texto con números: Todas aquellas cadenas de texto que están conformadas por letras y números; v. gr., *R42, K23U1,* etc.

4.3 Algoritmos

Para explicar el comportamiento del algoritmo, es necesario explicar a detalle un aspecto del lenguaje español que dificulta la implementación del analizador sintáctico.

Dado que la labor de un analizador sintáctico es la de etiquetar cada una de las palabras que forman parte de una oración para obtener las estructuras sintácticas correctas en base a las reglas gramaticales, es correcto afirmar que el número de estructuras sintácticas se basará en la cantidad de reglas gramaticales que tenga la gramática del lenguaje que se está usando.

Dentro del lenguaje, existen una serie de elementos que vuelven más compleja o simplifican la relación entre los conjuntos de palabras que conforman las oraciones. Identificar las posibles combinaciones entre los elementos de las oraciones se vuelve menos complejo cuando, en el lenguaje, dichas posiciones están fijas [5].

Hablando particularmente del español, éste se considera como uno de los lenguajes con mayor complejidad debido al alto grado de libertad con el que se cuenta para formar oraciones; por lo que su análisis sintáctico resulta ser difícil.

Al igual que otros lenguajes, el español cuenta con una estructura sintáctica básica para construir oraciones, la cual se compone de:

- Sujeto (S).
- Verbo (V).
- Complemento (C).

Sin embargo, dicha estructura no siempre se cumple en ese preciso orden de aparición. Se muestra la flexibilidad del lenguaje español con un ejemplo a continuación:

Julieta corre por el parque (SVC).

Variantes:

- Por el parque corre Julieta (CVS).
- Corre Julieta por el parque (VSC).
- Por el parque Julieta corre (CSV).
- Corre por el parque Julieta (VCS).
- Julieta por el parque corre (SCV).

Esto genera un grado de complejidad mayor para el diseño del analizador sintáctico. Debido a esto, el analizador sintáctico debe tomar en consideración esta característica del español, para determinar la estructura sintáctica adecuada para cada oración.

Considerando esta característica del idioma español, fue necesario construir un algoritmo, que haciendo uso de las reglas gramaticales propuestas, fuese capaz de evaluar todas las posibles formas en las que se puede reducir una oración.

El algoritmo se encarga de buscar todas las posibles formas en las que se puede reducir cada una de las variaciones que tiene una oración. Esto es debido a que los elementos que se toman en cuenta para reducir un segmento de la oración no siempre son los mismos.

Tomemos como ejemplo la oración “¿Cuál es el libro más barato de tipo Business?”, como podemos apreciar en la Figura 4.6 y la Figura 4.7, cada uno de los componentes léxicos ha sido etiquetado de la misma forma; sin embargo, la forma en la que se reducen es diferente. En la primera figura, los elementos restantes no pueden ser reducidos a una condición de salida viable, mientras que en la segunda figura sí.

<table>
<tr><td>¿Cuál</td><td>es</td><td>el</td><td>libro</td><td>más</td><td>barato</td><td>de</td><td>tipo</td><td>Business?</td></tr>
<tr><td>pro</td><td>ver</td><td>art</td><td>sus</td><td>adv</td><td>adj</td><td>pre</td><td>sus</td><td>sus</td></tr>
<tr><td>SNom</td><td colspan="3">SVer</td><td colspan="2">SAdj</td><td colspan="2">SPre</td><td>SNom</td></tr>
<tr><td>SNom</td><td colspan="3">SVer</td><td colspan="2">SAdj</td><td colspan="3">SPre</td></tr>
</table>

Figura 4.6 Reducción no posible

¿Cuál	es	el	libro	más	barato	de	tipo	Business?
pro	ver	art	sus	adv	adj	pre	sus	sus
SNom	ver	SNom		SAdj		SPre		SNom
SNom	ver	SNom				SPre		SNom
SNom	SVer					SPre		SNom
SNom	SVer							SNom
SNom	SVer							

Figura 4.7 Reducción posible

Lo anterior demuestra que para afirmar que una oración es o no sintácticamente correcta, se deben analizar todas las posibles formas en las que ésta puede ser reducida. Esto se logra con un algoritmo exhaustivo tal como el que estamos proponiendo en este proyecto.

4.3.1 Algoritmo principal

El algoritmo principal (ver Algoritmo 4.1) consta de cinco procesos, los cuales son:

1. Codificar las reglas a su forma numérica y obtener los índices de búsqueda, es decir, para producir las estructuras mostradas en la sección 4.2 Estructuras de datos (Figura 4.3 y Figura 4.4).
2. Insertar los datos de la consulta en la estructura mostrada en la sección 4.2 Estructuras de datos(Tabla 4.4).
3. Encontrar el número total de variaciones para la consulta y generar dichas variaciones (ver sección 4.3.2 Construcción de variaciones).
4. Copiar cada una de las variaciones en la estructura que será evaluada por el algoritmo de reducción (ver sección 4.3.3 Algoritmo de reducción.
5. Realizar la reducción de cada variación (ver sección 4.3.3 Algoritmo de reducción).

De los procesos antes descritos, el proceso 1 sólo se realiza una sola vez sin importar el número de consultas que contenga el archivo. Mientras que los procesos del 2 al 5 dependen de qué tantas formas de reducción tenga una variación y a su vez, cuántas variaciones pueda tener una consulta.

```
    Algoritmo Principal
1   GenerarReglas()        //1
2   while Oraciones do
3       for i = 0 to palabrasEnOracion do
4           generarListaDePalabras()        //2
5           variacionesTotales *= listaPalabras[i].cantidadCategorías
6       end for
7       generarVariaciones(listaPalabras, variacionesTotales)        //3
8       for j = 0 to variaciones do
9           for k = 0 to variaciones[j].length do
10              arbolProfundidad[0][k] = variaciones[j][k]
11          end for
12          arbolProfundidad[0][listaPalabras.length] = listaPalabras.length        //4
13          reduccionDeReglas(0, 0, 0)        //5
14      end for
15  end while
```

Algoritmo 4.1 Pseudocódigo del algoritmo principal

4.3.2 Construcción de variaciones

El número de variaciones totales que posee una consulta está determinado por la multiplicación de la cardinalidad del conjunto de categorías gramaticales de cada una de las palabras que constituyen la consulta a analizar; es decir, los valores que aparecen en la columna *No. categorías* de la estructura mostrada en la Tabla 4.4, lo que da lugar al total de variaciones posibles. Dicho cálculo se expresa de la manera mostrada en la expresión (4.1).

$$Total\ de\ variaciones\ = \prod_{i=0}^{n} |categorías_i| \qquad (4.1)$$

Por ejemplo, tomando en cuenta la expresión (4.1) y aplicándola en la oración de la Figura 4.8, se tiene como resultado la expresión (4.2).

$$Total\ de\ variaciones = 3 \times 1 \times 1 \times 1 \times 1 \times 1 \times 2 \times 2 = 12 \quad (4.2)$$

Una vez obtenido el número total de posibles variaciones para una determinada oración, se procede a construir las variaciones. En este ejemplo, se construirán todas las variaciones para la consulta mostrada en la Figura 4.2.

Este proceso produce i variaciones, en este caso $i = 12$ (el valor obtenido de la expresión (4.2)) de longitud n, donde n es la cantidad de palabras que contiene una oración, en este caso $n = 8$. En cada variación, cada una de las palabras toma el valor numérico de la posible categoría gramatical que puede poseer, como se muestra en la Tabla 4.4. Esto da como resultado una estructura como la que se muestra en la Figura 4.9.

i	0	1	2	3	4	5	6	7
0	5	1	2	7	2	7	4	5
1	5	1	2	7	2	7	4	2
2	5	1	2	7	2	7	3	5
3	5	1	2	7	2	7	3	2
4	2	1	2	7	2	7	4	5
5	2	1	2	7	2	7	4	2
6	2	1	2	7	2	7	3	5
7	2	1	2	7	2	7	3	2
8	3	1	2	7	2	7	4	5
9	3	1	2	7	2	7	4	2
10	3	1	2	7	2	7	3	5
11	3	1	2	7	2	7	3	2

Figura 4.9 Variaciones construidas para la consulta "Lista el número de pasajeros de cada vuelo"

4.3.3 Algoritmo de reducción

En el Algoritmo 4.2 se describe el funcionamiento del algoritmo de reducción propuesto. El algoritmo que usa es búsqueda exhaustiva, el cual se ejecuta de manera recursiva con el fin de encontrar todas las posibles formas de reducir una oración. Consta de cuatro procesos, los cuales son:

1. Se aplica la regla gramatical a la variación, lo que genera una nueva variación para que sea analizada (ver sección 4.3.3.1 Copiar variación aplicando regla).
2. Con el fin de reducir elementos semejantes, para que el algoritmo analice menos elementos, se aplica una reducción consecutiva explicada en la sección 4.3.3.2 Reducción consecutiva.
3. Se busca una regla contenida en el listado que sea capaz de reducir el segmento a analizar (ver sección 4.3.3.3 Buscar regla).
4. Si se encuentra una regla para reducir el segmento a analizar, se ejecuta de nuevo esta función enviando como parámetros; la regla que se va a aplicar, el índice donde se aplicará la regla y la posición de la variación a la cual se le aplicará dicha regla.

```
    reduccionDeReglas(regla, i, profundidad)
1   if regla != -1
2           aplicarRegla()                      //1
3           reducciónConsecutiva()              //2
4   for regla tamaño k = 3 to 1
5           Buscar regla de tamaño k para todos los elementos de la variación
                    de i = 0 to (elementosEnVariación – k + 1)      //3
6           if regla != -1
7                   reduccionDeReglas(regla, i, profundidad)        //4
8   end for
9   if !reducción
10          Evaluar condición de salida esperada para variación actual
```

Algoritmo 4.2 Pseudocódigo del algoritmo de reducción

4.3.3.1 Copiar variación aplicando regla

La función *aplicarRegla* toma la variación anterior y aplica la regla indicada por la variable *regla* en la posición i de la variación.

Como se muestra en la Figura 4.10 las etiquetas *art* y *sus* son reducidas a la etiqueta *SNom*. Una vez que estas dos etiquetas son remplazadas por esta nueva etiqueta, una variación reducida es generada.

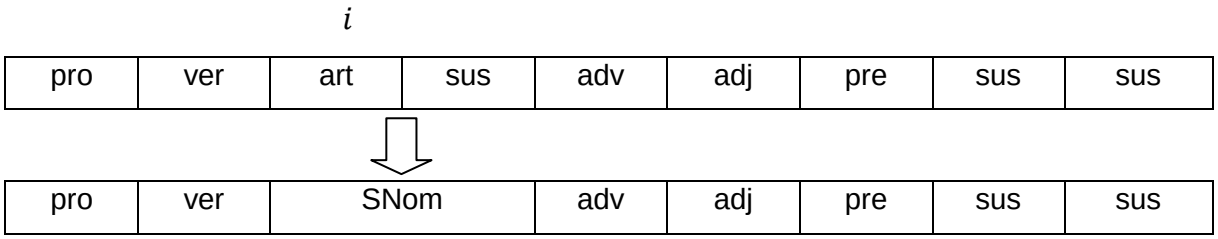

Figura 4.10 Aplicación de regla

4.3.3.2 Reducción consecutiva

Como se menciona en la sección 4.1.1 Construcción de las reglas gramaticales las reglas gramaticales están construidas en base a un esquema de tres elementos como máximo, los cuales son reducidos a un solo elemento. A fin de aprovechar al máximo el conocimiento de la gramática española y mantener un número mínimo de reglas, se optó por llevar al algoritmo de reducción una característica de ciertas categorías gramaticales y sintagmas.

La característica que se observó es la siguiente: ciertas etiquetas gramaticales iguales que se encuentran de manera consecutiva pueden ser reducidas a una sola etiqueta gramatical. Las categorías gramaticales que comparten esta propiedad son las siguientes:

- Sustantivo
- Adjetivo
- Verbo

Estas tres categorías gramaticales fungen como núcleo para sus respectivos sintagmas y a su vez heredan esta propiedad a este tipo de frases. Además, para la propiedad de la que hablamos el sintagma preposicional no es excepción. De modo tal que las frases que comparten esta propiedad son:

- Sintagma nominal
- Sintagma adjetival
- Sintagma preposicional
- Sintagma verbal

A pesar de que esta propiedad no puede ser considerada una generalidad para todo el lenguaje español, ha demostrado ser benéfica para realizar las reducciones de las consultas sujetas a prueba. A continuación se detalla cómo es que esta propiedad funciona.

Tomemos como ejemplo la oración: "*Ana Sofía aguardaba con ansias el transporte público*", donde los elementos que poseen esta propiedad son sustantivos.

Ana	Sofía	aguardaba	con	ansias	el	transporte	público
sus	sus	ver	pre	sus	art	sus	adj

En este ejemplo, el segmento “Ana Sofía” contiene dos sustantivos y ambos son nombres propios. Para fines gramáticos, “Ana Sofía” es considerada el sujeto en la oración. En lugar de tener dos elementos y buscar reglas gramaticales para reducirlos, lo cual resulta en un incremento en las iteraciones del algoritmo, se reducen ambos elementos a ser considerados como uno solo.

sus	sus	ver	pre	sus	art	sus	adj
sus		ver	pre	sus	art	sus	adj

Realizar el procedimiento anterior reduce el número de elementos que contiene una variación, así mismo, reduce el número de iteraciones que el algoritmo de reducción tiene que realizar.

Ahora veamos cómo se comporta esta propiedad para algunos de los elementos más relevantes.

Verbo:

La propiedad en el verbo nos permite trabajar con tiempos compuestos y la aparición de verbos auxiliares antes de verbos principales.

José	iba	corriendo	por	el	parque
sus	ver	ver	pre	art	sus

En la oración, “iba corriendo” es la acción que realiza el sujeto, la cual está compuesta por un verbo auxiliar y un verbo principal. La oración se puede reducir a la siguiente forma:

sus	ver	ver	pre	art	sus
sus	ver		pre	art	sus

Adjetivo:

Los adjetivos generalmente califican o determinan a los sustantivos. Cuando los adjetivos se encuentran de manera consecutiva pueden ser reducidos, ya que su objetivo es describir el sujeto u objeto.

¿Podrías	comprar	globos	azules	rojos	blancos	y	amarillos?
ver	ver	sus	adj	adj	adj	con	adj

En la oración, todos los colores que se muestran están calificando a *globos*; por lo tanto, se puede reducir el número de elementos que participan.

ver	ver	sus	adj	adj	adj	con	adj
ver		sus	adj			con	adj

Sintagma nominal:

Los sintagmas nominales consecutivos suelen fungir como elementos del sujeto o como elementos del complemento en la oración.

