EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLOGICO DE CIUDAD MADERO

DIVISION DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

MAESTRIA EN CIENCIAS DE LA COMPUTACION

"POR MI PATRIA Y POR MI BIEN"

TESIS

ANALIZADOR SINTÁCTICO DE INGLÉS PARA UNA INTERFAZ

DE LENGUAJE NATURAL A BASES DE DATOS

Que para obtener el Grado de
Maestro en Ciencias de la Computación

Presenta
ITIC Julieta Vanelly Orta Camacho
G19073016
No. CVU de CONACyT 1007979

Director de Tesis
Dr. José Antonio Martínez Flores
No. CVU de CONACyT 202705

Codirector de Tesis
Dr. Rodolfo Abraham Pazos Rangel

Cd. Madero, Tamaulipas, México Diciembre, 2021

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Ciudad Madero
Subdirección Académica
División de Estudios de Posgrado e Investigación

Cd. Madero, Tam. 06 de diciembre de 2021

OFICIO No. : U.169/21
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

C. JULIETA VANELLY ORTA CAMACHO
No. DE CONTROL G19073016
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, se acordó autorizar la impresión de su tesis titulada:

"ANALIZADOR SINTÁCTICO DE INGLÉS PARA UNA INTERFAZ DE LENGUAJE NATURAL A BASES DE DATOS"

El Jurado está integrado por los siguientes catedráticos:

PRESIDENTE:	DR.	JUAN JAVIER GONZÁLEZ BARBOSA
SECRETARIA:	DRA.	CLAUDIA GUADALUPE GÓMEZ SANTILLÁN
VOCAL:	DR.	JOSÉ ANTONIO MARTÍNEZ FLORES
SUPLENTE:	DR.	RODOLFO ABRAHAM PAZOS RANGEL
DIRECTOR DE TESIS:	DR.	JOSÉ ANTONIO MARTÍNEZ FLORES
CO-DIRECTOR:	DR.	RODOLFO ABRAHAM PAZOS RANGEL

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
Excelencia en Educación Tecnológica®
"Por mi patria y por mi bien"®

MARCO ANTONIO CORONEL GARCÍA
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE CIUDAD MADERO
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

c.c.p.- Archivo
MACG 'jar'

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos,
C.P. 89440 Cd. Madero, Tam. Tel. 01 (833) 357 48 20, ext. 3110
e-mail: depi_cdmadero@tecnm.mx
tecnm.mx | cdmadero.tecnm.mx

ESPACIO 100% LIBRE DE PLÁSTICO

México 2021 Año de la Independencia

Contenido

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

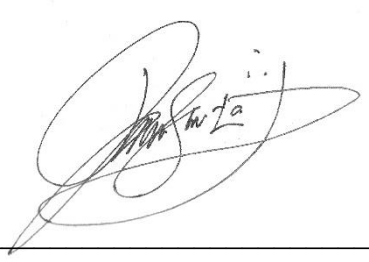
Declaración de originalidad

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 ajenas, indico explícitamente los datos de los autores y las publicaciones.

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

2 de diciembre de 2021, Cd. Madero, Tamps.



ITIC Julieta Vanelly Orta Camacho

Dedicatoria

El presente trabajo de tesis de maestría la dedico a mi madre Julieta Camacho Martínez y a mi padre Miguel Ángel Orta Castro por ser mi inspiración, mi motivación y mi orgullo. Gracias por enseñarme que ningún sueño es imposible de alcanzar y por brindarme siempre las oportunidades, sus consejos, los recursos y principalmente la confianza para lograr todas mis metas.

A mi abuelita Cástula Castro Damez por su amor, apoyo y ser un ejemplo de fortaleza en mi vida.

Especialmente al Ing. Raúl Salvador Medina Alatorre por acompañarme y apoyarme a lo largo de este camino y crear momentos inolvidables.

Agradecimientos

Deseo expresar mi profundo agradecimiento a mis padres Julieta y Miguel, y a Raúl por motivarme a ser una mejor persona cada día.

A los miembros de mi comité tutorial de tesis: Dr. José Antonio Martínez, Dr. Rodolfo Abraham Pazos Rangel, Dra. Claudia Guadalupe Gómez Santillán, Dra. Guadalupe Castilla Valdez y Dr. Juan Javier González Barbosa. Por sus observaciones, sugerencias y tiempo, las cuales fueron de gran ayuda en el desarrollo de esta tesis.

Al Dr. José Antonio Martínez y al Dr. Rodolfo Abraham Pazos Rangel por su tiempo y apoyo para dirigir este proyecto.

A la Dra. Claudia Guadalupe Gómez Santillán por sus consejos y orientación.

Agradezco a mis compañeros de generación Melissa Castillo Pérez, Eduardo David Martínez Hernández, Manuel Barrón Santiago, Luis Mario Velasco Ocejo, Arsenio Jesús Lizardi Duran y Alejandro Castellanos Álvarez, por su amistad, apoyo y confianza brindada a lo largo de éste trabajo.

A todas las personas y nuevos amigos que conocí en el Instituto Tecnológico de Ciudad Madero en especial a Sandra Gonzales de la Cruz, Saúl Mata Alvarado, Cintia Dennyse Ibarra Hernandez y Andrés Adolfo Verástegui Ollervides.

Quiero expresar mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo brindado para llevar a cabo este proyecto.

Resumen

El principal objetivo de las interfaces de lenguaje natural a bases de datos es simplificar el acceso a la información almacenada en BDs para que cualquier usuario tenga acceso a ella. El Instituto Tecnológico de Ciudad Madero (ITCM) cuenta con una interfaz en español que ha mostrado tener un buen funcionamiento, descrita en [Verástegui, 2020].

Esta interfaz cuenta con un nuevo Analizador Sintáctico-Semántico (AS-S), el cual involucra en sus reglas de producción información semántica y sintáctica. Esto le permite dar tratamiento a valores de búsqueda de difícil detección como los valores de búsqueda que involucran columnas booleanas y palabras del idioma, reduciendo los tiempos de procesamiento.

En este proyecto se prueba que los métodos propuestos en [Verástegui, 2020], para el AS-S de español, pueden ser migrados a otro idioma, específicamente el inglés. Se comprobó que los métodos pueden ser implementados para trabajar con consultas en LN en inglés, permitiendo ampliar la funcionalidad de la ILNBD desarrollada en el ITCM al lograr que trabaje con consultas en LN en español e inglés.

En ese proyecto se realizaron cuatro contribuciones principales: La **primera** de ellas es la definición de un lexicón de inglés con suficiente información que permite la implementación del analizador para la interfaz. La **segunda** contribución es la adaptación de los procesos y la implementación de un analizador léxico para las particularidades de la lengua inglesa, lo cual permite la identificación temprana de los valores de búsqueda, y da tratamiento a la ambigüedad léxica en consultas en inglés.

La **tercera** contribución es la definición de las reglas de producción del AS-S basadas en las estructuras gramaticales de la lengua inglesa. Éstas incluyen tanto información sintáctica como semántica, lo cual permite realizar reducciones que facilitan el análisis de consultas en inglés. Por último, la **cuarta** contribución fue la implementación del AS-S para el idioma inglés.

Se realizaron diversas pruebas de funcionalidad y comparativas al AS-S de inglés. Los resultados demuestran que los métodos presentados en [Verástegui, 2020] funcionan de forma adecuada con consultas en otro idioma, específicamente inglés.

Abstract

The main objective of the natural language interface to a database is to simplify the access to the data stored in DBs so that any user may have access to it. The Instituto Tecnológico de Ciudad Madero (ITCM) currently has one interface for the Spanish language that has displayed a good performance.

This interface has a new Syntactic-Semantic Parser (S-SP), which involves both semantic and syntactic information in its production rules. This allows it to handle hard to detect search values such as values that involve Boolean columns or language words, reducing the processing time.

In this project it is shown that the methods proposed in [Verástegui, 2020], for the S-SP for Spanish, can be migrated to other languages, specifically English. It was proven that the methods can be implemented to work on queries in NL in English, allowing to expand the functionality of the NLIDB developed at the ITCM by allowing it to work with NL queries in both Spanish and English.

In this project four main contributions were made: The **first** one of them is the definition of an English lexicon with enough information for allowing the implementation of the parser for the interface. The **second** contribution is adapting the processes and the implementation of the lexical analyzer for the peculiarities of the English language, which allows the early detection of search values, and treating lexical ambiguity in English queries.

The **third** contribution is defining the production rules for the S-SP based on the grammatical structures of the English language. These include both syntactic and semantic information, which allows to perform reductions that simplify the analysis of English queries. Finally, the **fourth** contribution was implementing the S-SP for the English language.

Different functional and comparative tests were performed on the English S-SP. The results show that the methods presented in [Verástegui, 2020] work adequately on queries in other language, specifically English.

CAPÍTULO 1

Introducción

En la actualidad, la tecnología forma parte vital del día a día de la sociedad. Esto se debe a que gran parte de las actividades cotidianas implican la interacción con aplicaciones y softwares, los cuales generan una gran cantidad de información que es almacenada en Bases de Datos (BDs) para su manejo y análisis.

En años recientes, la información almacenada en las BDs se ha convertido en un recurso valioso para las industrias, ya que les brinda recursos para la toma de decisiones, predicciones, etc. Sin embargo, para tener acceso a dicha información, se requiere que el usuario cuente con conocimiento especializado en un lenguaje de consultas para BDs, como *Structured Query Language* (SQL).

Para acceder a la información almacenada en las BDs, es necesario que el usuario solicite la información de tal forma que la computadora pueda interpretar la solicitud y mostrar la información requerida. Desafortunadamente, la gran mayoría de los usuarios, por ejemplo: ejecutivos, gerentes, CEOs (*Chief Executive Officers*), etc., no cuentan con el conocimiento necesario para formular consultas que les permita obtener la información requerida de las BDs, ya que los lenguajes de consulta a BDs tienen un grado de complejidad con el que la mayoría de los usuarios casuales no están familiarizados.

Desde la creación de las primeras computadoras, la forma en que los humanos interactúan con ellas ha ido evolucionando, desde tarjetas perforadas, hasta lenguajes de programación de alto nivel. El siguiente paso lógico fue lograr que estos sistemas consiguieran comprender la forma de comunicación nativa de los humanos, es decir, el lenguaje natural (LN).

Las interfaces de lenguaje natural a bases de datos (ILNBDs) surgen como una herramienta conveniente, la cual permite lograr el objetivo de que los usuarios no especializados puedan comunicarse y tener acceso a la información recopilada en las BDs a través de consultas formuladas en su lenguaje nativo.

En [Pazos, 2013] se menciona que las primeras ILNBDs surgieron en los 60s. Sin embargo, debido a las limitaciones en la tecnología de la época y a la complejidad del LN, dichas interfaces no lograron el desempeño esperado. Actualmente, los avances en la tecnología facilitan la implementación, así como la explotación de información almacenada en BDs por parte de las

ILNBDs. Sin embargo, la complejidad del LN continúa siendo la mayor limitante para esta tecnología [Verastegui, 2020].

Es importante mencionar que el LN surgió y evolucionó de forma natural y espontánea entre un grupo de humanos a través de su uso y repetición sin ninguna premeditación o planeación consciente, es decir, no existió intervención humana en términos de su diseño y reglas de funcionamiento, el cual puede tomar diferentes formas tales como el hablado, escrito o cantado [Lyons, 1991].