Los	perros	y	los	gatos	son	buenas	mascotas
art	sus	con	art	sus	ver	adj	sus

En la oración, *los perros* y *los gatos* son divididos por una conjunción; sin embargo, *y los gatos* es un sintagma nominal que funge como complemento, así pues:

art	sus	con	art	sus	ver	adj	sus
SNom		SNom			ver	adj	sus
SNom					ver	adj	sus

Sintagma preposicional:

Los sintagmas preposicionales consecutivos suelen fungir como complementos.

Por	la	senda	con	mi	hermano	iba	corriendo
pre	art	sus	pre	adj	sus	ver	ver

En la oración, *por la senda* y *con mi hermano* son complementos a una oración implícita, donde existen dos verbos de manera consecutiva.

pre	art	sus	pre	adj	sus	ver	ver
SPre			SPre			ver	
SPre						ver	

Gracias a esta propiedad, el número de elementos que se deben analizar en el análisis sintáctico disminuye considerablemente.

4.3.3.3 Buscar regla

Las funciones mostradas en las líneas 4 a 6 del Algoritmo 4.2 están implementadas con un par de ciclos *for.* Dentro de estos ciclos se encuentra una función que verifica si, a partir de la posición i, existe una regla de k elementos que reduzca a los elementos consecutivos de i a la derecha.

Una vez que se selecciona el segmento en base a la posición i, la variable k puede tomar los valores de 3 (ver Figura 4.11), 2 (ver Figura 4.12) o 1 (ver Figura 4.13).

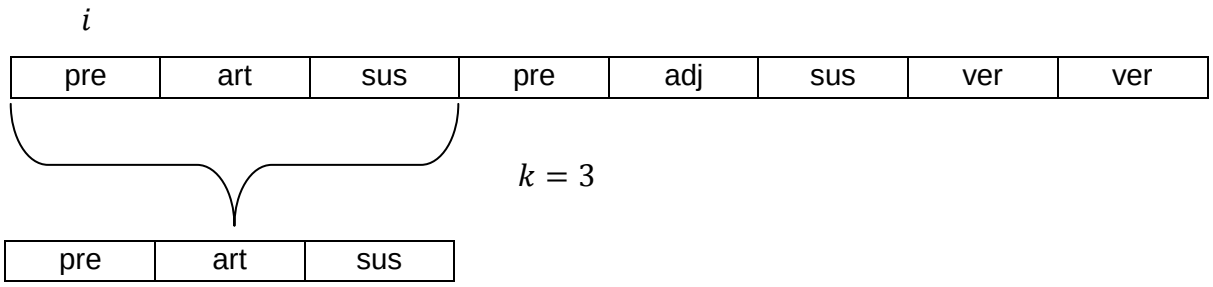

Figura 4.11 Selección del segmento para k = 3 elementos

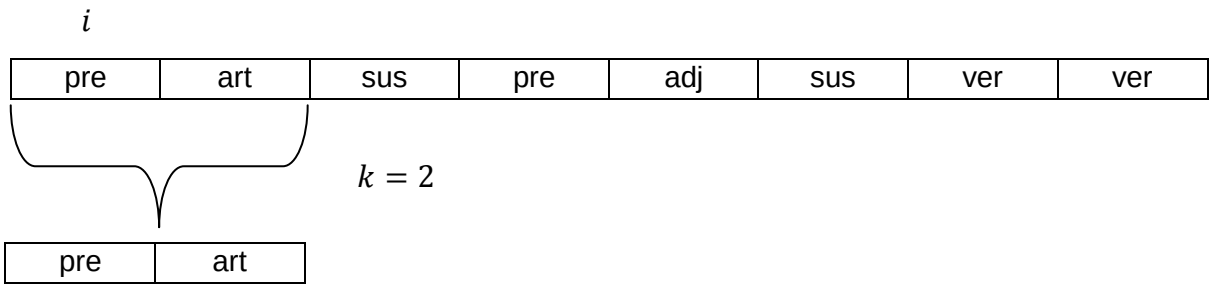

Figura 4.12 Selección del segmento para k = 2 elementos

En dado caso que no se encuentre una regla gramatical dentro del compendio que coincida con el segmento que se va a analizar, el índice i avanza una posición y busca de nuevo.

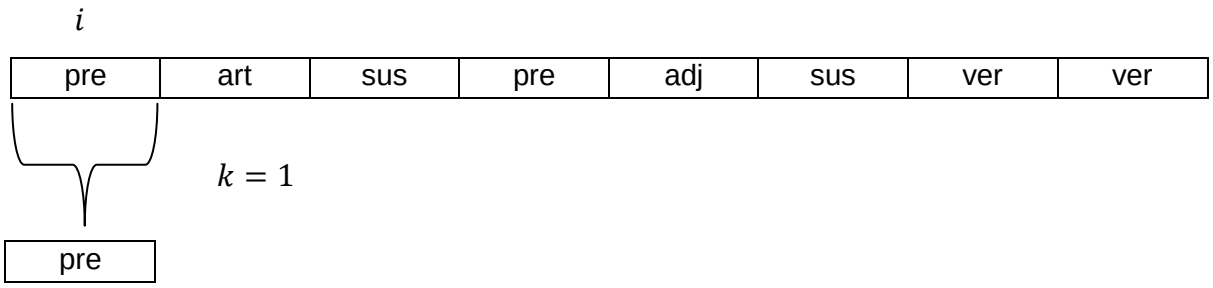

Figura 4.13 Selección del segmento para k = 1 elemento

4.3.3.4 Evaluación de la condición de salida

En este punto, la condición de salida es evaluada. Las posibles variaciones de etiquetas gramaticales que indican una correcta reducción, y por ende una oración sintácticamente correcta, se muestran a continuación:

1. SNom
2. SVer
3. SNom SVer
4. SVer SNom

4.3.4 Espacio de búsqueda

Hans van de Koot en 1992 define el problema de reconocimiento gramatical de palabras (Word Grammar Recognition (WGR)) como [20]:

Dada una oración x y una WG G, ¿es $x \in L(G)$?

A continuación se presenta el siguiente teorema, el cual define la complejidad del problema WGR.

Teorema: WGR es NP-completo.

La prueba de este teorema consiste en efectuar la transformación de 3SAT (el cual es un problema NP-completo conocido) a WGR, cuyos detalles se encuentran en [20].

Informalmente, el teorema anterior implica que es poco probable encontrar un algoritmo de tiempo polinomial que realice el análisis sintáctico exacto de una expresión en LN.

Como se mencionaba en secciones anteriores, éste es un algoritmo exhaustivo, el cual busca todas las posibles formas en las que una consulta puede ser reducida. Para la implementación del algoritmo usado en este proyecto la cantidad total de configuraciones a revisar es $|V| = \left\lfloor \frac{n}{3} \right\rfloor 3^{\left\lfloor \frac{n}{3} \right\rfloor}$

Tomemos como ejemplo la oración: “¿Cuántos ríos hay en Chicago?” y etiquetemos gramaticalmente cada uno de sus elementos léxicos:

¿Cuántos	ríos	hay	en	Chicago?
pro	sus	ver	pre	sus

Denominaremos a esta expresión como cadena productora (CP). Esta expresión es la base de todas las posibles reducciones que se generarán. El algoritmo procede a analizar todos los segmentos de 3, 2 y 1 elementos de la CP en busca de una regla gramatical para reducir los elementos del segmento.

En la Tabla 4.5 mostramos cómo la CP tiene 12 segmentos a analizar, de los cuales, sólo 7 presentan una coincidencia de regla gramatical, y por lo tanto, generan una reducción.

Tabla 4.6 Cadenas generadas a partir de la cadena productora original

CG	Elementos	Categorías					Reducción
1	3	**pro**	**sus**	**ver**	pre	sus	
1R	3	pro	sus	ver	pre	sus	**No**
2	2	**pro**	**sus**	ver	pre	sus	
2R	2	**SNom**		ver	pre	sus	**Sí**
3	1	**pro**	sus	ver	pre	sus	
3R	1	**SNom**	sus	ver	pre	sus	**Sí**
4	3	pro	**sus**	**ver**	**pre**	sus	
4R	3	pro	sus	ver	pre	sus	**No**
5	2	pro	**sus**	**ver**	pre	sus	
5R	2	pro	sus	ver	pre	sus	**No**
6	1	pro	**sus**	ver	pre	sus	
6R	1	pro	**SNom**	ver	pre	sus	**Sí**
7	3	pro	sus	**ver**	**pre**	**sus**	
7R	3	pro	sus	**SVer**			**Sí**
8	2	pro	sus	**ver**	**pre**	sus	
8R	2	pro	sus	ver	pre	sus	**No**
9	1	pro	sus	**ver**	pre	sus	
9R	1	pro	sus	**SVer**	pre	sus	**Sí**
10	2	pro	sus	ver	**pre**	**sus**	
10R	2	pro	sus	ver	**SPre**		**Sí**
11	1	pro	sus	ver	**pre**	sus	
11R	1	pro	sus	ver	pre	sus	**No**
12	1	pro	sus	ver	pre	**sus**	
12R	1	pro	sus	ver	pre	**SNom**	**Sí**

Y para $\forall\ v\ \in V$ se calculan:

$$total\ de\ reducciones\ analizadas = \frac{|v|!}{a!\ b!\ c!}$$

donde:

a es la cantidad de grupos de tres elementos

b es la cantidad de grupos de dos elementos

c es la cantidad de grupos de un elemento

El cálculo se hace cada vez que se realiza una reducción hasta $N - 1$ veces donde N es el número de elementos que posee una variación.

El algoritmo toma la primera coincidencia de reducción y la vuelve la nueva CP para realizar este proceso nuevamente. Para este ejemplo particular, la primera coincidencia se encuentra en la reducción 2. Por lo tanto, la nueva CP es 2R:

SNom	ver	pre	sus

En la Tabla 4.7 se muestran las cadenas que resultaron de CP = 2R. De nuevo, el algoritmo toma la primer coincidencia y continúa evaluando todos los segmentos posibles.

Tabla 4.7 Cadenas generadas a partir de la cadena productora 2R

CG	Elementos	Categorías				Reducción
13	3	**SNom**	**ver**	**pre**	sus	
13R	3	SNom	ver	ver	pre	**No**
14	2	**SNom**	**ver**	pre	sus	
14R	2	SNom	ver	pre	sus	**No**
15	1	**SNom**	ver	pre	sus	
15R	1	SNom	ver	pre	sus	**No**
16	3	SNom	**ver**	**pre**	**sus**	
16R	3	SNom	**SVer**			**Sí**
			⋮			
			⋮			
21	1	SNom	ver	pre	**sus**	
21R	1	SNom	ver	pre	**SNom**	**Sí**

Esto da una idea de la cantidad de reducciones totales que el algoritmo analiza con el fin de obtener todas las posibles reducciones de una consulta.

CAPÍTULO 5: PRUEBAS EXPERIMENTALES

En este capítulo se describen las características de hardware y software del equipo en el que fue realizada la experimentación, además de los resultados experimentales y un breve análisis de los mismos.

5.1 Ajustes de la experimentación

Todas las pruebas fueron realizadas en una laptop con las especificaciones mostradas en la Tabla 5.1. Mientras que el software usado tanto para la implementación como para las pruebas, se encuentra descrito en la Tabla 5.2.

Tabla 5.1 Especificaciones del equipo

Sistema	Descripción
Procesador	Intel Core i5 3210M @ 2.50 GHz
Memoria	4 GB
Sistema operativo	Windows 7 Home Premium 64x

Tabla 5.2 Especificaciones del software

Software	Descripción
Entorno	Eclipse Java EE IDE (Juno Release)
Lenguaje	JAVA
JAVA	Versión 1.7

5.2 Iteraciones del algoritmo

Tomemos como ejemplo la oración: “Lista el número de pasajeros de cada vuelo”. Cada una de las consultas del ANEXO D: Corpus de consultas cuenta con este formato.

@

8			
Lista el número de pasajeros de cada vuelo			
Lista	ver	sus	adj
el	art		
número	sus		
de	pre		
pasajeros	sus		
de	pre		
cada	pro	adj	
vuelo	ver	sus	

@@

Primeramente, se realiza el proceso de llenar la estructura y codificar los elementos léxicos a sus identificadores numéricos.

Palabra	Expresion regular	Categorías	No. categorías
Lista	palabra	[5][2][3]	3
el	palabra	[1]	1
número	palabra	[2]	1
de	palabra	[7]	1
pasajeros	palabra	[2]	1
de	palabra	[7]	1
cada	palabra	[4][3]	2
vuelo	palabra	[5][2]	2

Posteriormente se generan todas las variaciones de la oración.

i	0	1	2	3	4	5	6	7
0	5	1	2	7	2	7	4	5
1	5	1	2	7	2	7	4	2
2	5	1	2	7	2	7	3	5
3	5	1	2	7	2	7	3	2
4	2	1	2	7	2	7	4	5
5	2	1	2	7	2	7	4	2
6	2	1	2	7	2	7	3	5
7	2	1	2	7	2	7	3	2
8	3	1	2	7	2	7	4	5
9	3	1	2	7	2	7	4	2
10	3	1	2	7	2	7	3	5
11	3	1	2	7	2	7	3	2

Después, se procede a tomar una a una las variaciones generadas para ser sujetas a la reducción. Para fines de explicación de este ejemplo, usaremos la variación 3.

Lista	el	número	de	pasajeros	de	cada	vuelo
ver	art	sus	pre	sus	pre	adj	sus
5	1	2	7	2	7	3	2

Se analizan los segmentos (5, 1, 2) (ver Figura 5.1), (5, 1) (ver Figura 5.2) y (5) (ver Figura 5.3), donde no se encuentran reglas para reducir, por lo que el resultado es el que se muestra en la Figura 5.4.