Por otro lado, la lingüística es la ciencia encargada del estudiar cada uno de los aspectos del LN tales como sus estructuras, su evolución histórica, su procesamiento, su funcionamiento, etc. [Halliday, 2006]. El problema principal en el estudio del LN es la dificultad de concretar una formalización completa de la misma. Es importante enfatizar este punto, ya que es precisamente esta imposibilidad de formalización la que se convierte en el problema principal del procesamiento de LN.

El objetivo principal de las ILNBDs es simplificar el acceso de los usuarios a la información almacenada en BDs, evitando la necesidad de aprender un lenguaje de consulta a BDs, ya que el usuario introduce una consulta en LN. Esta consulta es traducida por la interfaz a un lenguaje de consulta a BDs para que posteriormente la interfaz envíe la consulta traducida en SQL a un servidor de BDs para obtener la información solicitada.

Un ejemplo de este tipo de interfaces es la ILNBD que se ha estado desarrollando en el Instituto Tecnológico de Ciudad Madero (ITCM) [Aguirre, 2014], donde se menciona que una ILNBD basada en sintaxis debe contar mínimo con tres tipos de análisis principales: análisis léxico, análisis sintáctico y análisis semántico.

El análisis léxico se encarga de etiquetar cada una de las palabras (*tokens*) con su categoría gramatical. El análisis sintáctico se encarga de agrupar las palabras en frases, lo cual facilita la identificación de las relaciones entre el conjunto de palabras, facilitando la interpretación de la consulta en LN. Para finalizar, se considera que el análisis semántico es el de mayor importancia, ya que su objetivo es comprender el significado de la consulta.

En [Verástegui, 2020] se continuó trabajando en una nueva versión de la ILNBD, la cual ahora cuenta con un nuevo Analizador Sintáctico-Semántico (AS-S) que involucra en sus reglas de producción información semántica y sintáctica. Esto le permite dar tratamiento a valores de difícil detección, como los valores de búsqueda que involucran columnas booleanas y palabras del idioma, reduciendo los tiempos de procesamiento y por ende los de respuesta por parte del nuevo analizador.

El presente trabajo tiene como objetivo principal probar que los métodos desarrollados en [Verástegui, 2020], para la ILNBD en español, pueden ser migrados a otro idioma, en el caso de este trabajo el inglés, así como la adaptación e implementación de éstos a dicho idioma. En particular, la adaptación de reglas de producción sintáctico-semánticas basadas en las estructuras gramaticales de la lengua inglesa.

1.1 Antecedentes

Este proyecto forma parte del proyecto titulado "Interfaz de Lenguaje Natural a Bases de Datos en Español para Usuarios de Internet" [Aguirre, 2014], el cual comenzó a desarrollarse en el CENIDET desde 2001 y se ha continuado en el ITCM desde 2002. Su objetivo es implementar una interfaz que permita a usuarios casuales e inexpertos formular consultas a BDs, usando expresiones en un lenguaje no restrictivo como el español. En [Aguirre, 2014] se resalta que una de las características más importantes de esta ILNBD es ser independiente de dominio, es decir, con ella se puede trabajar con más de una BD.

Este proyecto cuenta con cuatro proyectos como antecedentes. El primero de ellos es el proyecto titulado **"Traductor de Lenguaje Natural Español a SQL para un Sistema de Consultas a Bases de Datos"** [González, 2005]. Algunas de sus aportaciones son las siguientes:

- Un modelo de traducción para una ILNBD.
- Técnica para que el traductor de la interfaz sea portable a cualquier dominio.
- Generación de un diccionario de domino de forma automática, usando los metadatos de la BD que será consultada.

El segundo es el proyecto de tesis de maestría titulado **"Implementación de un Analizador Sintáctico del Idioma Español para una Interfaz de Lenguaje Natural a Bases de Datos"** [Mellado, 2014], el cual incluye las siguientes aportaciones:

- Diseño e implementación de un analizador sintáctico para el idioma español.
- Diseño y creación de un compendio de 59 reglas de producción basadas en la información del idioma español proporcionada por la RAE.
- Un número reducido de reglas de producción que permiten analizar un mayor número de oraciones.
- Incremento o modificación de las reglas de producción sin la necesidad de editar el código fuente.

El tercer antecedente es el proyecto de tesis de doctorado titulado **"Modelo Semánticamente Enriquecido de Bases de Datos para su Explotación por Interfaces de Lenguaje Natural"** [Aguirre, 2014] cuyas aportaciones son las siguientes:

- Concepción y diseño de un modelo semánticamente enriquecido de bases de datos para implementar un Diccionario de Información Semántica (DIS) en una ILNBD.
- Un DIS que permite mejorar el desempeño de la ILNBD, ya que cuenta con un desempeño del 90% de consultas correctamente contestadas.
- Diseño e implementación de un módulo de traducción basado en capas de funcionalidad, como el usado en el modelo OSI para redes de comunicación, para el procesamiento de las consultas. La Figura 1.1 muestra las capas de funcionalidad de este módulo.
- Un módulo de análisis semántico que permite una mayor flexibilidad y modularidad a la ILNBD para aplicar estrategias más complejas de procesamiento.

- El desarrollo de una herramienta (*wizard*) que permite al administrador de la DB configurar el DIS de forma más sencilla de tal manera que permite realizar la traducción correcta a SQL de la consulta en LN.
- El *wizard* utiliza diferentes heurísticas para encontrar errores en los procesos de traducción, proporcionando sugerencias al administrador sobre la información que debe modificar o añadir al DIS.

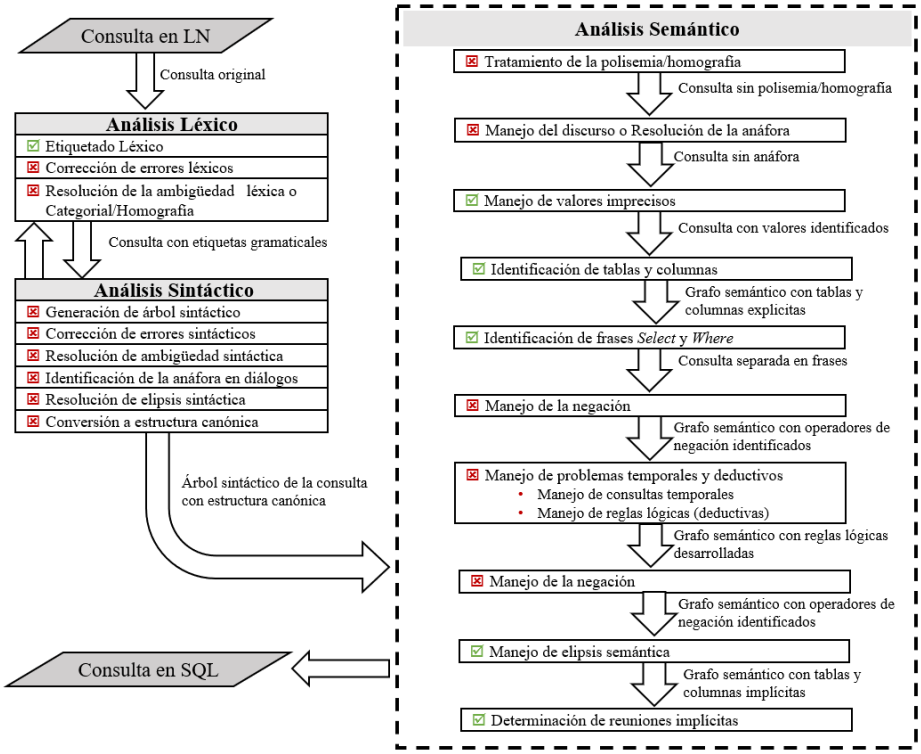


Figura 1.1. Capas de funcionalidad del módulo de traducción

- El *wizard* permite que el administrador pueda ser un usuario sin conocimientos sobre conceptos gramaticales o del funcionamiento interno de la ILNBD.

El último es el proyecto de tesis doctoral titulado **"Tratamiento de los Problemas de Valores de Búsqueda de Difícil Detección en la Traducción de Consultas de Lenguaje Natural**" [Verástegui, 2020] cuya interfaz sirve como base para este proyecto y cuenta con las siguientes aportaciones:

- Desarrollo de un analizador sintáctico exhaustivo que mostró que es imposible realizar un análisis exhaustivo en tiempo polinomial debido a la variabilidad del lenguaje español.
- Diseño e implementación de un nuevo método sintáctico-semántico.
- **Desarrollo de una versión totalmente nueva de un analizador sintáctico-semántico**, el cual utiliza reglas sintáctico-semánticas que generan reducciones o agrupaciones de palabras.
- Tratamiento a valores de búsqueda, específicamente, valores que involucran columnas booleanas y palabras del idioma en español.

- Detección de valores de búsqueda que antes eran difíciles de detectar sin la necesidad de utilizar diccionarios (lexicones).
- Detección temprana de valores de búsqueda de difícil detección, lo cual permite que el tiempo de procesamiento de éstos sea completamente independiente del tamaño de la BD.
- Traducción de consultas en LN que involucran columnas booleanas, palabras del idioma español y valores de difícil detección con una precisión del 100%, probando con el corpus *Air Travel Information Services* (ATIS).

1.2 Planteamiento del Problema

En 2020 en el ITCM, se concluyó la tesis de doctorado "Tratamiento de los Problemas de Valores de Búsqueda de Difícil Detección en la Traducción de Consultas de Lenguaje Natural" [Verástegui, 2020]. El propósito de dicha tesis consistió en mejorar el desempeño de la ILNBD de dominio independiente (que es aplicable a cualquier BD), la cual fue desarrollada en 2014 en la tesis de doctorado "Modelo Semánticamente Enriquecido de Bases de Datos para su Explotación por Interfaces de Lenguaje Natural" [Aguirre, 2014].

Es importante destacar que la última versión de la ILNBD obtiene un desempeño del 100% de consultas correctamente contestadas (consultas que involucran columnas booleanas, palabras del idioma y valores de difícil detección en español) según los resultados obtenidos con un corpus de pruebas de la BD ATIS. Esto la hace competitiva con las mejores ILNBDs independientes de dominio desarrolladas por otros investigadores [Verástegui, 2020].

Debido a que la ILNBD desarrollada en el ITCM ha mostrado ser competitiva, una forma de ampliar el rango de individuos que tengan acceso a ella es aumentando su usabilidad. El inglés es un idioma con una cobertura de usuarios más amplia que el español, tanto en el ámbito científico como el social.

El problema que se aborda en este proyecto es el desarrollo de una nueva versión para inglés del analizador sintáctico de la ILNBD del ITCM, aprovechando los métodos propuestos en [Verástegui, 2020], tomando en cuenta las particularidades del idioma inglés.

Para resolver los problemas que se enfrentaron al desarrollar la nueva versión del analizador para inglés, fue necesario realizar las siguientes actividades:

1. Familiarizarse con la versión para español del analizador sintáctico.
2. Crear un lexicón para la ILNBD que contenga lexemas del idioma inglés.
3. Analizar las estructuras de la gramática inglesa, ya que el analizador de español se basa en reglas de producción sintáctico-semánticas que incluyen información semántica en su diseño, lo cual requiere tener conocimientos de la gramática del idioma con el que se trabajó (inglés en el caso de este proyecto).
4. Adaptar a las particularidades del inglés cada uno de los procesos del análisis sintáctico, específicamente, el análisis léxico (particularmente el postprocesamiento) y el análisis sintáctico.