Lista	**el**	**número**	de	pasajeros	de	cada	vuelo
ver	**art**	**sus**	pre	sus	pre	adj	sus
5	**1**	**2**	7	2	7	3	2

Figura 5.1 Segmento (5, 1, 2)

Lista	**el**	número	de	pasajeros	de	cada	vuelo
ver	**art**	sus	pre	sus	pre	adj	sus
5	**1**	2	7	2	7	3	2

Figura 5.2 Segmento (5, 1)

Lista	el	número	de	pasajeros	de	cada	vuelo
ver	art	sus	pre	sus	pre	adj	sus
5	1	2	7	2	7	3	2

Figura 5.3 Segmento (5)

ver	**art**	**sus**	**pre**	**sus**	**pre**	**adj**	**sus**
5	1	2	7	2	7	3	2

Figura 5.4 Variación resultante1

En este caso, se analizan los segmentos (1, 2, 7) (ver Figura 5.5), (1, 2) (ver Figura 5.6) y (1) (ver Figura 5.7), donde el segmento (1, 2) coincide con una regla gramatical. Dicha reducción produce como resultado que estos números se reduzcan al elemento 10 resultando la variación mostrada en la Figura 5.8.

ver	**art**	**sus**	**pre**	sus	pre	adj	sus
5	**1**	**2**	**7**	2	7	3	2

Figura 5.5 Segmento (1, 2, 7)

ver	**art**	**sus**	pre	sus	pre	adj	sus
5	**1**	**2**	7	2	7	3	2

Figura 5.6 Segmento (1, 2)

ver	art	sus	pre	sus	pre	adj	sus
5	**1**	2	7	2	7	3	2

Figura 5.7 Segmento (1)

ver	art	sus	pre	sus	pre	adj	sus
ver	**SNom**		**pre**	**sus**	**pre**	**adj**	**sus**
5	10		7	2	7	3	2

Figura 5.8 Variación resultante 2

Ahora se analizan los segmentos (7, 2, 7) (ver Figura 5.9), (7, 2) (ver Figura 5.10) y (7) (ver Figura 5.11), donde el segmento (7, 2) coincide con una regla gramatical. Dicha regla gramatical produce como resultado que estos números se reduzcan al elemento 14 dando como resultado la variación mostrada en la Figura 5.12.

ver	SNom	**pre**	**sus**	**pre**	adj	sus
5	10	**7**	**2**	**7**	3	2

Figura 5.9 Segmento (7, 2, 7)

ver	SNom	**pre**	**sus**	pre	adj	sus
5	10	**7**	**2**	7	3	2

Figura 5.10 Segmento (7, 2)

ver	SNom	**pre**	sus	pre	adj	sus
5	10	**7**	2	7	3	2

Figura 5.11 Segmento (7)

ver	art	sus	pre	sus	pre	adj	sus
ver	SNom		pre	sus	pre	adj	sus
ver	**SNom**		**SPre**		**pre**	**adj**	**sus**
5	10		14		7	3	2

Figura 5.12 Variación resultante 3

En este caso se analizan los segmentos (7, 3, 2) (ver Figura 5.13), (7, 3) (ver Figura 5.14) y (7) (ver Figura 5.15), donde el segmento (7, 3, 2) coincide con una regla gramatical. Esta regla gramatical produce como resultado que estos números se reduzcan al elemento 14 dando como resultado la variación mostrada en la Figura 5.16.

ver	SNom	SPre	**pre**	**adj**	**sus**
5	10	14	**7**	**3**	**2**

Figura 5.13 Segmento (7, 3, 2)

ver	SNom	SPre	**pre**	**adj**	sus
5	10	14	**7**	**3**	2

Figura 5.14 Segmento (7, 3)

ver	SNom	SPre	**pre**	adj	sus
5	10	14	**7**	3	2

Figura 5.15 Segmento (7)

ver	art	sus	pre	sus	pre	adj	sus
ver	SNom		pre	sus	pre	adj	sus
ver	SNom		SPre		pre	adj	sus
ver	**SNom**		**SPre**		**SPre**		
5	10		14		14		

Figura 5.16 Variación resultante 4

Además, tenemos el segmento (14, 14), el cual por el método de reducción consecutiva puede ser reducido a un solo elemento como se muestra en la Figura 5.17.

ver	art	sus	pre	sus	pre	adj	sus
ver	SNom		pre	sus	pre	adj	sus
ver	SNom		SPre		pre	adj	sus
ver	SNom		SPre		SPre		
ver	**SNom**		**SPre**				
5	10		14				

Figura 5.17 Variación resultante 5

Dicha redución produce que los elementos a analizar ahora sean: (5, 10, 14) (ver Figura 5.18), (5, 10) (ver Figura 5.19) y (5) (ver Figura 5.20). En estas circunstancias, el segmento (5,10) coincide con una regla gramatical, la cual da como resultado que estos números se reduzcan al elemento 12 dando como resultado la variación mostrada en la Figura 5.21.

ver	SNom	SPre
5	10	14

Figura 5.18 Segmento (5, 10, 14)

ver	SNom	SPre
5	10	14

Figura 5.19 Segmento (5, 10)

ver	SNom	SPre
5	10	14

Figura 5.20 Segmento (5)

ver	art	sus	pre	sus	pre	adj	sus
ver	SNom		pre	sus	pre	adj	sus
ver	SNom		SPre		pre	adj	sus
ver	SNom		SPre		SPre		
ver	SNom		SPre				
SVer			**SPre**				
12			14				

Figura 5.21 Variación resultante 6

Por último quedan dos elementos (12, 14) los cuales son sujetos a reducción, donde se identifican los segmentos (12, 14) (ver Figura 5.22) y (12) (ver Figura 5.23).

SVer	SPre
12	14

Figura 5.22 Segmento (12, 14)

SVer	SPre
12	14

Figura 5.23 Segmento (12)

En estas circunstancias, el segmento (12, 14) coincide con una regla gramatical, la cual da como resultado que estos números se reduzcan al elemento 12 dando como resultado la variación mostrada en la Figura 5.24.

ver	art	sus	pre	sus	pre	adj	sus
ver	SNom		pre	sus	pre	adj	sus
ver	SNom		SPre		pre	adj	sus
ver	SNom		SPre		SPre		
ver	SNom		SPre				
SVer			SPre				
SVer							
12							

Figura 5.24 Variación resultante 7

Una vez que se vuelve a analizar la oración en búsqueda de una posible reducción, el algoritmo, al no ser capaz de encontrar una reducción somete la variación resultante a la condición de salida. En este caso el elemento 12 es una condición de salida válida. Por lo tanto, esta variación produce un árbol sintáctico correcto.

5.3 Resultados

Se realizaron pruebas utilizando las consultas especificadas en el ANEXO D: Corpus de consultas. A continuación se menciona la procedencia de los corpus de consulta usados para este proyecto.

El corpus de consulta CFA es un extracto obtenido a partir de los corpus de consultas de las bases de datos de Northiwnd [21], Pubs [21] y Geobase [22]. El corpus de consultas de ATIS es un extracto obtenido traducido al español a partir del original [23], el cual se encuentra en idioma inglés. El corpus de consultas de Pubs fue tomado de un trabajo realizado en el CENIDET en el año 2005 a cargo de Javier González [21].

A continuación se encuentra una descripción de la Tabla 5.3, la cual contiene los resultados.

- La primera columna indica el nombre del corpus de consultas.
- La segunda columna muestra el número de consultas de cada uno de los corpus que logró cumplir con alguna de las cuatro condiciones de salida en alguna de sus variaciones.
- La tercera columna contiene el total de consultas que tenía cada uno de los corpus.
- La cuarta columna muestra el valor de la segunda columna como un porcentaje respecto al valor de la tercera columna.
- La quinta columna es la más importante, denota el porcentaje de consultas de la cuarta columna cuyos árboles sintácticos fueron los correctos para que la oración tuviese sentido.

Tabla 5.3 Resultados de las pruebas

Corpus	No. de consultas		Porcentaje de consultas con condición de salida aceptada	Porcentaje de certidumbre de los resultados
	Condición de salida aceptada	Totales		
CFA	**30**	**31**	**97%**	**100%**
ATIS	**68**	**70**	**97%**	**100%**
Pubs	**66**	**69**	**96%**	**100%**

A cada una de las consultas se le determinó un tiempo fijo de ejecución de 5 segundos. Pasado este tiempo, si el algoritmo no era capaz de proporcionar al menos un árbol sintáctico, dicha consulta se considera “inalcanzable”. Este término significa que, debido a que el algoritmo es exhaustivo, toma demasiado tiempo considerar todas las posibles reducciones.

Pese a que puede existir una forma posible de reducir una de las variaciones que tenga alguna oración, puede ser que la reducción correcta sea encontrada al final de todas las posibles.

5.3.1 Pruebas negativas

Como especifica Belkan [24], los casos de prueba son construidos sistemáticamente a partir de elementos individuales, mientras que los corpus de prueba son extraídos de manera natural de los textos. Cualquiera que sea el caso, se deben tomar en cuenta ciertas consideraciones al momento de realizar la experimentación y reportar los resultados.

Uno de los aspectos que se destacan es la consideración de ejemplos negativos o gramaticalmente incorrectos. Belkan explica que este tipo de ejemplos no ocurre naturalmente en un corpus; sin embargo, someter a este tipo de ejemplos en verificadores gramaticales o correctores de lenguaje puede evidenciar información de gran importancia.

El motivo detrás de esto es que, hace una distinción confiable acerca del desempeño de estas herramientas con entradas correctas así como con entradas incorrectas.

A fin de confirmar el porcentaje de acierto del analizador sintáctico, se sometió su desempeño a pruebas negativas las cuales cuentan con las siguientes características:

- Las consultas usadas serían extraídas de los corpus de consulta CFA, ATIS y Pubs.
- Se elegirían 20 consultas de manera aleatoria de cada uno de los corpus.
- Las consultas deberían contener al menos un error gramatical en su redacción.

En un inicio se pensó en eliminar el verbo de la oración; sin embargo, ya que la mayoría de las consultas empleadas en los corpus de prueba resultan ser oraciones imperativas, eliminar el verbo, el cual está ubicado al inicio de la oración daba como resultado una estructura de una oración nominal.

Se optó por incluir de manera arbitraria palabras que de algún modo causaran un error gramatical en la consulta. De todas las categorías gramaticales reconocidas por la RAE, se seleccionó al adverbio para provocar dicho error como se muestra a continuación.

Dame	los	**títulos**	**de**	los	libros	

Dame	los	**títulos**	**rápidamente**	**de**	los	libros
ver	art	sus	adv	pre	art	sus

El adverbio funge como modificador del verbo y el adjetivo, su colocación arbitraria en algún segmento de la consulta provocaría un error debido a la rigidez de las reglas gramaticales para esta categoría.

Sin embargo, la colocación arbitraria del adverbio puede causar que la consulta tenga en algunos casos al menos un árbol sintáctico, por ejemplo, tomemos la consulta “Lista el número de gente en cada vuelo”.

Lista	el	número	de	gente	en	cada	vuelo
ver	art	sus	pre	sus	pre	adj	sus

Introduciendo el adverbio en la posición indicada (más abajo) y procediendo a realizar la reducción paso a paso, se llega a un punto donde no es posible seguir reduciendo el árbol sintáctico. Para esta consulta, en la que el adverbio está en esa posición, se genera un error de reducción, por lo que la prueba negativa es aceptable.

Lista	el	número	de	gente	**pronto**	en	cada	vuelo
ver	art	sus	pre	sus	**adv**	pre	adj	sus

ver	art	sus	en	sus	adv	en	adj	sus
SVer			SPre		adv	SPre		
SVer					adv	SPre		

Sin embargo, si se coloca el adverbio en alguna otra posición arbitraria (en este ejemplo está localizado al inicio de la consulta) puede darse el caso que se genere un árbol sintáctico que satisfaga la condición de salida.

Pronto	lista	el	número	de	gente	en	cada	vuelo
adv	ver	art	sus	pre	sus	pre	adj	sus

adv	ver	art	sus	pre	sus	pre	adj	sus
SVer		SNom		SPre		SPre		
SVer		SNom		SPre				
SVer				SPre				
SVer								

El analizador sintáctico presentó el error gramatical en 80% a 85% de las ocasiones, mientras que en el resto de los casos presentaba estructuras gramaticales, las cuales cumplían con algún conjunto de reglas gramaticales produciendo así un árbol sintáctico. En conclusión, ya que el corpus de esta prueba contenía algunas oraciones correctas, el algoritmo detectó correctamente los casos negativos (80-85%) como positivos.

Esto significa que debido a que la inclusión de un adverbio en la consulta se dio de manera arbitraria, se dio el caso donde la posición del adverbio provocó la generación de un árbol sintáctico correcto. La manera de evitar esto sería colocar el adverbio en una posición específica, en la cual se asegurara que generara un error gramatical, dando así como resultado una prueba negativa aceptable.

CAPÍTULO 6: CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo se presentan las conclusiones de este proyecto después de haber realizado un análisis de la gramática española para posteriormente diseñar e implementar un analizador sintáctico para consultas a bases de datos en lenguaje natural.

También se mencionan las distintas áreas de oportunidad en el tema de análisis sintáctico para el procesamiento del lenguaje natural, haciendo uso del idioma español.

6.1 Conclusiones

En este trabajo se diseñó e implementó un analizador sintáctico para el lenguaje español, logrando así cumplir el objetivo general de este proyecto de tesis. Como resultado del proyecto, se concluye lo siguiente:

1. Se implementó un analizador sintáctico que permite incrementar y modificar las reglas gramaticales de producción sin la necesidad de editar el código (Java).

2. Pese a no ser capaz de eliminar la ambigüedad sintáctica que llegue a existir en una oración, el analizador sintáctico es capaz de determinar todos los árboles sintácticos de una oración.

3. Se propone el diseño de una base de datos de un lexicón, el cual contiene todas las categorías gramaticales reconocidas por la RAE así como todos sus accidentes gramaticales (ver ANEXO C: Diseño del lexicón).

4. Se logró diseñar y realizar un compendio de reglas gramaticales basándose en información de la RAE [14] [17], logrando así obtener un menor número de reglas gramaticales que sea capaz de analizar un mayor número de oraciones.

5. Se redactó un compendio de la gramática española, el cual está validado por información de la RAE [14] [17] (ver ANEXO B: Fundamentos de gramática).

Se destaca el desarrollo del compendio de la gramática española, ya que gracias a este conocimiento se logró impactar la manera en la que se desarrolló tanto el diseño de la base de datos del lexicón, como el diseño e implementación

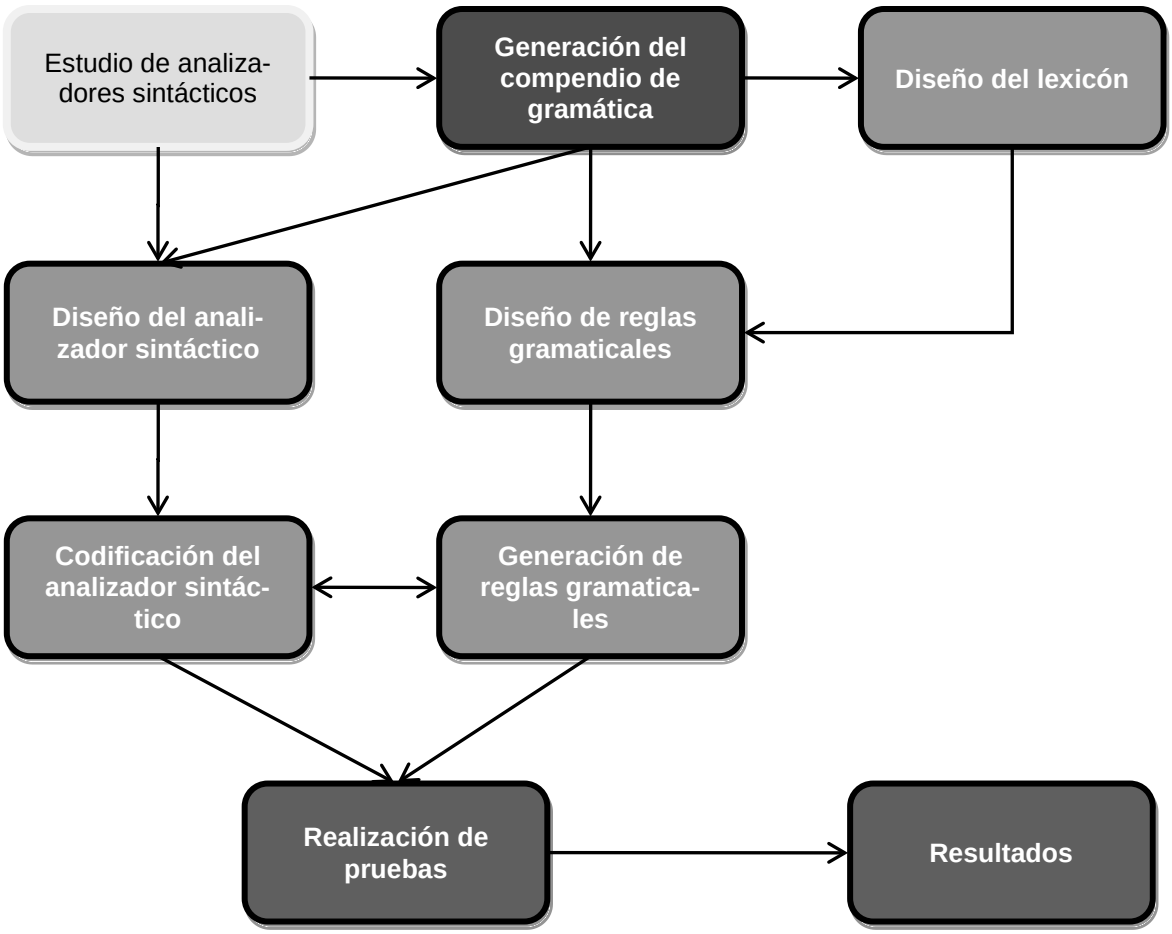

Figura 6.1 Sinergia de actividades realizadas

de las reglas gramaticales y algoritmo de reducción como se ilustra en la Figura 6.1. El resultado fue un mayor porcentaje de acierto en las consultas.

6.2 Trabajos futuros

Actualmente existe un largo camino por recorrer en el área del lenguaje natural del español. Algunas áreas de oportunidad en las que se pueden enfocar otros proyectos, a fin de complementar este trabajo, son:

1. Realizar un estudio abordando la problemática de la implementación de un enfoque gramatical en un entorno computacional. Los enfoques gramaticales existentes suelen ser vistos desde un punto de vista lingüístico que involucra un pensamiento y procesamiento humano; sin embargo, la implementación de dichos enfoques gramaticales en un entorno computacional, el cual sea capaz de obtener los mismos resultados involucra un estudio más detallado.

2. Optimizar el algoritmo de reducción propuesto en este proyecto de tesis. El algoritmo que se propone hace una búsqueda de todos los árboles sintácticos posibles de una oración y no se encuentra optimizado. Esto implica un tiempo de ejecución muy largo, no sólo si la oración no tiene un árbol sintáctico sino también si el árbol sintáctico correcto se encuentra al final de todas las posibles opciones. Se propone realizar una optimización del algoritmo de reducción haciendo uso de un análisis semántico superficial o el uso de programación dinámica.

3. Desarrollar un lexicón que contenga todas las palabras de la lengua española haciendo uso del diseño propuesto en el ANEXO C: Diseño del lexi-

ción. Esto permitiría contar con un correcto etiquetado léxico de las oraciones que son introducidas en el analizador sintáctico, así mismo, la información de los accidentes gramaticales de cada una de las palabras proveerá de datos de relevancia para el resto del procesamiento de la consulta.

4. Implementar un módulo que permita al analizador sintáctico ser capaz de procesar locuciones.

5. Ampliar la funcionalidad del analizador sintáctico para que permita procesar signos de puntuación; verificar congruencias de género, número, persona verbal, tiempo verbal, etc.; y procesar excepciones a las reglas gramaticales.

ANEXO A: MARCO TEÓRICO

Con el paso de los años, el volumen de información generada por los sistemas informáticos se ha ido incrementando de manera exponencial. Debido a este fenómeno los desarrolladores se han dado a la tarea de elaborar mecanismos que sean capaces de acceder y procesar dicha información de manera más rápida.

Sin embargo, dichas herramientas requieren un conocimiento sobre algún lenguaje formal específico para su uso (como SQL), tal hecho provocó que se desarrollaran nuevas herramientas que no dependieran de un lenguaje formal para acceder a la información.

A.1 Procesamiento del lenguaje

A continuación se detallan ciertos conceptos que serán de ayuda para el entendimiento del procesamiento lenguaje.

A.1.1 Lenguaje

Conjunto de sonidos articulados o símbolos con que el hombre manifiesta lo que piensa o siente. Cuando se habla de lenguajes se pueden diferenciar dos clases muy bien definidas [2] [24]:

- Los lenguajes naturales como el español, inglés, francés, etc.
- Los lenguajes formales como los lenguajes de programación (p. ej., SQL), el lenguaje de la lógica matemática, etc.

A.1.2 Lenguaje natural

Lenguaje hablado o escrito por humanos, opuesto a un lenguaje de programación utilizado para programar o comunicarse con computadoras. Existen dos campos en el estudio del entendimiento del lenguaje natural [24]:

- Entendimiento del lenguaje escrito, que utiliza el conocimiento léxico, sintáctico y semántico del lenguaje, unido a la información o conocimiento del dominio.
- Entendimiento del lenguaje oral, que comprende todo lo del campo anterior junto con toda la fonología.

A.1.3 Lenguaje formal

Un lenguaje formal es un lenguaje artificial creado por el hombre, el cual está formado por símbolos y fórmulas y tiene como objetivo fundamental formalizar la programación de computadoras o representar simbólicamente un conocimiento [24].

A.1.4 Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (PLN o NLP en inglés) es la capacidad que tiene una computadora para procesar la información comunicada, no simplemente las letras o los sonidos de un lenguaje formal [4].

A.2 Interfaces de lenguaje natural

Las interfaces de lenguaje natural son mecanismos de comunicación entre una persona y una máquina a través de lenguaje natural, donde los distintos fe-

nómenos lingüísticos sirven como controladores para la creación, modificación y selección de datos. Dichas interfaces cuentan con un modelo general descrito en la Figura A.1.

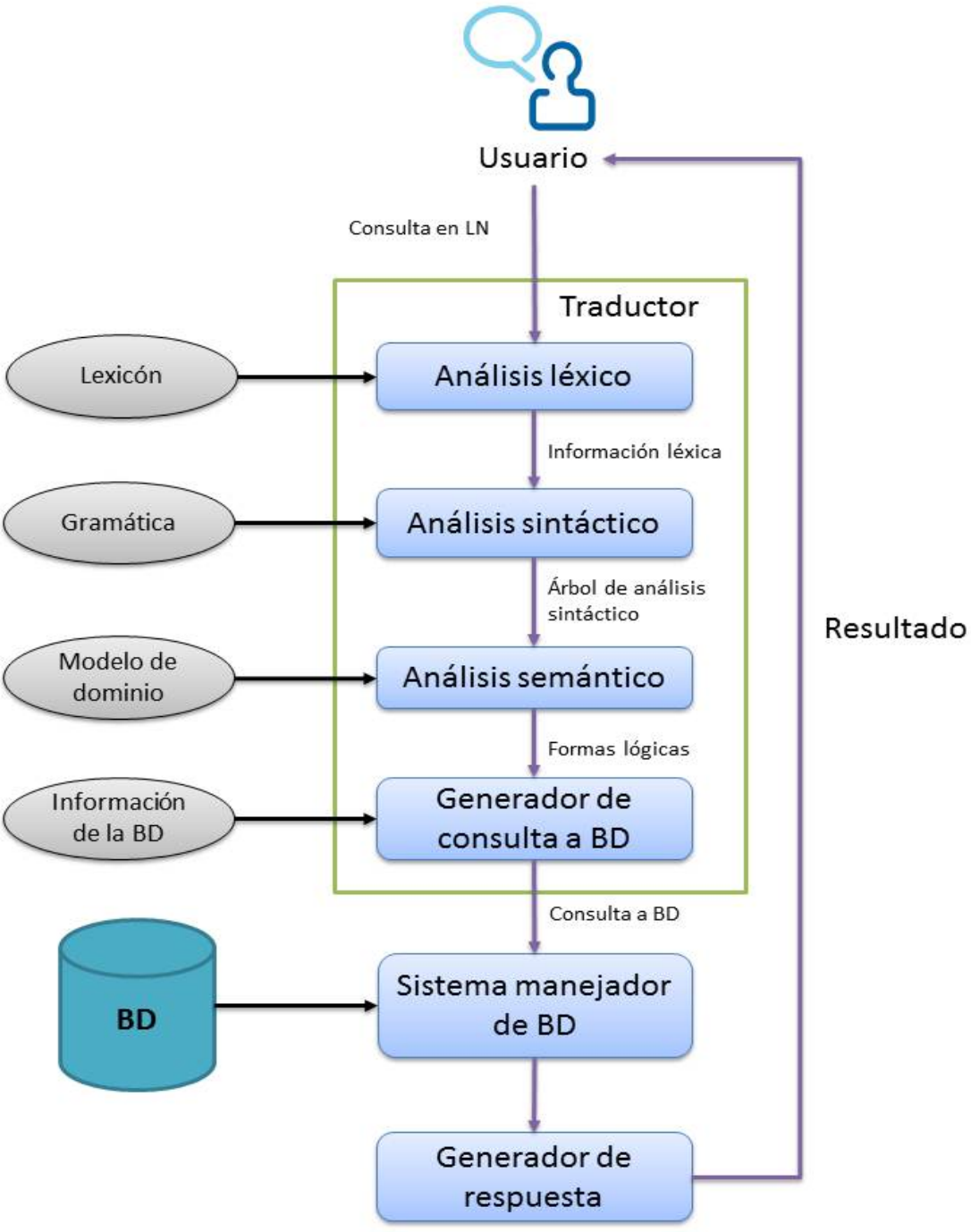

Figura A.1 Modelo general de una interfaz de lenguaje natural

A.2.1 Análisis léxico

La información léxica está contenida en el lexicón, es decir, en el conjunto de unidades léxicas pertenecientes a un sistema lingüístico. Dicha información consta de la etiqueta relativa a la categoría gramatical de cada unidad lingüística (sustantivo, verbo, pronombre, etc.) y de una o varias etiquetas correspondientes a cada uno de los rasgos de subcategorización que hacen posible que cada unidad lingüística seleccione otra u otras a la hora de combinarse formando las distintas oraciones posibles de una lengua [25].

A.2.2 Análisis sintáctico

La sintaxis trata la combinación de las palabras en la frase. Los problemas principales de los que se ocupa la sintaxis se refieren al orden de las palabras, a los fenómenos de reacción (es decir, la manera en que ciertas palabras imponen a otras variaciones de número, género, etc.) y las funciones que las palabras puedan cumplir en la oración [25].

Algunas de las tareas que se encarga de realizar el análisis sintáctico son las siguientes:

- Generación del árbol sintáctico.
- Corrección de errores sintácticos.
- Resolución de la ambigüedad sintáctica.
- Identificación de los diálogos.
- Conversión a la estructura canónica.

A.2.3 Análisis semántico

La semántica proporciona el significado de las palabras según el contexto. Gran parte de la información semántica de una unidad léxica ya se encuentra contenida en forma de rasgos semánticos en la descripción de dicha unidad; es decir, la información semántica es responsable de la correcta combinación de unidades léxicas en un discurso [25].

A.3. Los analizadores sintácticos

Ahora bien centraremos nuestra atención en los analizadores sintácticos y su implementación, la cual está determinada principalmente por dos métodos: el descendente (Top-Down) y el ascendente (Bottom-Up). Cualquiera que sea el caso, ambos métodos analizan un símbolo a la vez de izquierda a derecha usando las reglas de la gramática formal.

A.3.1 Analizador sintáctico ascendente

Comienza tomando cada palabra que constituye la oración de entrada y la etiqueta como terminación del árbol a construir. Este método conoce la parte derecha de las reglas de producción y trata de sustituirla por la parte izquierda que denota la regla que produce.

A.3.2 Analizador sintáctico descendente

Comienza tomando el símbolo inicial que representa la oración y en base a las reglas de producción construye una estructura que representa la secuencia de palabras para construir la oración de entrada. El proceso se visualiza como un ár-

bol, en donde se construyen las estructuras parciales una a una desde el símbolo inicial y continúa de manera descendente hasta encontrar la estructura sintáctica adecuada para la oración de entrada.

A.4 Procesamiento de texto

Habiendo detallado la metodología que siguen los analizadores sintácticos para su creación, es necesario especificar las normas o reglas que siguen para que sean lingüísticamente correctos. Ya que las frases son secuencias gramaticales, éstas deben obedecer ciertas leyes gramaticales.

A.4.1 Sintaxis

Establecer métodos que determinen únicamente las secuencias gramaticales en el procesamiento lingüístico de textos, ha sido el objetivo de los formalismos gramaticales de la lingüística computacional. En ella se han considerado dos enfoques principales para describir la gramaticalidad de las oraciones.

A.4.2 Enfoque de constituyentes

Los constituyentes y la suposición de la estructura de frase, sugerida por Lonard Bloomfield en 1933, es el enfoque en el que las oraciones se analizan mediante un proceso de segmentación y clasificación. Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso para cada parte dividiéndola en subconstituyentes, y así sucesivamente hasta que las partes sean las partes de la palabra indivisibles dentro de la gramática (morfemas) [26] como aparece en la Figura A.2.