5. Definir reglas de producción sintáctico-semánticas que contengan información semántica de la lengua inglesa.
6. Rediseñar e implementar programas para los procesos del análisis léxico y sintáctico.
7. Probar el desempeño del analizador con las modificaciones realizadas para el idioma inglés.

1.3 Objetivos de la Investigación

Objetivo General:

OG. Desarrollar un Analizador Sintáctico-Semántico para el idioma inglés para una interfaz de lenguaje natural a bases de datos que aproveche los métodos desarrollados para el analizador del idioma español descritos en [Verástegui, 2020].

Objetivos Específicos:

OE 1. Definir y crear un lexicón para el idioma inglés.
OE 2. Estudiar el analizador desarrollado para el idioma español.
OE 3. Definir las reglas de producción sintáctico-semánticas para el idioma inglés.
OE 4. Implementar un analizador sintáctico-semántico para el idioma inglés.

1.4 Hipótesis

Las hipótesis de investigación descritas como parte del proceso de solución abordado en este proyecto son las siguientes:

H1. Es posible adaptar e implementar los métodos propuestos en [Verástegui, 2020], para un analizador de español, para desarrollar un analizador sintáctico de inglés para una ILNBD.

H2. Es posible adaptar o definir reglas de producción para un analizador sintáctico de inglés que incluyan tanto información semántica como información sintáctica, las cuales permitan el análisis de consultas en el idioma inglés.

1.5 Justificación

Todos los días se generan enormes cantidades de información útil para la toma de decisiones, la cual es almacenada en BDs para un manejo más eficiente. Esta situación genera la necesidad de crear un lenguaje, aplicación o interfaz de consultas simple y eficaz que permita que usuarios sin conocimientos de lenguajes formales de consultas, como SQL, tengan acceso a la información almacenada en las BDs.

Existen diferentes formas para que los usuarios puedan acceder a la información almacenada en BDs. Las interfaces graficas son una opción cuya ventaja principal es la de no requerir un conocimiento profundo para su manejo. Desafortunadamente, la información a la que se tiene acceso es muy limitada debido a que está restringida al propósito específico para el cual fue diseñada la interfaz.

Una forma más flexible de acceder a los datos sin limitaciones es a través de consultas en SQL, sin embargo, sólo los usuarios expertos en este lenguaje pueden formular consultas eficazmente, lo cual excluye a los usuarios inexpertos.

Tomando en cuenta lo anterior, las interfaces de lenguaje natural (ILNs) son una solución viable para este problema, al brindar facilidad y flexibilidad para acceder a la información. En el mercado actual existen muy pocas ILNs para consultas a BDs, y las que existen tienen un desempeño con una amplia área de oportunidad para mejorar [Pazos, 2021].

Como se menciona en los antecedentes, el ITCM cuenta con una ILNBD para el idioma español. Su nuevo analizador sintáctico le permite tener un desempeño del 100% de consultas correctamente contestadas cuando las consultas involucran columnas booleanas, palabras del idioma y valores de difícil detección en español. Además, el analizador obtiene resultados muy aceptables en la interpretación y respuesta de consultas en dicho idioma [Verástegui, 2020].

Surgió entonces la necesidad de probar si es posible migrar las estrategias implementadas en el analizador actual **a un idioma diferente** y probar si se obtienen resultados similares. El inglés al ser uno de los idiomas más hablados[1], tanto de forma nativa como de lengua común no nativa, es la opción más adecuada, considerando que la ILNBD del ITCM no cuenta con un analizador sintáctico para este idioma.

Considerando lo anterior, este proyecto busca adaptar al idioma inglés los métodos propuestos en [Verástegui, 2020] para el analizador de español. Esto es con el fin de probar si dichos métodos pueden o no ser adaptados a otros idiomas y comprobar si es posible obtener resultados similares en el análisis de consultas en LN en inglés. Por otro lado, la ILNBD del ITCM contará con una primera versión, tanto de un analizador como de las reglas de producción para inglés, lo cual le permitirá procesar consultas en este idioma.

1.6 Beneficios

En lo que se refiere a este proyecto, los tres principales beneficios a resaltar son los siguientes:

1. Adaptar el analizador sintáctico-semántico (AS-S) propuesto en [Verástegui, 2020] para el idioma inglés.
2. La ILNBD del ITCM contará con un AS-S que responda consultas en el idioma inglés.
3. Contar con el código del algoritmo del AS-S de inglés para poder trabajar con él en trabajos futuros.

[1] https://www.thehistoryofenglish.com/history_today.html

1.7 Alcances y Limitaciones

Los alcances de este proyecto son criterios que deben cumplirse durante el desarrollo de la investigación para alcanzar los objetivos del proyecto. Los alcances se describen en los siguientes puntos:

- Las consultas a analizar deben estar escritas en lenguaje natural en inglés.
- La definición del corpus utilizado para la creación del lexicón.
- Creación de un lexicón para el idioma inglés con información gramatical para el correcto análisis de consultas en LN.
- Los analizadores léxico y sintáctico-semántico, descritos en [Verástegui, 2020], deben funcionar de forma correcta para el idioma inglés.
- La definición de las reglas de producción para inglés para la versión del AS-S para este idioma.
- Las reglas de producción podrán ser modificadas sin comprometer el funcionamiento del analizador.

Las limitaciones de este proyecto permiten delimitar los alcances de la investigación y se describen en los siguientes puntos:

- El AS-S no será multilenguaje, específicamente, el idioma a analizar es únicamente el inglés.
- La comunicación es únicamente en lenguaje escrito.
- Las consultas deben ser léxica y sintácticamente correctas.
- Las consultas deben ser oraciones del tipo interrogativo e imperativo en inglés.
- La nueva versión de la ILNBD no generará código en SQL, es decir, no contestará consultas.

CAPÍTULO 2

Marco Teórico

En este capítulo se abordarán los conceptos y métodos principales para el desarrollo del analizador de inglés. De igual forma se requiere definir algunos términos relevantes para el correcto entendimiento de los temas involucrados en este proyecto. Adicionalmente, se definen de manera breve las técnicas a implementar en este proyecto.

2.1 Procesamiento de Lenguaje Natural

A fin de profundizar en el tema del procesamiento del LN, se presentan algunas definiciones básicas importantes para facilitar la comprensión del tema:

***Lenguaje Formal*:** "Conjunto de cadenas de símbolos formados de acuerdo con algunas reglas que determinan cómo se pueden combinar los símbolos de una colección determinada" [Oxford, 2020a].

***Lenguaje Natural*:** "Un lenguaje que se ha desarrollado de forma natural por las personas que lo usan para comunicarse, en lugar de un lenguaje inventado o un código de computadora" [Oxford, 2020b].

***Oración*:** "Es la expresión lingüística de una proposición. Intervienen en ella un sujeto del cual se afirma algo y la afirmación que se hace respecto a ese sujeto" [Sapir, 1954]. "Estructura gramatical formada por la unión de un sujeto y un predicado" [RAE, 2021].

***Unidad Léxica (token)*:** Es una unidad del lenguaje aislable compuesta por letras, además del guion y del apóstrofe, la cual se encuentra situada entre dos espacios. Ésta puede tener o carecer de significado propio [Martínez de Sousa, 1995]. Una sola palabra, una parte de una palabra o una cadena de palabras que forman los elementos básicos del léxico de un idioma (vocabulario) [Lewis, 1997].

***Categoría gramatical*:** También conocida como **categoría sintáctica**, es la clase de palabras que corresponden tradicionalmente a las partes del discurso (en inglés *parts of speech*, v. gr. sustantivo, verbo, preposición, etc.) [Brinton, 2010].

***Sintaxis*:** De acuerdo con el diccionario Oxford, la sintaxis está definida lingüísticamente como "el **arreglo de palabras y frases para formar oraciones en un lenguaje**" y, a su vez, computacionalmente como "las reglas que establecen cómo se deben usar palabras y frases en un leguaje computacional" [Oxford, 2021a].

***Regla Gramatical*:** "Regla lingüística para la sintaxis de expresiones gramaticales" [Farlex, 2020].

***Información semántica*:** "Responde a las relaciones que se establecen entre significados" [Azcoaga, 1989]. La información semántica puede ser instruccional o factual, la información instruccional instruye o da pie a que algo se lleve a cabo sin tener un valor de verdad, la información factual hace alusión a hechos, por lo tanto, tiene un valor de verdad [Floridi, 2015].

***Procesamiento de Lenguaje Natural*:** "El procesamiento, o tratamiento por computadora, de lenguaje natural" [Cohen, 2004]. "Consiste en tomar una oración escrita o hablada y procesarla para obtener su significado, por lo que el sistema obtiene la secuencia de palabras (oración), y el resto del proceso para el entendimiento puede ser dividido en varias etapas: análisis léxico, análisis morfológico, análisis sintáctico, análisis semántico y análisis contextual. Cada una de estas etapas cumple una tarea necesaria para llegar a un entendimiento del LN" [Cervantes, 2005].

***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…) 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" [Saiz, 2003].

***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 rección (es decir, la manera en que ciertas palabras imponen a otras variaciones de número, género…) y las funciones que las palabras puedan cumplir en la oración" [Saiz, 2003].

***Análisis Semántico*:** La semántica es el subcampo **que estudia el significado**. Puede abordar el significado en los niveles de palabras, frases, oraciones o unidades más grandes del discurso. Una de las cuestiones cruciales que une diferentes enfoques de la semántica lingüística es la de la relación entre forma y significado [Kroeger, 2019]. "La información semántica es responsable de la correcta combinación de unidades léxicas en un discurso" [Saiz, 2003].

***Analizador Sintáctico*:** "Podemos distinguir dos clases de analizadores sintácticos: analizador superficial y analizador completo. Los componentes sintácticos aislados se identifican en el analizador superficial. No se establecen relaciones sintácticas entre ellos, por lo que el costo computacional es bajo, a costa de disminuir la profundidad del análisis en la oración. Un analizador completo, por el contrario, rechaza cualquier oración que no pueda analizar globalmente. Sin embargo, proporciona información mucho más valiosa, ya que establece vínculos funcionales entre los diferentes elementos sintácticos que constituyen la oración" [Verástegui, 2020].

2.2 Lexicón

Antes de abordar el tema, se presentan algunas definiciones básicas importantes para facilitar su comprensión:

***Base de Datos*:** "Una BD es una colección continua de información que es utilizada en sistemas y aplicaciones de una compañía o por usuarios" [Date, 2003].

***Lexicón*:** Un lexicón es una colección ordenada de lexemas (palabras) pertenecientes a un sistema lingüístico, el cual puede incluir información gramatical de las palabras que lo forman [Mellado, 2014].

***Consulta en Lenguaje Natural*:** "Una consulta expresada de forma escrita o hablada en inglés, francés o cualquier otro idioma hablado de manera normal" [PCMag, 2020]. "Una consulta en LN consta sólo de términos normales en el idioma del usuario sin ninguna sintaxis o formato especial" [Oracle, 2011].