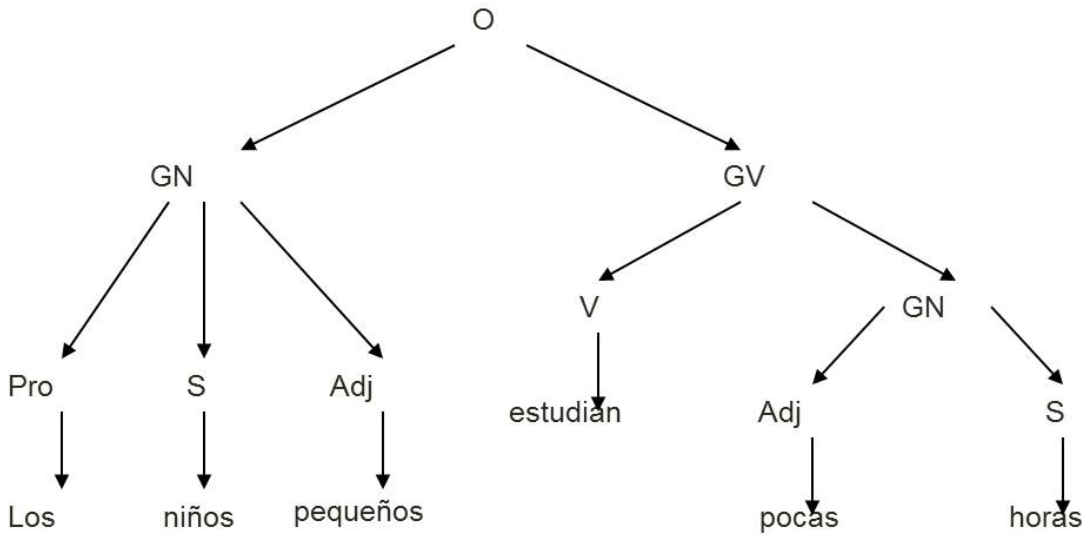

Figura A.2 Enfoque de constituyentes

A.4.3 Enfoque de dependencias

El primer intento real para construir una teoría que describiera las gramáticas de dependencias fue el trabajo de Lucien Tesnière en 1959. Las dependencias se establecen entre pares de palabras, donde una es principal o rectora y la otra está subordinada a (o dependiente de) la primera. Si cada palabra de la oración tiene una palabra propia rectora, la oración entera se ve como una estructura jerárquica de diferentes niveles, como un árbol de dependencias. La única palabra que no está subordinada a otra es la raíz del árbol [26] como aparece en la Figura A.3.

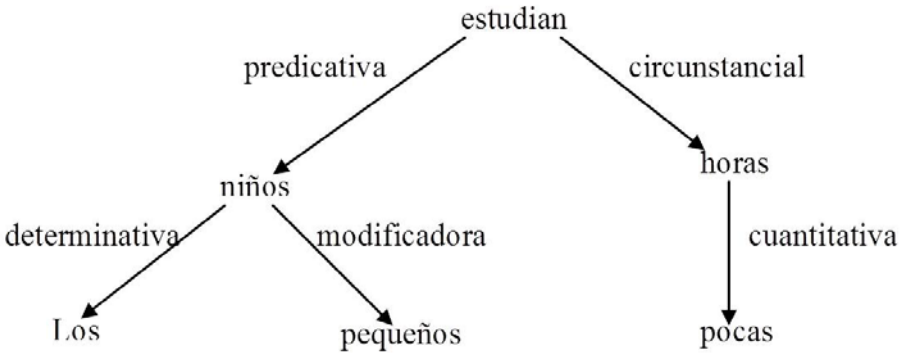

Figura A.3 Enfoque de dependencias

A.5 Gramáticas

Ahora se abordarán algunas de las distintas gramáticas que se pueden utilizar para implementar un analizador sintáctico. Cabe aclarar que el término gramática empleado en este apartado hace referencia al tipo de análisis que se realiza sobre la oración con el fin de elaborar el árbol sintáctico, a diferencia del concepto usual de gramática de un idioma específico.

A.5.1 Gramática de estructura de frase generalizada (GPSG)

Fue iniciada por Geral Gazdar en 1981, y desarrollada por él y un grupo de investigadores. La idea central de la GPSG es que las gramáticas usuales de estructura de frase independientes del contexto pueden mejorarse en formas que no enriquecen su capacidad generativa, pero que las hacen adecuadas para la descripción de la sintaxis de lenguajes naturales. Al situar la estructura de frase, otra vez, en un lugar principal, consideraban que los argumentos que se habían aducido contra las CFG, como una teoría de sintaxis, eran argumentos relacionados

con la eficiencia o la elegancia de la notación y no realmente con la cobertura del lenguaje.

La GPSG propone sólo un nivel sintáctico de representación, que corresponde a la estructura superficial, y reglas que no son de estructura de frase, en el sentido en que no están en una correspondencia directa con partes del árbol [26].

A.5.2 Gramática de léxica funcional (LFG)

Fue desarrollada por Bresnan en 1982 y Dalrymple en 1995. Comparte con otros formalismos la idea de que los conceptos relacionales, como sujeto y objeto, son de importancia central y no pueden definirse en términos de estructuras de árboles. La LFG considera que hay más en la sintaxis de lo que se puede expresar con árboles de estructura de frase. La LFG se ha centrado en el desarrollo de una teoría universal acerca de cómo las estructuras de constituyentes se asocian con los objetos sintácticos.

También considera la estructura de frase como una parte esencial de la descripción gramatical [26].

A.5.3 Gramática de estructura de frase dirigida por el núcleo-h (HPSG)

Iniciada por Pollard y Sag en 1987 y revisada por ellos mismos en 1994, evolucionó directamente de la GPSG con la intención de modificarla incorporando otras ideas y formalismos de los años ochenta. El nombre se modificó para reflejar el hecho de la importancia de la información codificada en los núcleos-h léxicos de las frases sintácticas, es decir, de la preponderancia del empleo de la marca head en el subconstituyente hijo principal [26].

La HPSG tiene las siguientes características:

- Posee una arquitectura basada en signos lingüísticos.
- Organiza la información lingüística mediante tipos, jerarquías de tipos y herencia de restricciones.
- Proyecta las frases mediante principios generales a partir de información con abundancia léxica.
- Organiza esa información léxica mediante un sistema de tipos léxicos.
- Factoriza las propiedades de frases en construcciones específicas y restricciones más generales.

A.5.4 Gramática categorial (GC)

Introducida por Ajdukiewicz, en 1935, adquirió importancia para los lingüistas cuando Montague en 1970 la usó como marco sintáctico de su método para analizar la semántica del lenguaje natural. La idea central de la GC es que una concepción enriquecida de categorías gramaticales puede eliminar la necesidad de muchas de las construcciones que se encuentran en otras teorías gramaticales (por ejemplo, de las transformaciones).

Una gramática categorial consiste, simplemente, en un diccionario y unas cuantas reglas que describen cómo pueden combinarse las categorías. Las categorías gramaticales se definen en términos de sus miembros potenciales para combinarse con otros constituyentes, por lo que algunos autores ven a la GC como una variación de la Gramática de Dependencias.

La suposición básica de la GC es que hay un conjunto fijo de categorías básicas, de las cuales se construyen otras categorías. Estas categorías básicas son: sustantivo, grupo nominal y oración; cada una de las categorías básica tiene características morfosintácticas determinadas por el lenguaje específico.

Los atractivos principales de la GC fueron su simplicidad conceptual y su adecuación a la formulación de análisis sintácticos y semánticos estrechamente ligados. Esto último debido a que se considera que restringe las asignaciones léxicas a expresiones básicas y a construcciones sintácticas potenciales, de tal forma que solamente se permiten las combinaciones de categorías sintácticas semánticamente significativas [26].

A.5.5 Gramática de restricciones (GR)

Toda la estructura relevante se asigna directamente de la morfología (considerada en el diccionario) y de mapeos simples de la morfología a la sintaxis (información de categorías morfológicas y orden de palabras, a etiquetas sintácticas). Las restricciones sirven para eliminar muchas alternativas posibles.

Los autores indican que su meta principal es el análisis sintáctico orientado a la superficie y basado en morfología de textos sin restricciones. Se considera sintaxis superficial, y no sintaxis profunda, porque no se asigna ninguna estructura sintáctica que no esté en correspondencia directa con los componentes léxicos de las formas de palabra que están en la oración.

Una idea relevante de la GR es poner en primer plano la descripción de ambigüedades, por lo que básicamente es un formalismo para escribir reglas de desambiguación. Divide el problema de análisis sintáctico en tres módulos: desambiguación morfológica, asignación de límites de cláusulas dentro de las oraciones, y asignación de etiquetas sintácticas superficiales. Las etiquetas indican la función sintáctica superficial de cada palabra y las relaciones de dependencia básica dentro de la cláusula y la oración [26].

A.5.6 Gramática de dependencias

Describe cómo los elementos se relacionan con otros elementos, y se concentra en las relaciones entre unidades sintácticas terminales, es decir, entre palabras [26]. Esta gramática genera árboles de dependencias, los cuales presentan las siguientes características:

- Muestran cuáles elementos se relacionan con cuáles otros y en qué forma.
- Revelan la estructura de una expresión en términos de ligas jerárquicas entre sus elementos reales, es decir, entre palabras.
- Se indican explícitamente los roles sintácticos, mediante etiquetas especiales.
- Contienen solamente nodos terminales, no se requiere una representación abstracta de agrupamientos [5].

A.5.7 Teoría de significado-texto (MTT)

La meta de la teoría es modelar la comprensión del lenguaje como un mecanismo que convierta los significados en los textos correspondientes y los textos en los significados correspondientes. Aunque no hay una correspondencia de uno a uno, ya que el mismo significado puede expresarse mediante diferentes textos, y un mismo texto puede tener diferentes significados.

La MTT emplea un mayor número de niveles de representación, tanto la sintaxis como la morfología y la fonología se dividen en dos niveles: profundo (D) y superficial (S). Está construida para una lengua en la que el orden de palabras es más flexible. El análisis sintáctico separa dependencias de orden de palabras de forma diferente. El orden de palabras (LI) conforma la sintaxis superficial, y las dependencias (DO) la sintaxis profunda [26].

ANEXO B: FUNDAMENTOS DE GRAMÁTICA

A continuación se detallan ciertos conceptos que serán de ayuda para el entendimiento de la gramática española.

B.1 La palabra

Una palabra es cada una de las porciones limitadas en una cadena hablada o escrita, que puede aparecer en varias posiciones y que a su vez está dotada de una función específica, posee un significado y espacios potenciales al inicio y al final. Dependiendo de la función que realice una palabra dentro de una oración se le asigna una categoría gramatical.

B.1.1 Categoría gramatical

Una categoría gramatical se encuentra definida por la Real Academia Española (RAE) [14] como cada una de las diferentes clases de oficio que desempeña una palabra dentro de una oración.

Las categorías gramaticales son las siguientes:

- Artículo.
- Sustantivo.
- Pronombre.
- Adjetivo.
- Verbo.
- Adverbio.

- Conjunción.
- Preposición.
- Interjección.

B.1.2 Accidente gramatical

En la gramática tradicional, las palabras a menudo experimentan una modificación flexiva variable para expresar valores de alguna categoría gramatical, como el género, el número, la persona o el tiempo [17].

Habiendo definido las categorías gramaticales y el accidente gramatical, se hará una breve explicación de cómo interactúan estos elementos en la gramática.

B.1.3 Artículo

Los artículos son una clase de palabras de carácter átono que indican si lo designado por el sustantivo o elemento sustantivado es o no es consabido.

La flexión del género y número da lugar a las siguientes cuatro formas del artículo definido: *el* y *la* para el masculino y el femenino singular, y *los* y *las* para masculino y femenino plural. A ellas debe añadirse la forma *lo*, carente de plural y tradicionalmente considerada como neutra.

Por otra parte, los artículos indefinidos se anteponen al nombre para indicar que éste se refiere a entidades no consabidas por los interlocutores. De igual manera presentando las variaciones en el singular, *un* y *una*, y en el plural, *unos* y *unas*.

Los artículos contractos, están formados a partir del artículo *el*, va unido con las preposiciones *a* y *de*. De esta manera se fusionan el artículo y la preposición y dan lugar a dos artículos contractos: *al* y *del*.

B.1.4 Pronombre

Los pronombres son una clase de palabra que sustituyen a otros términos para designar personas o cosas sin nombrarlas de forma directa, así mismo, puede fungir algunas veces como sustantivo.

- Personales: Que designa personas, animales o cosas mediante cualquiera de las tres personas gramaticales; v. gr., *yo, tú, él, nosotros, ellos, mi...*
- Demostrativos: Que señala personas animales o cosas; v. gr., *éste, ése, aquél, éstos, ésos, aquéllos...*
- Indefinidos: Que ocasionalmente alude a personas o cosas o expresa alguna noción que cuantifica; v.gr., *nada, algo, alguien, alguno, cualquier, cualquiera...*
- Relativos: Que desempeña una función en la oración a la que pertenece, inserta ésta en una unidad superior y tiene un antecedente, expreso o implícito; v. gr., *que, quien, cuyo, cual, cuantos...*
- Posesivos: Que denota posesión o pertenencia; v. gr., *mío, tuyo, suyo, cuyo...*
- Interrogativos: Que sin acompañar al nombre, permite construir enunciados interrogativos u oraciones interrogativas indirectas; v. gr., *qué, quién, cuánto, cuándo, cuál, dónde, cómo...*

B.1.5 Sustantivo

Clase de palabras que pueden funcionar como sujeto de la oración.

B.1.6 Adjetivo

Clase de palabras que califican o determinan al sustantivo. Un detalle al que se debe prestar especial atención en los adjetivos es la verificación de la concordancia en cuanto a la persona, género y número del sustantivo al que va ligado.

- Calificativo: Palabra que acompaña al sustantivo para expresar alguna cualidad de la persona o cosa nombrada; v. gr., *blanco, alto, azul.*
- Ordinal: Adjetivo numeral que expresa la idea de orden o sucesión; v. gr., *uno, segundo, cuarto*.
- Determinativo: Se limita de algún modo el alcance del nombre.
  - Posesivo: Indica la posesión, propiedad o pertenencia a una o varias personas o cosas de lo significado por el sustantivo a que se refiere; v. gr., *nuestra, mi, sus, tu*.
  - Gentilicio: Denota la procedencia geográfica de las personas o su nacionalidad; v. gr., *castellano, madrileño, andaluz, peruano.*

B.1.7 Verbo

Clase de palabras que pueden tener variación de persona, número, tiempo, modo y aspecto.

B.1.8 Adverbio

Clase de palabras invariables cuya función consiste en complementar la significación del verbo, de un adjetivo, de otro adverbio y de ciertas secuencias. De acuerdo a la RAE existen dos clasificaciones para los adverbios: los léxicos y los pronominales.

Los léxicos a su vez se dividen en: calificativos, de lugar, tiempo, temporales intransitivos y modales. Mientras que los pronominales se dividen en: deícticos, temporales, modo, cuantitativos, cuantitativos temporales, cuantitativos aspectuales, numerales, identificativos, identificativos polares, relativos e interrogativos/exclamativos.

B.1.9 Conjunción

Clase de palabras invariables que encabezan diferentes tipos de oraciones subordinadas o que unen vocablos o secuencias sintácticamente equivalentes. Se clasifican en los siguientes tipos:

- Copulativas.
- Disyuntivas.
- Adversativas.
- Concesiva.
- Causales.
- Condicionales.
- Comparativa.
- Consecutivas.
- Finales.
- Completiva.

B.1.10 Preposición

Clase de palabras invariantes que introducen elementos nominales u oraciones subordinadas sustantivas haciéndolos depender de alguna palabra anterior. Varias de ellas coinciden en su forma con prefijos.

B.1.11 Interjección

Clase de palabras que expresan alguna impresión súbita o un sentimiento profundo. Sirven también para apelar al interlocutor, o como forma de saludo, despedida, conformidad, etc.

B.2 La frase

Una frase es una expresión acuñada constituida generalmente por dos o más palabras, cuyo significado conjunto no se deduce de los elementos que la componen.

B.3 El sintagma

Conjunto de palabras estructuradas, relacionadas en torno a un núcleo. Generalmente, le corresponde un comportamiento sintáctico unitario. Un sintagma puede caer en cierta clase dependiendo del núcleo que lo conforme:

- Nominal.
- Verbal.
- Preposicional.
- Adjetival.
- Adverbial.

En torno al núcleo del sintagma existen otras palabras que lo detallan y complementan. Este es el caso de los determinantes, modificadores o complementos.

- El determinante tiene como función dentro del sintagma concretar o limitar la extensión de un sustantivo.
- El modificador limita el significado de una palabra.
- El complemento precisa el significado de una o varias palabras. Dependiendo de las palabras a las que afecte el complemento, éste se puede clasificar en una de las siguientes categorías:
  - Complemento referido a un núcleo no verbal.
  - Complemento referido a un núcleo verbal.
  - Complemento referido a un sustantivo y a un verbo.

B.3.1 Sintagma nominal

El sintagma nominal está construido en torno a un nombre o sustantivo. La estructura de un sintagma nominal es la siguiente:

- Núcleo: Un sustantivo, pronombre o palabra sustantivada.
- Actualizador: Pueden ser un determinante o un cuantificador.
- Complemento: El cual puede ser a su vez un sintagma nominal o en su defecto un sintagma adjetival o preposicional.

Tomando como ejemplo la oración: *Ese vestido tan bonito de tu armario*, se describirá la estructura del sintagma nominal en la Figura B.1.

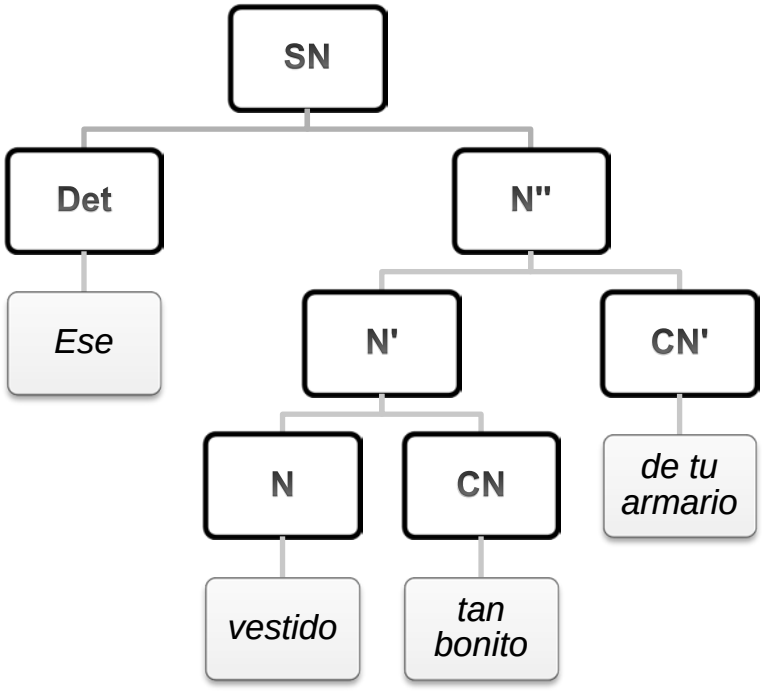

Figura B.1 Sintagma nominal

B.3.2 Sintagma verbal

El sintagma verbal está construido en torno a un verbo. La estructura de un sintagma verbal es la siguiente:

- Núcleo: El verbo en forma simple, en forma compuesta o una perífrasis verbal.
- Modificadores: Los complementos del verbo: atributo, complemento directo o complemento indirecto.

Se describirá la estructura del sintagma verbal en la Figura B.2.

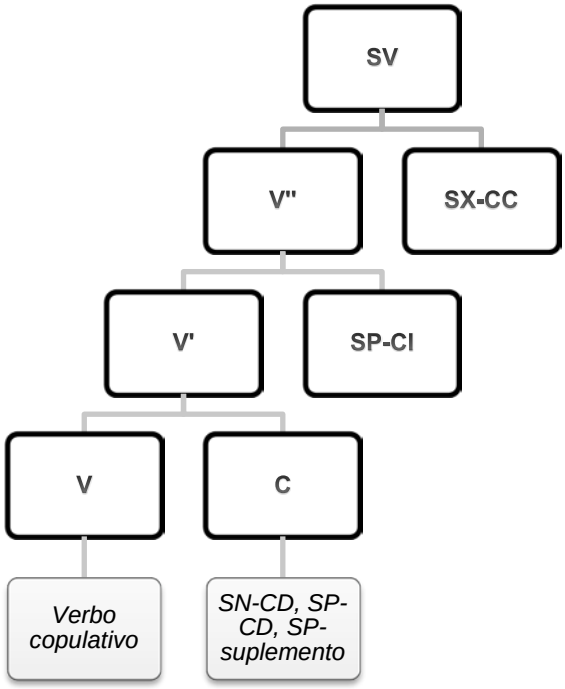

Figura B.2 Sintagma verbal

B.3.3 Sintagma preposicional

El sintagma preposicional es encabezado por una preposición. Tomando como ejemplo la frase: *En tu casa de la playa*, se describirá la estructura del sintagma preposicional en la Figura B.3.

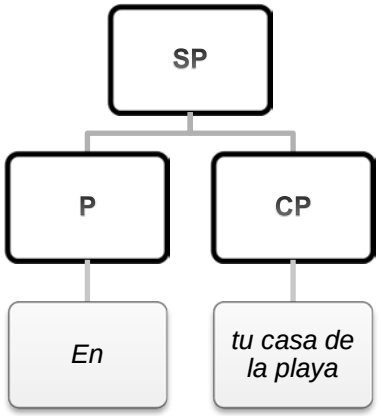

Figura B.3 Sintagma preposicional

B.3.4 Sintagma adjetival

El sintagma adjetival está construido en torno a un adjetivo. La estructura de un sintagma adjetival es la siguiente:

- Núcleo: Un adjetivo.
- Modificadores adjetivales: Pueden ser adverbios de grado; v. gr., *mucho, muy, bastante, demasiado, harto, más, menos, algo, nada, poco, un poco, medio, un tanto, tan, cuán*; o adverbios modificadores terminados en -*mente*.
- Complementos adjetivales: Suelen estar conformados por un sintagma preposicional.

Tomando como ejemplo la oración: *Muy fácil para esa gente*, se describirá la estructura del sintagma adjetival en la Figura B.4.

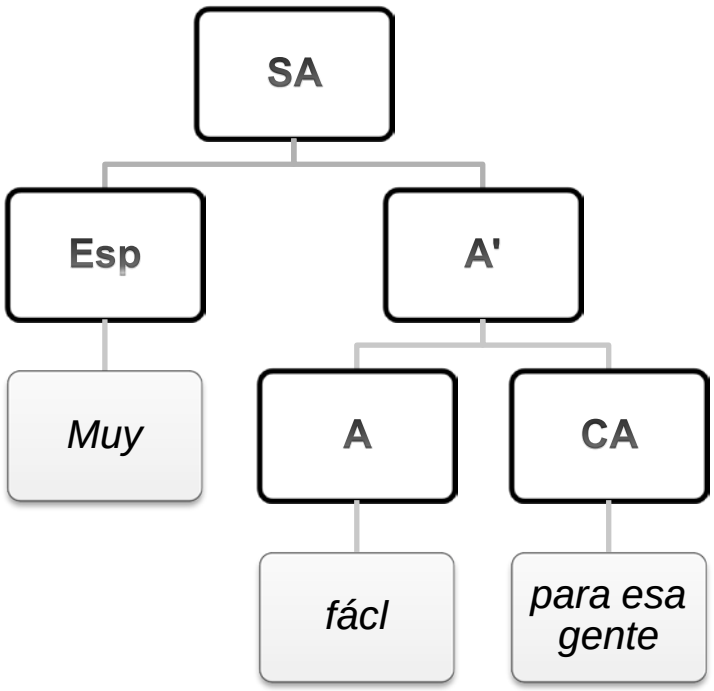

Figura B.4 Sintagma adjetival

B.3.5 Sintagma adverbial

El sintagma adverbial está construido en torno a un adverbio. La estructura de un sintagma adverbial es la siguiente:

- Núcleo: Un adverbio.
- Cuantificador: Adverbio cuantificador.
- Complemento adverbial: Sintagma preposicional que completa el significado del adverbio.

Tomando como ejemplo la oración: *Bastante cerca de mi colegio*, se describirá la estructura del sintagma adverbial en la Figura B.5.

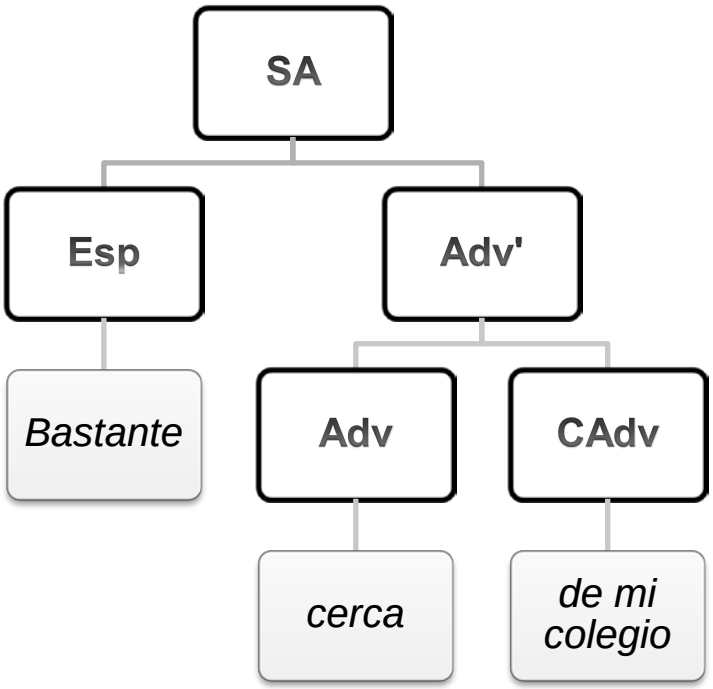

Figura B.5 Sintagma adverbial

B.4 La locución

Las locuciones son las unidades fraseológicas del sistema de la lengua, llamadas también "modismos", "frases hechas", etc. Casares en 1950 rechazaba el término modismo, por considerarlo poco claro y mal delimitado. Las denominaciones de frase hecha o expresión fija son muy amplias. Los rasgos distintivos de las locuciones son los siguientes:

- Fijación interna.
- Unidad de significado.
- Fijación pasemática (tipo de fijación externa que consiste en que determinadas unidades lingüísticas se emplean según el papel del hablante en el acto comunicativo).

Estas unidades no constituyen enunciados completos y funcionan como elementos oracionales. Son sintagmas fijos, porque no permiten la modificación, la sustitución, la adición de complementos o cualquier otra modificación de la estructura. Las locuciones se asemejan a las palabras compuestas, ambas presentan cohesión semántica y morfosintáctica. Es decir, pueden cumplir las mismas funciones sintácticas en la frase. Ni las palabras compuestas ni las locuciones permiten modificaciones parciales de sus elementos que las constituyen. No se les pueden intercalar otros elementos ni alterar su orden.

La diferencia entre las palabras compuestas y las locuciones consiste en la ortografía. Las palabras compuestas se suelen considerar como unidades léxicas formadas por la unión gráfica la que no es indispensable y acentual, y las locuciones no muestran unión ortográfica [27].

Aclarado esto, definimos una locución como una combinación fija de varios vocablos que funciona como una determinada clase de palabras [17] [14]. Se distinguen varios tipos de locuciones según su función gramatical:

- Locución adjetiva.
- Locución adverbial.
- Locución conjuntiva.
- Locución cuantificadora o intensificadora.
- Locución interjectiva.
- Locución preposicional.
- Locución pronominal.
- Locución sustantiva.

B.4.1 Compendio de locuciones

A continuación se muestran algunas de las locuciones encontradas en el lenguaje español.

- A bombo y platillo.
- A diestro y siniestro.
- A la virulé (tener un ojo).
- A ojo de buen cubero.
- A troche y moche.
- Acercar / Arrimar el ascua a su sardina.
- Aguar la fiesta.
- Ahogarse en un vaso de agua.
- Andar (se) con pies de plomo.
- Al pie de la letra.
- Ancha es Castilla.
- Apretarse el cinturón.
- Apuntarse a un bombardeo.

- Armar la de Dios es Cristo.
- Arrimar el hombro.
- Bailar el agua (a alguien).
- Caer la cara de vergüenza.
- Caérsele el alma a los pies.
- Cantar las cuarenta (al lucero del alba).
- Coger el toro por los cuernos.
- Como quien oye llover.
- Corriente y moliente.
- Cortar las alas.
- Coser y cantar (ser).
- Costar un ojo de la cara / un riñón / un huevo (vulgar).
- Dar de baja.
- Dar en bandeja.
- Dar la cara.
- Dar la espalda.
- Dar la nota.
- Dar vela en este entierro.
- De tomo y lomo.
- De uvas a peras.
- Dorar la píldora (a alguien).
- Echar balones fuera.
- Echar chispas.
- Echar las campanas al vuelo.
- Echar leña al fuego.
- Echar sal en la herida.
- Echar toda la carne en el asador.
- Echar una mano.
- En un abrir y cerrar de ojos / En un santiamén.
- Escurrir el bulto.