En las ILNBDs, un lexicón constituye uno de los primeros componentes a definir, dado que su función primordial es ser la base para la interpretación de las consultas en LN [Cervantes, 2005]. La exactitud, al realizar la identificación de las categorías gramaticales de las palabras que conforman una consulta en LN y el grado de comprensión del significado de ésta, dependen de la información almacenada en la BD del lexicón.

La ILNBD en español del ITCM cuenta con un lexicón que almacena todas las palabras del idioma español (aproximadamente 1,132,218 palabras), divididas en trece categorías gramaticales: adjetivos, adjetivos plurales, adverbios, artículos, conjunciones, interjecciones, participios, preposiciones, pronombres, sustantivos, sustantivos plurales, verbos y verbos auxiliares [Aguirre, 2014].

En el Capítulo 4 de este proyecto, se aborda la arquitectura definida para un lexicón de inglés que utiliza como guía la estructura del lexicón de la interfaz en español.

2.2.1 Fuente de Lexemas para el Lexicón de Inglés

A diferencia del idioma español, el cual cuenta con una institución cultural dedicada a la regularización lingüística del lenguaje hispanohablante, la lengua inglesa no cuenta con un órgano regulador oficial del idioma como lo mencionan [Brock, 2016] y [The Boston Language Institute, 2013]. No obstante, existen diferentes autoridades altamente reconocidas que definen de forma general las estructuras del idioma inglés, por ejemplo: Oxford University y Cambridge University.

Adicionalmente a autoridades como Oxford University y Cambridge University, existen diferentes corpus predefinidos de lexemas con alto reconocimiento en la literatura. Uno de los más reconocidos y de mayor tamaño es "The Corpus of Contemporary American English (COCA)" [Davies, 2009] o "Corpus de Inglés Americano Contemporáneo". Dado que en este proyecto se trabajó con el corpus COCA, se presenta enseguida una visión general sobre este corpus.

2.2.2 Corpus COCA

COCA es un corpus de acceso gratuito de inglés americano y es el único con un gran tamaño, aproximadamente mil millones de palabras, con un equilibrio variado entre los géneros de las fuentes de extracción de sus palabras. Es un corpus que se encuentra en constante evolución, ya que se actualiza constantemente (última actualización marzo 2020) [Davies, 2008].

Fue lanzado en 2008 y pertenece a un grupo de corpus denominado "BYU Corpora", los cuales fueron creados por Mark Daves, profesor de lingüística en la Universidad de Brigham Young [Davies, 2010]. Es uno de los corpus más utilizados, debido a que se encuentra estrechamente relacionado con muchos otros corpus de inglés.

El corpus COCA recopila sus palabras de un amplio grupo de textos recientes, los cuales pertenecen a diferentes áreas y géneros. Esto le permite contar con más de mil millones de palabras extraídas de diversos textos (20 millones de palabras cada año entre 1990-2020) de ocho géneros diferentes [Davies, 2008]. La Tabla 2.1[2] muestra los diferentes géneros incluidos en el corpus.

Tabla 2.1. Descripción de los géneros del corpus COCA

Género	No. textos	No. palabras	Definición
Hablado	44,803	127,396,932	Transcripción de conversaciones no escritas de más de 150 programas de radio y televisión. Por ejemplo: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC) y Oprah.
Ficción	25,992	119,505,305	Cuentos cortos y obras de teatro de revistas literarias, revistas infantiles, revistas populares, primeros capítulos de libros de primera edición desde 1990 al presente y fan-fictions.
Revistas	86,292	127,352,030	Cuenta con casi 100 revistas diferentes con una buena combinación entre dominios específicos como noticias, salud, hogar y jardinería, mujeres, finanzas, religión, deportes, etc.
Textos Académicos[3]	26,137	120,988,361	Más de 200 revistas diferentes que cubren una gama completa de disciplinas académicas con un buen equilibrio entre educación, ciencias sociales, historia, humanidades, derecho, medicina, filosofía/religión, ciencia/tecnología y negocios.
Periódicos	90,243	122,958,016	Periódicos de todo Estados Unidos, incluidos USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Tiene una buena mezcla de diferentes secciones del periódico como locales, noticias, opinión, deportes, finanzas, etc.
Web (Genl)	88,989	129,899,427	Clasificado en los géneros web de académico, argumento, ficción, información, instrucción, legal, noticias, personal, promoción, páginas web revisadas (por Serge Sharoff). Tomado de la porción estadounidense del corpus GloWbE.
Web (blog)	98,748	125,496,216	Textos que Google clasificó como blogs. Más clasificado en los géneros de web académico, argumento, ficción, información, instrucción, legal, noticias, personal, promoción, revisión de páginas web. Tomado de la porción estadounidense del corpus GloWbE.
TV/Películas	23,975	129,293,467	Subtítulos de OpenSubtitles.org, y más tarde los corpus de TV y Películas. Los estudios han demostrado que el lenguaje de estos programas y películas es aún más coloquial y básico que los datos en los "corpus hablados" actuales.

[2] https://www.english-corpora.org/coca/help/coca2020_overview.pdf

[3] Actualizado hasta marzo de 2020 [Davies, 2008].

2.2.3 Categorías del Lexicón de Español

El lexicón de español cuenta con trece tablas, una por categoría gramatical, llamadas respectivamente según la categoría gramatical a la que hacen alusión: adjetivos, adjetivos plurales, adverbios, artículos, conjunciones, interjecciones, participios, preposiciones, pronombres, sustantivos, sustantivos plurales, verbos y verbos auxiliares [Aguirre, 2014].

Para definir las categorías gramaticales que debe contener la versión para inglés del lexicón, se requiere determinar si existen estructuras gramaticales equivalentes para inglés. En [Oxford, 2020c], se establece que existen nueve categorías gramaticales en el idioma inglés, las cuales se muestran en la Tabla 2.2.

Tabla 2.2. Categorías gramaticales de inglés

adjective – adjetivo	noun – sustantivo
adverbs – adverbio	preposition – preposición
conjunction – conjunción	pronoun – pronombre
determiner – determinante	verb – verbo
exclamation – exclamación	

2.2.4 Etiquetado Usado en el Lexicón

El lexicón almacena tanto las palabras del idioma como información gramatical de las mismas que brindan información adicional para el etiquetado de las consultas en LN durante el análisis léxico.

El corpus COCA utiliza para el etiquetado gramatical del *Part of Speech* (*PoS*) el sistema "Constituent Likelihood Automatic Word-tagging System" (CLAWS) o sistema de etiquetado automático de palabras de probabilidad constituyente en su versión "CLAWS7 Tagset" [Garside, 1987].

El sistema de etiquetado CLAWS7 aporta información sobre los accidentes gramaticales de las palabras tales como género, tipo, número, persona, etc., permitiendo agregar más columnas a las diferentes tablas del lexicón. La Tabla 2.3 muestra las etiquetas divididas por categorías gramaticales e indica en la primera columna la etiqueta seguida de una breve definición de cada una [Rayson, 1998].

Tabla 2.3. Etiquetas del lexicón de inglés

Adjectives – Adjetivos	
JJ	Se utiliza para la clase principal de adjetivos, aquéllos que se pueden usar de manera predicativa o atributiva (con o sin el mismo significado).
JJR	Se usa para adjetivos comparativos, p. ej., `whiter` - *más blanco*.
JJT	Se usa para adjetivos superlativos, p. ej., `whitest` - *el más blanco*.
JK	(catenativo). Se usa para poder, capacidad y voluntad en oraciones, p. ej., "*Will you be* `able_JK` *to manage?*", pero no cuando se usa como adjetivo general, p. ej., "*Your son is very* `able_JJ`".

Adverbs – Adverbios	
RA	Se usa después de la cabeza nominal, p. ej., `else` - *si no*, `galore` - *en abundancia*.
REX	Se usa en adverbios que introducen construcciones aposicionales p. ej., `namely` - *a saber*, `viz.` - *a saber*.
RG	Se usa para adverbios de grado, p. ej., `very` - *muy*, `so` - *muy*, `too` - *también*.
RGA	Se usa para adverbios de grado postnominal / adverbial / adjetivo, p. ej., `indeed` - *de hecho*, `enough` - *suficiente*.
RGQ	Adverbios de grado wh-, p. ej., `how` - *cómo*.
RGQV	Adverbios de grado how-, p. ej., `however` - *sin embargo*.
RGR	Adverbios de grado comparativos, p. ej., `more` - *más*, `less` - *menos*.
RGT	Adverbios de grado superlativo, p. ej., `most` - *el más*, `least` - *el menos*.
RL	Adverbios locativos, p. ej., `alongside` - *junto*, `forward` - *adelante*.
RP	Se usa para adverbios o partículas preposicionales, p. ej., `in` - *adentro*, `up` - *arriba*, `about` - *aproximadamente*.
RPK	prep. adv., catenative, p. ej., `about` in `be about to` - *a punto* en *estar a punto de*.
RR	Se usa para los adverbios generales, p. ej., `actually` - *en realidad*.
RRQ	wh- adverbio general, p. ej., `where` - *donde*, `when` - *cuando*, `why` - *por qué*, `how` - *cómo*.
RRQV	wh-ever adverbio general, p. ej., `wherever` - *donde sea*, `whenever` - *cuando sea*.
RRR	Adverbio general comparativo, p. ej., `better` - *mejor*, `longer` - *más largo*.
RRT	Adjetivo general superlativo, p. ej., `best` - *el mejor*, `longest` - *más largo*.
RT	Adverbios de tiempo, p. ej., `now` - *ahora*, `tommorrow` - *mañana*.
Articles – Artículos	
AT	Artículo, p. ej., `the` - *el*, `no` - *no*.
AT1	Artículo singular, p. ej., `a` - *un*, `an` - *un*, `every` - *cada*.
Pronouns – Pronombres	
APPGE	Pronombre posesivo, p. ej., `my` - *mi*, `your` - *tu*, `our` - *nuestro*, etc.
PNQO	p. ej., `whom` - *de quién*.
PNQS	p. ej., `who` - *quién*.
PNQV	p. ej., `whoever` - *quien sea*, `whomever` - *a quien sea*, `whomsoever` - *cualquiera que sea*, `whosoever` - *quienquiera que*.
PNX1	Pronombre indefinido reflexivo, p. ej., `oneself` - *uno mismo*.
PP	Pronombre personal posesivo nominal, p. ej., `mine` - *mío*, `yours` - *tuyo*.
PPH1	`It` - *eso*.
PPHO1	`Him` - *él*, `her` - *ella*.
PPHO2	`Them` - *ellos*.
PPHS1	`He` - *él*, `she` - *ella*.
PPHS2	`They` - *ellos*.
PPIO1	`Me` - *yo*.
PPIO2	`Us` - *nosotros*.
PPIS1	`I` - *yo*.
PPIS2	`We` - *nosotros*.
PPX1	Pronombre personal reflexivo singular, p. ej., `yourself` - *tú mismo*, `itself` - *sí mismo*.
PPX2	Pronombre personal reflexivo plural, p. ej., `yourselves` - *ustedes mismos*, `ourselves` - *nosotros mismos*.
PPY	`You` - *tú, ustedes*.
Conjunctions – Conjunciones	
BCS	Antes de una conjunción, p. ej., `in order` - *en orden*, `that` - *eso*, `even` - *incluso*, `if` - *si*, etc.
CC	Conjunciones coordinantes, p. ej., `and` - *y*, `or` - *o*.
CCB	Conjunciones coordinantes, p. ej., `but` - *pero*.
CS	Conjunción subordinada, p. ej., `if` - *si*, `because` - *porque*, `unless` - *a menos que*.