- Estar a verlas venir.
- Estar en Babia.
- Estar entre la espada y la pared.
- Estar entre Pinto y Valdemoro.
- Estar para comérselo.
- Hacer de tripas corazón.
- Hacer estragos.
- Hacer la vista gorda.
- Hacer leña del árbol caído.
- Hacer mutis por el foro.
- Hacer oídos sordos.
- Hacerse el sueco.
- Ir de la Ceca a la Meca.
- Ir hecho un brazo de mar.
- Irse por los cerros de Úbeda.
- Lavarse las manos.
- Levantarse con el pie izquierdo.
- Limpio de polvo y paja (estar algo).
- Llegar y besar el santo.
- Llevar a alguien por la calle de la Amargura.
- Llevar la batuta.
- Llevar la voz cantante.
- Llevar las de perder.
- Matar el gusanillo.
- Meter baza.
- Meter la pata.
- Montar en cólera.
- Montar un pollo.
- Morderse la lengua.
- Nadar y guardar la ropa (saber).

- No andarse con chiquitas.
- No caberle a alguien el corazón en el pecho.
- No dar pie con bola.
- No dar palo al agua.
- No llegarle a uno la camisa al cuerpo.
- No tener dónde caerse muerto.
- No ver tres en un burro.
- No tener ni pies ni cabeza.
- Pasar las de Caín.
- Pedir peras al olmo.
- Perder la cabeza.
- Pisar los talones (a alguien).
- Poner algo entre comillas.
- Poner el cascabel al gato.
- Poner las peras al cuarto.
- Poner los puntos sobre las íes.
- Poner una pica en Flandes.
- Por arte de birlibirloque.
- Quedar en agua de borrajas.
- Ser a la buena de Dios.
- Ser de rompe y rasga.
- Ser (el) cabeza de turco.
- Ser carne de cañón.
- Ser del año de la pera / de Maricastaña / de la Polca.
- Ser la repanocha.
- Ser más feo que Picio.
- Ser más papista que el Papa.
- Ser un lobo con piel de cordero.
- Ser una mosquita muerta.
- Ser pan para hoy y hambre para mañana.

- Sin tapujos (hacer algo).
- Tener azogue.
- Tener dos dedos de frente.
- Tener el baile de San Vito.
- Tener en la punta de la lengua.
- Tener los pies en el suelo.
- Tener manga ancha.
- Tener mano izquierda.
- Tener (buen) ojo.
- Tener un nudo en la garganta.
- Tirar la casa por la ventana.
- Tomar las de Villadiego.
- Valer un Potosí.
- Vestir un santo para desnudar a otro.
- Ver las estrellas.
- Ver los toros desde la barrera.

B.5 La oración

Una oración está definida como un conjunto de palabras con que se expresa un sentido gramatical completo. La estructura básica de una oración está dada por un sujeto y un predicado [17].

Las oraciones pueden ser clasificadas en base al enfoque que denoten, como se muestra a continuación [14]:

- Oración enunciativa: Comunica un hecho positivo o negativo.

    - Afirmativa; v. gr., *Luis corre a la casa*.
    - Negativa; v. gr., *Luis no corre a la casa*.
- Oración interrogativa: Realiza una pregunta; v. gr., *¿Cocinas conmigo?*
- Oración imperativa: Denota un mandato, exhortación o ruego.
    - Verbo en imperativo; v. gr., *Recoge el tiradero*.
    - Verbo en presente; v. gr., *¡Ahora limpias!*
    - Verbo en futuro; v. gr., *Mañana leerás el libro*.
    - Verbo en subjuntivo (usando una negativa); v. gr., *No comas eso*.
    - Verbo en infinitivo; v. gr., *¡A callar!*
- Oración exclamativa: Denota emoción, alegría, asombro, etc.; v. gr., *¡Yo adoro el helado!*
- Oración dubitativa: Manifiesta una duda o inseguridad acerca del mensaje que se desea transmitir.
    - Uso de subjuntivos y/o adverbios de duda o probabilidad; v. gr., *Quizá llueva pronto*.
    - En forma de pregunta a uno mismo; v. gr., *¿Qué pasará en casa?*
- Oración personal: Aquélla que tiene un sujeto.
    - Explícito; v. gr., *Tu tía llamará pronto*.
    - Implícito; v. gr., *Corrieron en el parque*.
- Oración impersonal: Aquélla que no tiene sujeto.
    - Con un verbo que exprese un sentido meteorológico; v. gr., *Ayer nevó poco*.
    - Haciendo referencia a fenómenos naturales o el paso del tiempo; v. gr., *Pronto amanecerá*.
    - Tercera persona del singular del verbo haber; v. gr., *No hay pudor*.
    - Haciendo uso del pronombre *se*; v. gr., *Se espera el arribo del ministro*.
- Oración activa: El sujeto realiza una acción; v. gr., *Los estudiantes cumplen con sus tareas*.
- Oración pasiva: El sujeto recibe una acción.

    - Pasiva simple, la cual usa la ayuda del verbo auxiliar *ser*; v. gr., *Los deberes son desempeñados por los pasantes*.
    - Pasiva refleja, tiene significado pasivo pero forma activa, se incorpora el uso del pronombre *se*; v. gr., *Las decoraciones se terminaron*.
- Oración reflexiva: El sujeto realiza y recibe la acción, la oración es construida con los pronombres reflexivos: *me, te, se, nos, os, se*.
    - El pronombre realiza la función de complemento directo; v. gr., *Jesús se asea*.
    - El pronombre realiza la función de complemento indirecto; v. gr., *Jesús se asea la cara.*
- Oración recíproca: Puede tener varios sujetos y la acción es intercambiada entre cada uno de ellos, se construye con los pronombres *nos, os, se*.
    - El pronombre realiza la función de complemento directo; v. gr., *Pedro y Lucía se escriben*.
    - El pronombre realiza la función de complemento indirecto; v. gr., *Pedro y Lucía se escriben largas cartas*.
- Oración atributiva: Está construida por los verbos *ser, estar, parecer*, también denominados copulativos. La única función de dichos verbos es unir al sujeto con su atributo; v. gr., *Su tía es simpática*.
- Oración predicativa: Carece de verbo copulativo.
    - Transitiva, es acompañada de su objeto directo; v. gr., *Luis esconde los regalos*.
    - Intransitiva, no tiene complemento directo; v. gr., *Este invierno esquiaremos en Holanda*.

Cabe aclarar que una oración puede tener una o más clasificaciones; por ejemplo, la oración: *Su tía es simpática*; es atributiva y también es intransitiva. Sin embargo, también hay clases que son mutuamente excluyentes; v. gr., transitiva e intransitiva, ya que no es posible que una oración transitiva sea a su vez intransitiva.

ANEXO C: DISEÑO DEL LEXICÓN

El siguiente anexo contiene los agrupamientos de los distintos accidentes gramaticales de cada una de las categorías gramaticales. Estos agrupamientos fueron realizados con el fin de obtener un diseño para la implementación de la base de datos de un lexicón, el cual contendrá todas las palabras reconocidas por la RAE.

Otro objetivo de este diseño es el de contar con toda la información léxica de cada una de las palabras del lenguaje español. Dicha información podrá ser utilizada para verificar congruencias de género, número, persona verbal, tiempo verbal, etc.; y procesar excepciones a las reglas gramaticales.

Tabla C.1 Agrupamiento de accidentes gramaticales de los artículos (1 de 2)

Número		**Género**	
Etiqueta	**Significado**	**Etiqueta**	**Significado**
S	Singular	M	Masculino
P	Plural	F	Femenino
I	Invariable	N	Neutro
		E	Pendiente especificar

Tabla C.2 Agrupamiento de accidentes gramaticales de los artículos (2 de 2)

Tipo		**Complejidad**	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
D	Definido	S	Simple
I	Indefinido	C	Contracto

Tabla C.3 Agrupamiento de accidentes gramaticales de los pronombres (1 de 3)

Persona		Tipo	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
1S	1ra. persona singular	D	Demostrativo
1P	1ra. persona plural	P	Personal
2S	2da. persona singular	N	Numeral
2P	2da. persona plural	I	Indefinido
2aS	2da. persona alterna singular	Ip	Impersonal pasivo
2aP	2da. persona alterna plural	R	Relativo
3S	3er. persona singular	Ir	Interrogativo
3P	3er. persona plural	Po	Posesivo
E	Pendiente especificar	E	Pendiente especificar

Tabla C.4 Agrupamiento de accidentes gramaticales de los pronombres (2 de 3)

Género		Número	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
F	Femenino	S	Singular
M	Masculino	P	Plural
N	Neutro	I	Invariable
E	Pendiente especificar		

Tabla C.5 Agrupamiento de accidentes gramaticales de los pronombres (3 de 3)

Pronombres personales		Caso en los pronombres	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
A	Átonos	A	Acusativo
T	Tónicos	D	Dativo
		E	Pendiente especificar

Tabla C.6 Agrupamiento de accidentes gramaticales de los sustantivos (1 de 2)

Género		Número	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
F	Femenino	S	Singular
M	Masculino	P	Plural
N	Neutro	I	Invariable
E	Pendiente especificar		

Tabla C.7 Agrupamiento de accidentes gramaticales de los sustantivos (2 de 2)

Tipo	
ETIQUETA	**SIGNIFICADO**
C	Común
P	Propio
N	Neutro
E	Pendiente especificar

Tabla C.8 Agrupamiento de accidentes gramaticales de los adjetivos (1 de 2)

Persona		Tipo	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
1S	1ra. persona singular	C	Cardinal
1P	1ra. persona plural	O	Ordinal
2S	2da. Persona singular	Cf	Calificativo
2P	2da. persona plural	I	Indefinido
2aS	2da. persona alterna singular	P	Posesivo
2aP	2da. persona alterna plural	D	Determinativo
3S	3er. persona singular	G	Gentilicio
3P	3er. persona plural		
E	Pendiente especificar		

Tabla C.9 Agrupamiento de accidentes gramaticales de los adjetivos (2 de 2)

Género		Número	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
F	Femenino	S	Singular
M	Masculino	P	Plural
N	Neutro	I	Invariable
E	Pendiente especificar		

Tabla C.10 Agrupamiento de accidentes gramaticales de los verbos (1 de 3)

Persona		Tipo	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
1S	1ra. persona singular	I	Infinitivo
1P	1ra. persona plural	G	Gerundio
2S	2da. persona singular	P	Participio
2P	2da. persona plural		
2aS	2da. persona alterna singular		
2aP	2da. persona alterna plural		
3S	3er. persona singular		
3P	3er. persona plural		
E	pendiente especificar		

Tabla C.11 Agrupamiento de accidentes gramaticales de los verbos (2 de 3)

Modo verbal		Tiempo verbal	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
I	Indicativo	P	Presente
S	Subjuntivo	I	Imperfecto
Im	Imperativo	F	Futuro
E	Pendiente especificar	C	Condicional
		E	Pendiente especificar

Tabla C.12 Agrupamiento de accidentes gramaticales de los verbos (3 de 3)

Construcción		Tipo de verbo	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
T	Transitivo	P	Principal
I	Intransitivo	A	Auxiliar
P	Pronominal		

Tabla C.13 Agrupamiento de accidentes gramaticales de los adverbios.

Persona		Tipo	
ETIQUETA	**SIGNIFICADO**	**ETIQUETA**	**SIGNIFICADO**
C	Calificativo	D	Deícticos
L	Lugar	Te	Temporales
T	Tiempo	Mo	Modo
N	Nominal	C	Cuantitativo
TI	Temporal intransitvo	CT	Cuantitativo temporal
M	Modal	CA	Cuantitativo aspectual
		N	Numeral
		I	Identificativo
		IP	Identificativo polar
		R	Relativo
		IE	Interrogativo/exclamativo

ANEXO D: CORPUS DE CONSULTAS

El corpus de consulta de CFA es un extracto obtenido a partir de los corpus de consultas de las bases de datos de Northiwnd [21], Pubs [21] y Geobase [22].

Características de las bases de datos:

- Northwind
  - Base de datos de control de inventarios, contiene elementos como: artículos, órdenes, empleados, zonas de trabajo, etc.
- Pubs
  - Base de datos de registro de publicaciones, contiene datos como: nombres de libros, ISBNs, autores, fechas de publicaciones, editoriales, etc.
- Geobase
  - Base de datos que contiene información geográfica sobre Estados Unidos de América, contiene datos como: estados, ciudades, número de habitantes, carreteras, ríos, lagos, etc.

D.1 Corpus de consultas CFA

1. ¿Cuántos vuelos tiene cada aerolínea?
2. ¿Cuántos pasajeros tiene cada vuelo de Boston a San Francisco?
3. Lista el número de pasajeros de cada vuelo.
4. Lista el número de gente en cada vuelo.
5. ¿Cuantos ejemplares del libro The Busy se vendieron el 14 de Septiembre?
6. ¿Cuántas ventas se realizaron en el año de 1992?
7. ¿Cuántos autores son de la ciudad de Berkeley?
8. ¿Cuál es el número de ventas realizadas el 14/09/1994?
9. Dame el número de libros vendidos el 13/19/1994.
10. ¿Cuál es el libro más barato de tipo Business?

11. Proporcióname el número de productos que se hicieron en la orden 10248.
12. ¿Cuál es el promedio de población por estado?
13. ¿Cuántos estados hay?
14. ¿Cuántos estados hay en Colorado?
15. ¿Cuántos ríos hay en Chicago?
16. ¿Cuántas ciudades hay en US?
17. ¿Cuántas ciudades hay en Louisiana?
18. ¿Cuántos estados tienen frontera con Iowa?
19. Número de estados fronterizos del estado con capital Boston.
20. Número de estados fronterizos del estado más grande.
21. ¿Cuál es el punto más alto de cada estado cuyo punto bajo es el nivel del mar?
22. ¿Cuál es la ciudad más grande en un estado con río?
23. ¿Cuál es la mayor capital?
24. ¿Cuál es el estado más grande?
25. ¿Cuál es la ciudad más grande en Kansas?
26. Número de personas que viven en el estado más grande.
27. ¿Qué estado tiene la mayor población?
28. ¿Cuál es el río más corto?
29. Promedio de edad de las mujeres de la muestra FESI 2010.
30. Promedio de glucosa por sexo de estudiantes de entre 20 y 22 años.
31. Total de estudiantes por muestra por sexo.