CSA	`As` - *como* como conjunción.
CSN	`Than` - *que* como conjunción.
CST	`That` - *que* como conjunción.
CSW	`Whether` - *ya sea* como conjunción.
Nouns – Sustantivos	
ND1	Sustantivo singular de dirección, p. ej., `north` - *norte*, `southeast` - *sureste*.
NN	Nombre común, neutral para números, p. ej., `sheep` - *oveja*, `cod` - *bacalao*.
NNA	Sustantivo que sigue a un título, p. ej., `M.A.`
NNB	Sustantivo anterior a un título, p. ej., `Mr, Prof`.
NN1	Sustantivo común singular, p. ej., `book` - *libro*, `girl` - *niña*.
NN2	Sustantivo común plural, p. ej., `books` - *libros*, `girls` - *niñas*.
NNL1	Sustantivo locativo singular, p. ej., `street` - *calle*, `bay` - *bahía*.
NNL2	Sustantivo locativo plural, p. ej., `islands` - *islas*, `roads` - *caminos*.
NNO	Sustantivo neutral para número, p. ej., `dozen` - *docena*, `thousand` - *mil*.
NNO2	Sustantivo neutral para número plural, p. ej., `hundreds` - *cientos*, `thousands` - *miles*.
NNT	Sustantivo temporal, neutral para número.
NNT1	Sustantivo temporal singular, p. ej., `day` - *día*, `week` - *semana*, `year` - *año*.
NNT2	Sustantivo temporal singular, p. ej., `days` - *días*, `weeks` - *semanas*, `years` - *años*.
NNU	Unidad de medida, p. ej., `in.` - *pulgada*, `cc.` - *centímetro cúbico*.
NNU1	Unidad de medida singular, p. ej., `inch` - *pulgada*, `centimetre` - *centímetro*.
NNU2	Unidad de medida singular, p. ej., `inches` - *pulgadas*, `centimetres` - *centímetros*.
NP	Nombres propios, p. ej., `Phillipines` - Filipinas, `Mercedes`.
NP1	Nombres propios singulares, p. ej., `London, Jane, Frederick`.
NP2	Nombres propios plurales, p. ej. `Browns, Reagans, Koreas`.
NPD1	Sustantivo singular días de la semana, p. ej., `Sunday` - *Domingo*.
NPD2	Sustantivo plural días de la semana, p. ej., `Sundays` - *Domingos*.
NPM1	Sustantivo singular para meses, p. ej., `October` - *Octubre*.
NPM2	Sustantivo plural para meses, p. ej., `Octobers` - *Octubres*.
Prepositions – Preposiciones	
IF	`For` - *para* como preposición.
II	Preposición.
IO	`Of` - *de* como preposición.
IW	`With` - *con* como preposición.
Verbs – Verbos	
VB0	`Be` - *ser* o *estar*.
VBDR	`Were` - *era* o *estaba*.
VBDZ	`Was` - *fue* o *estaba*.
VBG	`Being` - *siendo* o *estando*.
VBM	`Am` - *soy* o *estoy*.
VBN	`Been` - *sido* o *estado*.
VBR	`Are` - *eres* o *estás*.
VBZ	`Is` - *es* o *está*.
VD0	`Do` - *hacer*.
VDD	`Did` - *hizo*.
VDG	`Doing` - *haciendo*.
VDN	`Done` - *hecho*.
VDZ	`Does` - *hace*.
VH0	`Have` - *tener*.
VHD	`Had` - *tenía* (tiempo pasado).
VHG	`Having` - *teniendo*.
VHN	`Had` - *tenía* (pasado participio).

VHZ	`Has` - *tiene*.
VM	Verbos como auxiliares, p. ej., `can` - *poder*, `will`, `would`, etc.
VMK	Catenativo modal, p. ej., `ought` - *deber*, `used` - *soler*.
VV0	Forma básica del verbo léxico, p. ej., `give` - *dar*, `work` - *trabajar*, etc.
VVD	Forma de tiempo pasado del verbo, p. ej., `gave` - *dío*, `worked` - *trabajó*, etc.
VVG	Forma –`ing` del verbo o gerundios, p. ej., `giving` - *dando*, `working` - *trabajando*, etc.
VVN	Forma del pasado participio del verbo, p. ej., `given` - *dado*, `worked` - *trabajado*, etc.
VVZ	Forma –`s` del verbo, p. ej., `gives` - *da*, `works` - *trabaja*, etc.
VVGK	Forma –`ing` en un verbo catenativo, p. ej., `going` en `be going to` - *estar yendo a, ir yendo a*.

2.3 Analizador de Español de la ILNBD

Para la ciencia de la computación, un análisis léxico corresponde al proceso de convertir una secuencia de caracteres a una secuencia de cadenas con significado asignado (*tokens*) para su futura interpretación. En el caso de las ILNBDs, la secuencia de caracteres inicial consiste en una oración en LN que funciona como una consulta introducida por un usuario en un interfaz para ser procesada, como lo muestra la Figura 2.1 [Aguirre, 2014].

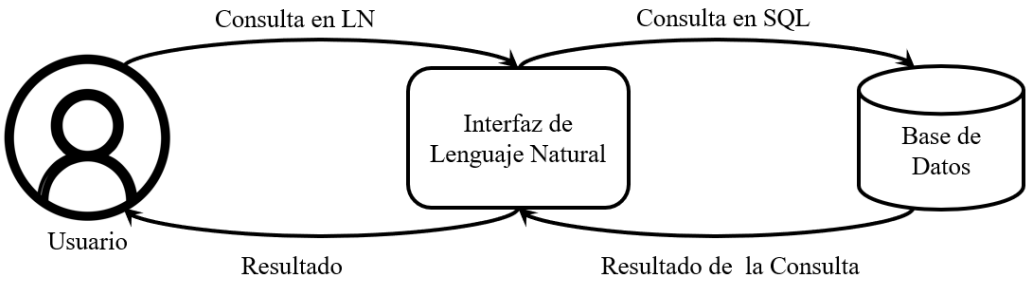


Figura 2.1. Flujo de una ILNBD

Como se mencionó en el Capítulo 1, la última versión de la ILNBD en español cuenta con dos analizadores principales para realizar el procesamiento de las consultas en LN: un analizador léxico y un analizador sintáctico-semántico [Verástegui, 2020].

2.4 Analizador Léxico

El objetivo del analizador léxico es el de etiquetar cada una de las palabras que conforman la consulta en LN con la categoría gramatical correspondiente. Éste cuenta con tres procesos principales que le permiten llevar a cabo su objetivo, como se muestra en la Figura 2.2.

Durante el desarrollo de la versión para español del analizador léxico (descrito en [Verástegui, 2020]), se concluyó que éste consumía demasiado tiempo, debido a que éste procesaba cada palabra buscándola en el lexicón, incluyendo los valores de búsqueda.

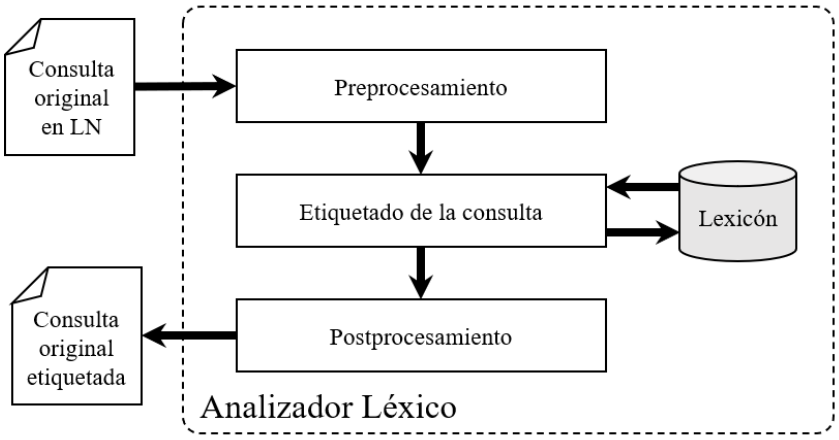


Figura 2.2. Funciones del analizador léxico

El proceso resultaba ineficiente, debido a que los valores de búsqueda generalmente no son palabras del idioma, por lo cual no se encuentran almacenadas en el lexicón. Para dar solución a este problema, en [Verástegui, 2020] se presentan tres funciones principales para el analizador léxico: preprocesamiento, etiquetado y postprocesamiento léxico.

2.4.1 Preprocesamiento Léxico

El **preprocesamiento** consiste en la aplicación de expresiones regulares definidas en [Verástegui, 2020] para la identificación de valores de búsqueda, descartándolos del proceso de búsqueda en el lexicón, ya que la mayoría de los valores de búsqueda no son palabras del idioma. Por lo tanto, no se encuentran almacenadas dentro del lexicón, reduciendo de esta forma el tiempo de procesamiento de la consulta en LN. La Tabla 2.4 muestra las expresiones regulares definidas en [Verástegui, 2020].

Tabla 2.4. Expresiones regulares para la identificación de valores de búsqueda

ID	Expresión regular	Definición
1	\b[a-z]+\b	Palabra
2	\b[A-Z][a-z]+([-]? [A-Z][a-z]+)?\b	Nombre propio
3	\b[\$]?[0-9][0-9]?[0-9]?([,][0-9][0-9][0-9])*[.][0-9]+\b	Decimal
4	\b[\$]?[0-9][0-9]?[0-9]?([,][0-9][0-9][0-9])*\b	Número entero
5	\b[0-9]?[0-9][:][0-9][0-9]\b	Hora
6	\b[0-9][0-9](([\/][0-9][0-9][\/][0-9][0-9][0-9][0-9])|([-][0-9][0-9] [-][0-9][0-9][0-9][0-9]))\b	Fecha
7	Las unidades léxicas (*tokens*) que no fueron identificadas en ninguna de las expresiones regulares anteriores.	Clave

Nota: la notación utilizada está descrita en [Cardiff School of Computer Science & Informatics, 1997].

Al utilizar expresiones regulares para la identificación de los valores de búsqueda, se reduce el tiempo de ejecución del análisis léxico, permitiendo que la búsqueda en el lexicón sea más sencilla. En consecuencia, únicamente las palabras identificadas con la etiqueta *word* son seleccionadas para buscarse en el lexicón.

Gracias a este enfoque, el analizador logra la traducción de consultas que incluyen valores de búsqueda de difícil detección con una exactitud del 100% utilizando el corpus de ATIS, el cual es utilizado como punto de referencia para la evaluación de la ILNBD [Verástegui, 2020].

2.4.2 Etiquetado Léxico

La función del **etiquetado léxico** consiste en buscar en el lexicón cada una de las unidades léxicas de la consulta en LN y que no fueron descargadas en la función anterior como valores de búsqueda. Esto es con la finalidad de definir todas las categorías gramaticales de cada palabra.

Para el proceso de etiquetado de las consultas, se requiere la implementación del Algoritmo 2.1 descrito en [Aguirre, 2014]. En el pseudocódigo del Algoritmo 2.1, *Q* identifica la consulta en LN introducida por el usuario, Q_i es una palabra (*token*) de la consulta, *n* es total de palabras en la consulta y *L* es la lista para almacenar las categorías gramaticales de la unidad léxica Q_i. En la línea 4, para cada unidad léxica de la consulta *Q*, se almacena en *L* una lista de las categorías gramaticales de Q_i; en las líneas 6 a 9, si la lista L no está vacía, todas las categorías gramaticales que se encuentran para la unidad léxica se almacenan en Q_i.

Algoritmo 2.1 Pseudocódigo del etiquetado léxico

```
1:  Q // Entrada de la consulta en LN
2:  n // Número total de unidades léxicas
3:  L // Lista de categorías gramaticales
4:  for i = 0 to n-1 do    // Para cada unidad léxica de la consulta Q
5:     L ← getGramaticalTag(Qi)    // Obtiene lista de categorías gramaticales de la unidad léxica
6:     if isNotEmpty(L)
7:        asignaEtiquetaGram(L, Qi)    // Asigna etiqueta a la unidad léxica Qi
8:     end if
9:  end for
```

Una vez concluido el etiquetado léxico, se obtiene como resultado la consulta original dividida en unidades léxicas que representan a cada una de las palabras que la conforman. Cada unidad léxica se encuentra etiquetada, sin embargo, ésta puede tener una o varias categorías gramaticales asignadas.

2.4.3 Postprocesamiento

El AS-S fue diseñado en base a reglas de producción, por lo cual requiere que cada una de las palabras que forman la consulta tengan asignada una sola categoría gramatical. Sin embargo, existen palabras que cuentan con dos o más categorías gramaticales (ambigüedad léxica).

En el **postprocesamiento**, se realiza una depuración de las categorías gramaticales por medio de una heurística propuesta en [Verástegui, 2020], la cual se encarga de definir qué categoría gramatical tiene la mayor probabilidad de ser la correcta con base en las palabras que tiene cerca. El postprocesamiento se aplica únicamente a aquellas palabras que tengan asignadas más de una categoría gramatical.

Para la aplicación del postprocesamiento, se requiere la implementación del Algoritmo 2.2 descrito en [Verástegui, 2020]. El Algoritmo 2.2 examina cada unidad léxica Q_i de la consulta, si se detecta que Q_i tiene más de una categoría gramatical, se extrae de un arreglo bidimensional la categoría gramatical con la mayor probabilidad de ser la correcta y se asigna a la unidad léxica Q_i.

Algoritmo 2.2 Pseudocódigo del postprocesamiento léxico

```
1:  procedimiento: CargaConsulta()
2:  for i = 0 to n-1 do    // Para cada unidad léxica de la consulta Q
3:     if TieneMultiplesCategorias(Qi) // Unidad léxica con múltiples categorías
4:        EstableceClase (Qi, BuscarCalificacion(i)  // Obtiene la categoría con el
          puntaje más alto, y se establece como la categoría correcta
5:     end if
6:  end for
```

Para elegir la categoría gramatical correcta en el algoritmo propuesto en [Verástegui, 2020], se requiere que tanto las palabras anteriores como posteriores sean examinadas para poder establecer una puntuación a cada una de las categorías gramaticales alternativas. Existen cuatro posibles puntuaciones: 0, 1, 2, 3.

Cada categoría alternativa inicia con una puntuación de 1, y dependiendo de un análisis sintáctico local de las palabras más cercanas, y con la aplicación de reglas gramaticales del idioma con el que se esté trabajando, la puntuación de la categoría gramatical que está siendo evaluada puede incrementarse a 2 o 3, o puede reducirse a 0.

Una vez finalizada la aplicación de las reglas del idioma con el que se esté trabajando, con todas las categorías alternativas evaluadas y con una puntuación asignada, se selecciona aquélla que tenga la puntuación más alta. Si existieran empates entre dos o más categorías, la palabra que se está evaluando se busca en el DIS, si la palabra se encuentra, se asigna la categoría gramatical que esté almacenada en el diccionario. En caso de que la palabra no se encuentre almacenada en el DIS, ésta se marca como palabra inútil.

Cuando se concluye el proceso del analizador léxico, se envía la consulta original etiquetada con una sola categoría gramatical al AS-S para que continúe el proceso de análisis de la consulta.

2.5 Analizador Sintáctico-Semántico

El analizador sintáctico es el algoritmo más importante para la correcta interpretación de las consultas en LN, ya que las reducciones gramaticales facilitan la comprensión del significado de

la consulta por medio de la agrupación de palabras que deben ir juntas, por ejemplo: nombres de bases de datos, valores de búsqueda, fechas y nombres [Verástegui, 2020].

Existen dos tipos de analizadores: los analizadores superficiales y los analizadores completos (profundos). Un analizador superficial utiliza componentes sintácticos mínimos de la oración que pueden ser analizados y que tienen algún significado, limitando la información sintáctica obtenida [Hardeniya, 2015]. No se toman en cuenta las relaciones sintácticas entre el conjunto de palabras de la oración, por lo tanto, su costo computacional es muy bajo, a expensas de reducir la profundidad el análisis de la oración.

En contraste, un analizador completo rechaza cualquier oración que no logre analizar de forma global. Adicionalmente, proporciona información más valiosa, ya que establece relaciones funcionales entre los diferentes componentes sintácticos que forman la oración [Faili, 2009].

Las oraciones en LN presentan una gran dificultad al momento de ser analizadas por los programas computacionales, ya que cuentan con una ambigüedad substancial en su estructura general y que su principal uso es el de transmitir significado (semántica) en un rango potencialmente ilimitado de posibilidades [Jurafsky, 1996]. Esto dificulta el diseño de las reglas para el análisis de oraciones en LN.

Otra forma de realizar el análisis sintáctico es por medio de la agrupación de palabras mediante la aplicación de reglas de producción, las cuales están basadas en las reglas gramaticales del LN. Éste es el caso de los analizadores descritos en [Verástegui, 2020].

Se han desarrollado tres versiones del analizador sintáctico para la ILNBD del ITCM. La primera versión era un analizador superficial de las consultas por medio de una heurística, la cual permite obtener una sola categoría gramatical para palabras con ambigüedad léxica. Adicionalmente ignora las palabras irrelevantes (algunos verbos, artículos y preposiciones), a menos que una definición en el DIS indique que son útiles para la identificación de una columna o tabla de la BD [Aguirre, 2014]. La Figura 2.3 muestra un ejemplo del funcionamiento de este analizador.

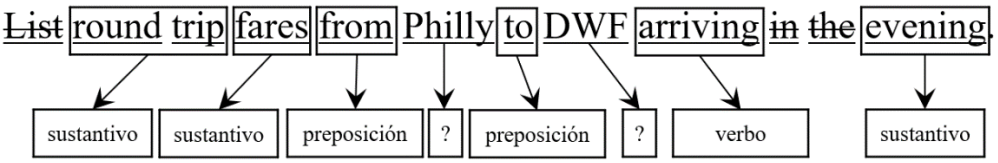


Figura 2.3. Ejemplo del analizador superficial

La segunda versión del analizador, desarrollada en el proyecto de tesis doctoral **Tratamiento de los Problemas de Valores de Búsqueda de Difícil Detección en la Traducción de Consultas de Lenguaje Natural a SQL**, realiza un análisis exhaustivo que genera todos los posibles árboles de análisis [Verástegui, 2020]. Sin embargo, este analizador dejó claro que el análisis sintáctico no puede ser efectuado sin el uso de información semántica.

Debido a lo anterior, se desarrolló la tercera y última versión de este analizador, el cual está basado en reglas de producción sintáctico-semánticas, las cuales integran información semántica en su diseño para la ejecución del análisis de consultas en LN [Verástegui, 2020].

2.5.1 Gramática del AS-S

***Gramática*:** Describe un lenguaje en términos de un conjunto de reglas para reconocer cadenas de lenguaje [Ford, 2004].

En [Verástegui, 2020] se definió una gramática (etiquetas gramaticales) para el análisis sintáctico, la cual cuenta con símbolos terminales genéricos para las diferentes categorías gramaticales generales, como se muestra en la Tabla 2.5.

Tabla 2.5. Símbolos terminales genéricos

Categoría gramatical	Símbolos genéricos generales
adjective	adj
adverb	adv
article	art
conjunction	con
verb	ver
preposition	prep
pronoun	pro
noun	nou

De igual forma se incorporan símbolos terminales categorizados para identificar algunas subcategorías gramaticales de adjetivos, pronombres y conjunciones, como se muestra en la Tabla 2.6.

Tabla 2.6. Símbolos terminales categorizados

Categoría gramatical	Símbolos categorizados	Categoría gramatical	Símbolos categorizados
positive adjectives	adj_cal	possessive adjective	adj_pos
comparative adjective	adj_comp	superlative adjective	adj_sup
demonstrative adjective	adj_dem	interrogative pronoun	pro_int
indefinite adjective	adj_ind	copulative conjunction[1]	con_cop
interrogative adjective	adj_int	disjuntive conjunction[2]	con_dis
quantitative adjective	adj_num	[1] identifica la conjunción *or* (o) [2] identifica la conjunción *and* (y)	

A su vez se definieron símbolos no terminales genéricos para las frases, como muestra la Tabla 2.7. Por último, se integran tres símbolos no terminales para sujeto (*Sbj*), complemento (*Com*) y oración (*Sentence*).

Tabla 2.7. Símbolos no terminales genéricos

Tipo de oración	Símbolos genéricos no terminales
Noun phrase	NouP
Prepositional phrase	PreP
Adjectival phrase	AdjP
Verb phrase	VerP

2.5.2 Reglas de Producción

El AS-S está basado en reglas de producción, basadas en las estructuras de la gramática del idioma español, las cuales combinan información sintáctica e información semántica en ellas. El AS-S evita la sobreproducción de árboles de análisis al asignar prioridades de aplicación a sus reglas de producción. Para tal efecto, el AS-S utiliza una técnica multipasada, es decir, una pasada por cada regla siguiendo la prioridad establecida para cada una.

La ejecución de las reglas de producción del AS-S se realiza por medio del Algoritmo 2.3, definido en [Verástegui, 2020], del cual se muestra el pseudocódigo a continuación.

Algoritmo 2.3 Pseudocódigo para la aplicación de las reglas de producción para español

```
1:  procedimiento: analysis (inputExpression)
    //todas las reglas están implementadas como métodos.
2:  inputExpression ← SN01_V (inputExpression) // inputExpression se actualiza dentro del método
3:  inputExpression ← SN02_V (inputExpression)
4:  inputExpression ← SN03_V (inputExpression)
5:  inputExpression ← SN01 (inputExpression)
6:  inputExpression ← SN02 (inputExpression)
7:  inputExpression ← SN03 (inputExpression)
8:  inputExpression ← SN06 (inputExpression)
9:  inputExpression ← SA01 (inputExpression)
10: inputExpression ← SA02 (inputExpression)
11: inputExpression ← SA03 (inputExpression)
12: inputExpression ← SA04 (inputExpression)
13: inputExpression ← SA07 (inputExpression)
14: inputExpression ← SN09 (inputExpression)
15: inputExpression ← SV01 (inputExpression)
16: inputExpression ← SN10_V (inputExpression)
17: inputExpression ← SP01 (inputExpression)
18: inputExpression ← SP02 (inputExpression)
19: inputExpression ← SP05 (inputExpression)
20: inputExpression ← SP01_V (inputExpression)
21: inputExpression ← SV03 (inputExpression)
22: inputExpression ← SV11 (inputExpression)
23: inputExpression ← SN10 (inputExpression)
24: inputExpression ← SN11 (inputExpression)
25: inputExpression ← SN12 (inputExpression)
```

La estructura utilizada para el AS-S es un arreglo bidimensional de tamaño $m \times n$, donde n es el número total de palabras de la consulta, y m es el doble de las reducciones generadas, considerando que cada reducción tiene el identificador en la parte superior (ver Tabla 2.8).