El corpus de consultas de ATIS es un extracto obtenido, el cual fue traducido al español a partir del original [23], el cual se encuentra en idioma inglés. La base de datos de ATIS maneja información acerca de aerolíneas, contiene datos como: Ciudades, aeropuertos, aviones, escalas, vuelos, tarifas, itinerarios, etc.

D.2 Corpus de consultas de ATIS

1. ¿Puedes decirme la tarifa para el vuelo número 16?
2. Costo de vuelo del vuelo número 144165.
3. ¿Puedo ver la aerolínea y el número de vuelo que saldría de San Francisco?
4. Primera clase aerolíneas Delta de Forth Worth a Philadelphia.
5. Dame una lista de todos los tipos de aeronaves.
6. Dame una lista de todas las aeronaves disponibles, la capacidad y el peso.
7. Dame una lista de todos los tamaños de equipo y velocidad.
8. Me gustarían las tarifas de vuelo para el vuelo 11 y 12.
9. Me gustaría la lista de precios del vuelo número 3.
10. Lista todas las restricciones de vuelos.
11. Lista todos los tipos de transporte terrestre.
12. Lista capacidad de equipo.
13. Lista tarifas de viaje redondo.
14. Lista los aeroplanos.
15. Lista las categorías de aeroplanos.
16. Lista las ciudades.
17. Lista las clases de códigos de vuelos.
18. Lista clases de servicio de vuelos.
19. Lista las descripciones de transporte.
20. Lista los tipos de aeroplanos para vuelos desde Fort Worth a Washington.
21. Nombra todos los aeropuertos.
22. Muéstrame el costo del vuelo 9.
23. Lista todos los vuelos.
24. Muéstrame el costo de clase Business para el vuelo número 1.

25. Tarifa de viaje redondo para el vuelo desde Atlanta.
26. Muéstrame los vuelos desde Oakland a Baltimore llegando después del mediodía.
27. Lista las tarifas para todos los vuelos saliendo después de las 1200 desde Boston a Baltimore.
28. Muestra los viajes que sólo salen de San José.
29. Dame una lista de todos los vuelos desde Dallas a Boston que lleguen antes de 7:00 am.
30. ¿Cuánto cuestan los vuelos número 1, 2, 3, 4, 5?
31. ¿Cuánto cuestan los vuelos desde Atlanta a San Francisco?
32. ¿Cuánto cuesta el vuelo número 90 y 888 desde Denver a Dallas Fort Worth?
33. ¿Cuánto cuesta volar desde Boston a Oakland sencillo?
34. ¿Cuánto cuesta un viaje redondo desde Boston a Dallas?
35. Necesito vuelos que lleguen antes del mediodía.
36. Lista todos los vuelos saliendo después del mediodía y llegando antes de las 7:00 pm.
37. Lista todos los vuelos saliendo de Denver a Pittsburgh después de las 5:00 pm y lista las tarifas.
38. Lista sólo vuelos llegando antes de las 7:00 pm.
39. Lista sólo vuelos saliendo de San Francisco.
40. Lista los vuelos que salen desde San Jóse California.
41. Sólo muéstrame los vuelos saliendo de San Francisco.
42. Lista sólo vuelos de clase económica saliendo después del mediodía.
43. Muéstrame vuelos que salgan después del mediodía.
44. Muéstrame las aerolíneas que vuelan desde Dallas a Denver.
45. Despliega el tiempo de salida.
46. Lista de todos los vuelos desde Atlanta a Boston.
47. Dame una lista de los vuelos desde Denver a San Francisco.
48. Lista todos los vuelos desde Denver a Pittsburg y lista las tarifas.
49. ¿Puedo ver vuelos desde San Francisco a Los Ángeles?

50. Muestra todos los vuelos y tarifas desde Fort Worth a Denver.
51. Vuelos desde Forth Worth a Philadelphia en Delta Airlines.
52. Vuelos desde San Francisco a Dallas.
53. ¿Desde Oakland a Boston qué tarifa es?
54. Da todos los vuelos desde Dallas a Boston a Denver.
55. Dame una lista de todos los vuelos desde Philadelphia a Atlanta y desde Atlanta a Dallas.
56. Dame la tarifa en clase Q desde Dallas a Atlanta.
57. Encuentra el costo de un vuelo sencillo desde Pittsburgh a Oakland.
58. Dame una tarifa de viaje redondo desde Atlanta a Baltimore.
59. Lista todos los vuelos desde Oakland a San Francisco mostrando los precios.
60. Lista vuelos desde Atlanta a San Francisco.
61. Lista tarifas de viaje redondo desde Fort Worth a Atlanta.
62. ¿Puedo tener una lista de tarifas desde Atlanta a Boston?
63. Tarifa sencilla desde Washington a Atlanta.
64. ¿Qué vuelos vespertinos están disponibles desde Washington a Boston con comidas?
65. Dame una lista de vuelos desde Philadelphia a Baltimore, en la mañana.
66. Dame los precios para todos los vuelos desde Dallas a Boston en la mañana.
67. Lista vuelos de tarde desde Atlanta a San Francisco.
68. Lista todos los vuelos saliendo después del mediodía y llegando antes de las 7:00 pm.
69. Lista todos los vuelos desde Dallas a Boston en la mañana.
70. Muéstrame la lista de vuelos desde Dallas a Denver en la mañana y muestra su costo.

El corpus de consultas de Pubs fue tomado de un trabajo realizado en el CENIDET en el año 2005 a cargo de Javier González [21]. La base de datos Pubs contiene información sobre el registro de publicaciones, tiene datos como: nombres de libros, ISBNs, autores, fechas de publicaciones, editoriales, etc.

D.3 Corpus de consultas de Pubs

1. Dame los títulos de los libros.
2. Visualiza los tipos de descuentos.
3. ¿Cuál es la dirección de la editorial y su ciudad?
4. ¿Cuál es el título del libro con identificador TC4203?
5. Selecciona el titulo donde el precio sea igual a $19.99 y el tipo sea bussines.
6. Lista los empleados con su respectivo cargo.
7. ¿Qué puesto ocupa cada empleado?
8. ¿A qué almacén pertenece la siguiente dirección 679 Carson st.?
9. ¿Qué empleado tiene como identificador H-B39728F?
10. ¿A qué almacén corresponde el identificador 7131?
11. ¿Qué autores viven en la ciudad de Oakland?
12. ¿Qué trabajador tiene como nivel de trabajo 227?
13. ¿Qué trabajador tiene su fecha de contratación como 13/02/1991?
14. ¿Qué libros son del tipo bussines?
15. ¿Quién es el autor del título The Busy?
16. Mostrar los libros cuyo precio es mayor a $19.99 y son de tipo bussines.
17. ¿Qué editorial se encuentra en Alemania?
18. ¿Quién es el autor del libro The Gourmet?
19. ¿Quién es el autor del libro The busy?
20. Selecciona todos los libros del autor Smith.
21. ¿Qué libros son de la editorial Algodata Infosystems?
22. Dame los títulos del autor Green.
23. ¿Qué puesto tiene el empleado Francisco?
24. ¿Qué puesto tiene el empleado Francisco Chang?

25. Selecciona el descuento para el almacén 8042.
26. ¿Qué puesto tiene el empleado Paolo y su fecha de contratación?
27. Dime los libros que fueron vendidos en la fecha 13/09/2004 y que son diferentes del tipo business.
28. ¿Cuál es la dirección del almacén Barnum's?
29. ¿Cuál es el número de teléfono del autor Cheryl?
30. ¿Cuál es la ciudad de la editorial New Moon Books?
31. ¿Cuál es el identificador del empleado Paolo Accort?
32. ¿Cuál es el nivel de trabajo de Philip Cramer?
33. ¿Cuál es el precio del identificador de editor 1389?
34. ¿Cuáles son los títulos de la editorial GGG&G?
35. ¿Cuál es la clave y el precio del libro You Can?
36. ¿Cuál es la dirección y el teléfono del autor del libro You Can?
37. ¿En qué ciudad se encuentra el autor Jonson White?
38. ¿Qué apellido tiene el empleado Pedro?
39. ¿En qué ciudad se encuentra el almacén Bookbeat?
40. ¿En qué ciudad se encuentra la editorial Lucerne Publishing?
41. ¿En qué estado se encuentra la editorial Ramona Publishers?
42. ¿En qué ciudad se ubica la tienda Erick The Read Books?
43. Titulo de los libros cuyos editores se encuentran en Texas.
44. ¿Qué nombre y dirección tiene el empleado que trabaja para la editorial GGG&G?
45. ¿Qué descripción tiene el puesto del empleado VPA30890F?
46. Obtener el nombre del almacén donde se encuentra el libro *Cooking With*.
47. ¿En qué fecha se realizó el contrato del empleado PTC11962M?
48. ¿Cuántos autores son de la ciudad de Berkeley?
49. ¿Cuál es el número de empleados de la editorial Scotney Book?
50. ¿Cuál es el número de ventas realizadas el 14/09/1994?
51. Dame el número de libros vendidos el 13/09/1994.
52. ¿Qué identificador tienen los títulos?
53. ¿A qué ciudad pertenece el código postal 89076?

54. ¿A qué ciudad corresponde la dirección 567 Pasadena Ave?
55. ¿Cuál es el adelanto del número de orden 6871?
56. Dame la fecha de contratación de Pedro.
57. ¿Qué títulos contiene cada editorial?
58. ¿Qué cantidad de Silicon Valley es vendida?
59. ¿En qué editorial trabaja Victoria Ashworth?
60. Nombre del almacén donde se encuentra The Busy.
61. ¿Cuántos números de ejemplares tiene el libro The Busy?
62. ¿Cuántos ejemplares del libro The Busy se vendieron el 14 de septiembre?
63. Todos los empleados que tengan un puesto.
64. ¿Cuál es el nombre de la editorial que no tenga estado y a su vez no esté en USA?
65. ¿Cuál es el título del libro que no tiene precio?
66. ¿Cuántas ventas se realizaron en el año de 1992?
67. Selecciona todas las editoriales del mismo país.
68. ¿Cuál es el libro más barato de tipo Business?
69. ¿Quién es el empleado que tiene más tiempo trabajando?

REFERENCIAS

[1] Marco Aguirre and Rodolfo Pazos, "Semantic model for improving the performance of natural language interfaces to databases," Instituto Tecnológico de Ciudad Madero, Cd. Madero, Tam., México, 2011.

[2] Antonio Cervantes, "Analizador sintáctico de oraciones en español usando el método de dependencias," CENIDET, Centro Nacional de Investigación y Desarrollo, Cuernavaca, Morelos, tesis de maestría 2005.

[3] Eckhard Bick, "A constraint grammar parser for spanish," Institute of Language and Communication, University of Southern Denmark, 2006.

[4] Alexander Gelbukh, "Procesamiento de lenguaje natural y sus aplicaciones," *Komputer Sapiens*, vol. 1, pp. 6-11, Enero 2010.

[5] Sofía Galicia, "Análisis sintáctico conducido por un diccionario de patrones de manejo sintático para lenguaje español," Instituto Politécnico Nacional, tesis de doctorado 2000.

[6] Sofía Galicia y Alexander Gelbukh, "Análisis sintáctico para el español basado en el formalismo de la teoría significado-texto," *Procesamiento del lenguaje natural*, no. 29, pp. 81-88, 2002. [Online]. http://hdl.handle.net/10045/1681

[7] Luis Losada, "Automatización del análisis sintáctico del español," 2003. [Online]. http://hdl.handle.net/10045/1594

[8] Kepa Bengoetxea, "Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera," 2007. [Online]. http://hdl.handle.net/10045/2932

[9] Jordi Carrera e Irene Castellón, "Gramáticas de dependencia en Freeling," *Procesamiento del Lenguaje Natural*, no. 41, pp. 21-28, 2008. [Online]. http://hdl.handle.net/10045/8056

[10] Elisabet Comelles, "Constituency and dependency parsers evaluation," Sociedad Española para el Procesamiento del Lenguaje Natural, Universidad de Barcelona, ELDA, Barcelona, España; Paris, Francia, 2010. [Online]. http://hdl.handle.net/10045/14706

[11] Analizador sintáctico con gramática de constituyentes, Septiembre 2012, http://clic.ub.edu/es/sintaxis.

[12] TUSIR, Noviembre 2012, gplsi.dlsi.ua.es/gplsi11/content/tusir-0.

[13] 3LB, Noviembre 2012, gplsi.dlsi.ua.es/gplsi11/content/3lb.

[14] Ignacio Bosque y Violeta Demonte, *Gramática descriptiva de la lengua española*, Real Academia Española, Ed.: ESPASA, 2000.

[15] Ruslan Mitov, *The oxford handbook of computational linguistics*, 1st ed.: Oxford, 2003.

[16] Michael Sipser, *Introduction to the theory of computation*, Segunda edición ed.: Thomson, 2006.

[17] Real Academia Española, Enero 2013, lema.rae.es/drae/.

[18] Antonio García Mejía, Análisis sintáctico y morfológico, 2004, http://angarmegia.com/analisis_gramatical.htm.

[19] Paul Deitel, *C# 2008 for Programmers*.: Pearson Education, 2008.

[20] Hans van de Koot, "Word grammar recognition is NP-hard," UCL Working Papers in Linguistics, 1992.

[21] Javier González, "Traductor de lenguaje natural español a SQL para un sistema de consultas de bases de datos," CENIDET, tesis doctoral 2005.

[22] Corpus de Geobase. (2013) Universidad de Texas. [Online]. ftp://ftp.cs.utexas.edu/pub/mooney/n1-ilp-data/restsystem/restqueries250

[23] Linguistic Data Consortium. (2013) Universidad de Pennsylvania. [Online]. http://catalog.ldc.upenn.edu/docs/LDC93S4B/tr_prmp.html

[24] Lorna Balkan, Klaus Netter, Doug Arnold, "Test suites for natural language processing," *Proceedings of Language Engineering Convention*, Julio 1994.

[25] Carlos Rojas, "Administrador de diálogo para una interfaz de lenguaje natural a bases de datos," CENIDET, tesis doctoral 2009.

[26] Manuel Sanz, *Procesamiento del lenguaje natural: presente y perspectivas futuras*.: Universidad de Alicante, 2003.

[27] Sofía Galicia y Alexander Gelbukh, *Investigaciones en análisis sintáctico para el español*, Primera ed. D.F., México: Instituto Politécnico Nacional, 2007.

[28] Magdaléna Holečková, "*Locuciones adverbiales en el español*", Departamento de Lenguas y Literaturas Románicas, Universidad de Masaryk en Brno, 2007.