Las reglas de producción fueron implementadas como métodos, ya que de esta forma se puede cambiar el orden de las reglas sin modificar el código de implementación de éstas.

El Algoritmo 2.3 recibe como entrada una secuencia de símbolos que representan la categoría gramatical de cada palabra (*inputExpression*). Cada línea del algoritmo llama al método de la regla específica que se pasa como parámetro. El método recorre la secuencia de símbolos (terminales y no terminales) e intenta encontrar una subsecuencia de símbolos que pueda ser reducida.

Si existe una subsecuencia de símbolos qué reducir, ésta es reducida, de lo contrario, continúa recorrido la secuencia de símbolos.

Finalmente, el método almacena cada reducción en una matriz. En la Tabla 2.8 se observa que cada celda almacena un identificador de la regla aplicada y los símbolos terminales y no terminales de las reducciones generadas. Los símbolos de cada reducción se almacenan en una fila independiente.

Por ejemplo, la Tabla 2.8 muestra que en la primera reducción el sexto símbolo terminal (*name*) fue reducido a NouP0V por el método de la regla SN01_V, mientras que, en la segunda reducción, el cuarto símbolo terminal (*nou*) fue reducido a NouP1 por el método de la regla SN01. Una vez que el método termina, el algoritmo continúa con el método de la siguiente regla, y así sucesivamente hasta concluir las reducciones al llegar al símbolo terminal *Sentence*.

Tabla 2.8. Secuencia de reducciones sintácticas

Paso:	**Símbolos (terminales y no terminales)**
0:	[pro_int, ver, art, nou, pre, name]
1:	SN01-V pro_int, ver, art, nou, pre, **NouP0V**
2:	SN02 pro_int, ver, art, **NouP0**, pre, **NouP0V**
3:	SN03 pro_int, ver, **NouP1**, pre, **NouP0V**
⋮	⋮
n:	Sentence

Al finalizar, la matriz contiene todas las reducciones que se puedan realizar a la expresión de entrada. La versión para español cuenta con las reglas terminales desactivadas, ya que no son de utilidad para los objetivos de ese proyecto.

El Algoritmo 2.4 muestra el pseudocódigo del método de la regla de producción SP01_V <PrePV> ::= <pre> <NouP1V> definida para el AS-S de español [Verástegui, 2020]. En esta regla

se indica que, en caso de la detección de una preposición (pre) y una frase nominal que contenga un valor de búsqueda (NouP1V), estos dos símbolos se reducen a PrePV (frase preposicional con valor de búsqueda). Una vez detectada la reducción (línea 3), *inputExpression* se actualiza (línea 4) y los símbolos reducidos son eliminados de *inputExpression* (línea 5). Para finalizar, el identificador de la reducción e *inputExpression* actualizado son almacenados en la matriz (líneas 6 y 7).

Algoritmo 2.4 Pseudocódigo de la regla SP01_V

```
1:  Procedimiento SP01_V (inputExpression) // Regla SPO1_V <PrePV> ::= <pre> <NouP1V>
2:  for i = 0 to n-1 do    // Para cada unidad léxica de la consulta Q
3:     if inputExpression [i] = ("pre") and inputExpression [i+1] = ("NouP1V") then
4:        updateInputExpression (i, "PrePV")
5:        reduce (inputExpression, i) // Elimina las unidades léxicas que fueron reducidas
6:        addMatrix ("SP01_V") // Inserta el identificador de la regla a la matriz
7:        addMatrix ("inputExpression") // Inserta la reducción a la matriz
8:     end if
9:  end for
10: return inputExpression
```

Antes de iniciar el proceso de reducción, se efectúa una búsqueda de etiquetas gramaticales iguales consecutivas. Cuando se detectan dos o más etiquetas iguales consecutivas, se realiza la consolidación de éstas, simplificando el análisis, ya que se reduce el número de elementos a analizar. De esta forma, el AS-S de español reduce el tiempo de procesamiento y permite obtener una alta exactitud en el análisis de las consultas en LN [Verástegui, 2020].

El AS-S de español reduce el tiempo de procesamiento (22 milisegundos) con respecto a otros analizadores (Freeling, 12.2 segundos) [Verástegui, 2020]. Además, en dicho documento se resolvieron problemas que usualmente se encuentran en el análisis de oraciones en LN tales como los siguientes: sobreproducción de árboles sintácticos, incoherencia semántica, desbordamiento de datos y tiempos de ejecución extensos [Verástegui, 2020].

Finalmente, es oportuno mencionar que las reglas de producción para inglés se muestran en la Tabla 4.3.

2.6 Estructuras Gramaticales de la Lengua Inglesa

Para poder realizar el análisis de una oración en inglés, es necesario tener un entendimiento profundo de las unidades estructurales básicas de esta lengua. Con ello se podrá determinar qué componentes de la oración en inglés son obligatorios, y qué constituyentes obligatorios u opcionales pueden estar en ella. En [Celce-Murcia, 1999a] se definen 16 reglas estructurales, las cuales pueden ser expandidas para especificar los elementos que forman la regla general.

La **primera regla** indica que una oración (S) puede ser expandida o reescrita de dos formas. La primera forma $S \rightarrow (\mathrm{sm})^n\, S'$ indica que la oración puede incluir opcionalmente uno o más

modificadores de oración (sm), por ejemplo: *perhaps*, *maybe*, *yes*, *no*, etc., donde *n* representa cualquier cantidad de modificadores. Además, debe incluir una oración núcleo obligatoria (S′).

La segunda forma S → SUBJ PRED indica la forma tradicional de representar una oración, la cual debe tener un sujeto (SUBJ) y un predicado (PRED). Esta regla se puede expresar de la siguiente forma:

Regla 1. $$S \rightarrow \begin{Bmatrix} (\mathrm{sm})^n & \mathrm{S}' \\ \mathrm{SUBJ} & \mathrm{PRED} \end{Bmatrix}$$

La **segunda regla** muestra de forma explícita que S′ puede ser expandida como una oración que cuenta con un sujeto y un predicado. En los casos en que la oración cuente con modificadores, ésta se expresa de la siguiente forma:

Regla 2. $$\mathrm{S}' \rightarrow \mathrm{SUBJ} \quad \mathrm{PRED}$$

La **tercera regla** muestra la representación del sujeto como una *noun phrase* (NP) o frase nominal, la cual se expresa de la siguiente forma:

Regla 3. $$\mathrm{SUBJ} \rightarrow \mathrm{NP}$$

La **cuarta regla** indica que las frases nominales pueden expandirse de formas muy complejas tales como NP → pro, la cual indica que una frase nominal puede ser un pronombre. Además, **NP** → $\boldsymbol{det^3}$ **N** indica que la frase nominal puede estar formada por un máximo de tres determinantes y un sustantivo. Esto implica que **NP** → **N**, la cual indica que la frase nominal puede consistir únicamente en un sustantivo (N).

De igual forma NP → N PP indica que una frase nominal puede estar formada por un sustantivo y una frase preposicional que no está funcionando como predicado de la oración, por ejemplo, *people on the street* (gente en la calle), es decir, que no esté un verbo *to be* entre N y PP *people **are** on the street* (gente está en la calle) [Finegan, 2008].

Así mismo una frase nominal puede ser aquélla formada por un objeto directo (NP′) y su predicado. En los siguientes ejemplos se muestra en negrita el objeto directo y subrayado el predicado, por ejemplo: *We elected **Rei** <u>treasurer</u>* (Elegimos a Rei tesorero), *Mari considers **Shinji** <u>pretty</u>* (Mari considera que Shinji es bonito), y *Asuka placed **the book** <u>on the table</u>* (Asuka puso el libro sobre la mesa).

Regla 4. $$\mathrm{NP} \rightarrow \left\{ \begin{array}{llll} (\mathrm{det})^3 & (\mathrm{AP}) & \mathrm{N} & (\mathrm{PP}) \\ \mathrm{pro} & & & \\ \mathrm{NP}' & \begin{Bmatrix} \mathrm{NP} \\ \mathrm{AP} \\ \mathrm{PP} \end{Bmatrix} & & \end{array} \right\}$$

Notas $(\mathrm{det})^3$ indica que el determinante puede aparecer 0, 1, 2 o 3 veces. La frase adjetival (AP) y la frase preposicional (PP) son opcionales.

La **quinta regla** muestra la expansión del símbolo AP o frase adjetival (AP), la cual puede estar formada opcionalmente por *n* cantidad de intensificadores (intens) que preceden al adjetivo, los cuales indican el grado o extensión en el que el adjetivo se aplica (***really very nice*** *clothes*, "ropa realmente muy bonita"). Además, indica que puede haber *n* cantidad de adjetivos (Adj) calificativos antes de un sustantivo principal (*the* ***big old yellow*** *bus*). También indica que una frase adjetival puede estar en la posición del predicado seguida de una frase preposicional, por ejemplo, *My* ***good-for-noting*** *cousing* (Mi primo bueno para nada), donde *good* es el adjetivo y *for nothig* es una frase preposicional (PP). Esta regla se expresa de la siguiente forma:

Regla 5. $\text{AP} \rightarrow (\text{intens})^n \quad \text{Adj}^n \quad \text{PP}$

La **sexta regla** expande las frases preposicionales como una preposición (Prep) seguida de una frase nominal.

Regla 6. $\text{PP} \rightarrow \text{Prep} \quad \text{NP}$

La **séptima regla** se encarga de extender el predicado de las oraciones. Esta regla expresa que obligatoriamente el predicado de cada oración en inglés está formado por un constituyente auxiliar (AUX) seguido de una frase verbal (VP). Adicionalmente, puede incluir un número *n* de adverbiales (Advl) opcionales en la posición final del predicado.

Regla 7. $\text{PRED} \rightarrow \text{AUX} \quad \text{VP} \quad (\text{Advl})^n$

La **octava regla** provee las tres posibles expresiones sintácticas de los adverbios al final del predicado, las cuales son las siguientes: cláusula adverbial, frase adverbial y frase preposicional.

Regla 8. $\text{Advl} \rightarrow \begin{Bmatrix} \text{Advl CL} \\ \text{Advl P} \\ \text{PP} \end{Bmatrix}$

La **novena regla** indica que una cláusula adverbial (Advl CL) incluye un subordinador adverbial (adv sub) seguido por una oración nueva, por ejemplo, ***before*** *their father could find them* (antes de que su padre pudiera encontrarlos), donde el subordinador adverbial está en negrita.

Regla 9. $\text{Advl CL} \rightarrow \textbf{adv sub} \quad \text{S}$

La **décima regla** indica que una frase adverbial (Advl P) obligatoriamente contiene un adverbio (Adv) y opcionalmente puede estar precedido por *n* cantidad de modificadores, por ejemplo, *very quickly* (muy rápido).

Regla 10. $\text{Advl P} \rightarrow (\text{intens})^n \quad \text{Adv}$

La **undécima regla** expresa de qué se componen los auxiliares (AUX). Éstos están conformados por el verbo imperativo (-imper), un verbo conjugado en un tiempo (T) o un modal (M). Si se tiene un verbo conjugado en un tiempo o modal, los auxiliares pueden estar seguidos de

otros aspectos auxiliares, por ejemplo: un frasal modal (pm), un perfecto (perf) y un progresivo (prog).

$$\text{Regla 11.} \qquad \text{AUX} \rightarrow \begin{Bmatrix} \begin{Bmatrix} \text{T} \\ \text{M} \end{Bmatrix} & (\text{pm}) & (\text{perf}) & (\text{prog}) \\ -\text{imper} & & & \end{Bmatrix}$$

La **duodécima regla** muestra los tiempos morfológicos del inglés. El tiempo pasado (-past) y el tiempo presente (-pres). En [Herring, 2016] se define que en el idioma inglés no existe un tiempo futuro como tal, ya que los verbos no se conjugan de cierta manera para reflejar este tiempo.

$$\text{Regla 12.} \qquad \text{T} \rightarrow \begin{Bmatrix} -\text{past} \\ -\text{pres} \end{Bmatrix}$$

La **decimotercera y decimocuarta reglas** expanden el perfecto y progresivo en sus verbos auxiliares y las inflexiones gramaticales que los acompañan.

$$\text{Regla 13.} \qquad \text{perf} \rightarrow \text{have} \ldots \text{pasado participio}$$
$$\text{Regla 14.} \qquad \text{prog} \rightarrow \text{be} \ldots -\text{ing}$$

La **decimoquinta regla** indica que una frase verbal (VP) se puede construir de cinco formas: (a) $\text{VP} \rightarrow \text{cop NP}$, es decir, por un verbo copulativo (cop) y una frase nominal, (b) $\text{VP} \rightarrow \text{cop AP}$, es decir, por un verbo copulativo y una frase adjetival, (c) $\text{VP} \rightarrow \text{cop PP}$, es decir, por un verbo copulativo y una frase preposicional, (d) $\text{VP} \rightarrow \text{V NP PP}$, es decir, por un verbo, una frase nominal y una frase preposicional, y (e) $\text{VP} \rightarrow \text{V}(\text{NP})^2$, la cual indica que la frase verbal está formada por un verbo y dos frases nominales, donde la primera funciona como objeto indirecto y la segunda como objeto directo. Estas formas se pueden expresar de la siguiente forma:

$$\text{Regla 15.} \qquad \text{VP} \rightarrow \begin{Bmatrix} \text{cop} & \begin{Bmatrix} \text{NP} \\ \text{AP} \\ \text{PP} \end{Bmatrix} & \\ \text{V} & (\text{NP})^2 & (\text{PP}) \end{Bmatrix}$$

Nota: la frase preposicional (PP) es opcional.

La **decimosexta regla** expresa cómo puede estar conformada la frase nominal del objeto directo (NP′), donde no existe un verbo explícito, no obstante, está presente una relación predictiva.

$$\text{Regla 16.} \qquad \text{NP}' \rightarrow \begin{Bmatrix} (\text{det})^3 & (\text{AP}) & \text{N} & (\text{PP}) \\ \text{pro} & & & \end{Bmatrix}$$

2.7 Etiquetador UCREL CLAWS

El sistema “Constituent Likelihood Automatic Word-tagging System” (CLAWS), o sistema de etiquetado automático de palabras de probabilidad constituyente, fue desarrollado por UCREL en

Lancaster University. CLAWS se ha ido desarrollando continuamente desde principios de 1980 [Garside, 1987].

Es uno de los sistemas de etiquetado de *Part of Speech* (POS) más utilizados por los corpus del idioma inglés, principalmente las versiones *C*5 y *C*7 *Tagset* [Newman, 2020]. Su conjunto de etiquetas POS corresponden a las partes tradicionales para el análisis del inglés (*article*, *infinitive*, *singular common noun*, etc.) y otras no tan comunes (v. gr., *after-determiner*).

UCREL cuenta con un etiquetador POS en web de acceso gratuito[4], el cual utiliza el sistema de etiquetas CLAWS descrito en la Subsección 2.2.4, en la Tabla 2.3. Éste permite utilizar dos versiones del sistema CLAWS: versiones 5 y 7. Permite introducir textos de hasta 100,000 palabras en inglés para ser etiquetadas como se muestra en la Figura 2.4.

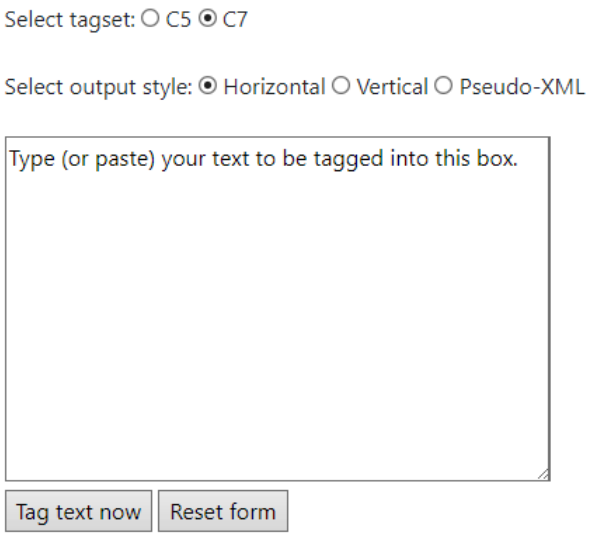


Figura 2.4. Etiquetador UCREL CLAWS

La Figura 2.5 muestra un ejemplo del proceso de etiquetado para la consulta *Which airlines have economy flights with discount from San Francisco to Dallas?* del corpus de pruebas. Como se puede observar, se obtiene el número de palabras etiquetadas y la versión del conjunto de etiquetas utilizado (versión C7).

Las etiquetas (descritas en la Tabla 2.3) aparecen en un formato predecible junto con las palabras a las que están asociadas, ambas conectadas por un guion bajo (_). Adicionalmente, se puede observar que CLAWS etiqueta los signos de puntuación trivialmente como el signo de puntuación que se esté utilizando, en el caso de la consulta, el de interrogación (?).

[4] Free CLAWS web tagger, http://ucrel-api.lancaster.ac.uk/claws/free.html

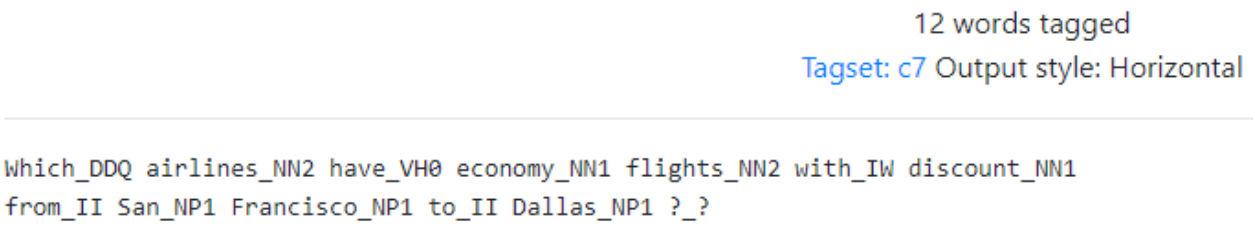


Figura 2.5. Etiquetado de consulta con CLAWS

2.8 Wolfram

Wolfram Alpha fue desarrollado por Wolfram Research, el cual es un motor de cálculo de respuestas por medio de inferencias a partir de un conjunto de información básica. Cuenta con un conjunto de herramientas muy variadas que abarca disciplinas desde matemáticas, finanzas, salud y medicina, nutrición, ingeniería, física, música, etc. [Wolfram, 2021]. Fue desarrollado por Stephen Wolfram y lanzado en marzo de 2009 [Cellan-Jones, 2009].

Wolfram cuenta con una función de análisis de oraciones llamada *TextStructure* [Wolfram, 2019]. Ésta genera el etiquetado de cada una de las palabras de la oración, además genera agrupaciones de elementos léxicos, generando reducciones que muestran la estructura gramatical de la oración en LN [Wolfram, 2021]. La Figura 2.6 muestra el análisis de la consulta *Which airlines have economy flights with discount from San Francisco to Dallas?* utilizando la función *TextStructure*.

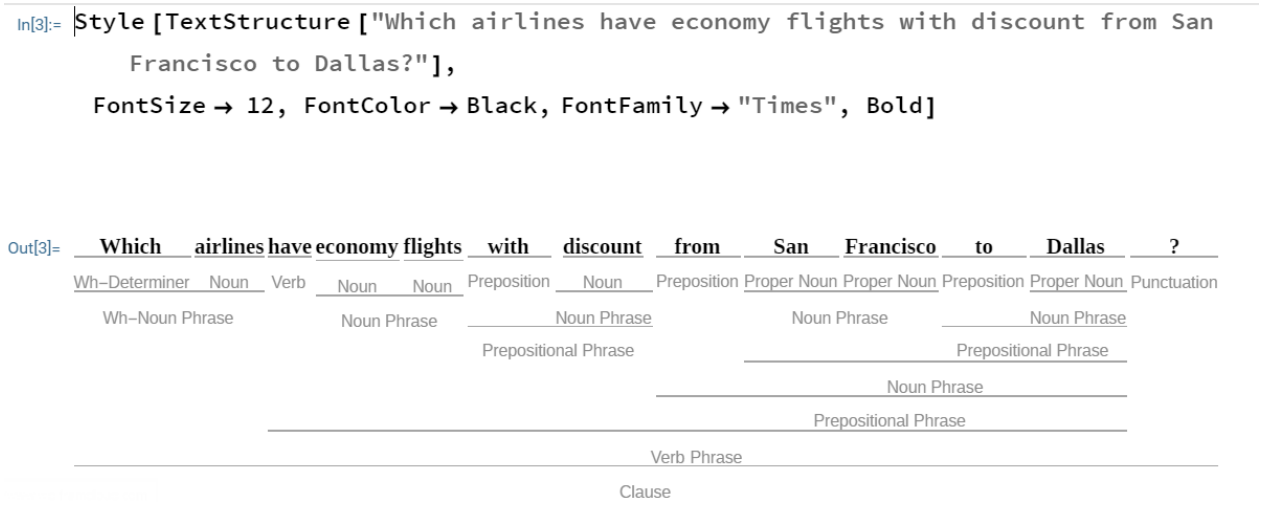


Figura 2.6. Ejemplo de función *TextStructure* de Wolfram

CAPÍTULO 3

Estado del Arte

En este capítulo se efectuará una breve presentación de los diferentes sistemas más populares de las interfaces de LN a base de datos que se han desarrollado a lo largo de los últimos 50 años. Además, se incluye la recopilación del estado del arte, incluyendo las interfaces más relevantes del estado del arte que están basadas en sintaxis.

3.1 Historia

A lo largo de los últimos 50 años, se han realizado muchos intentos de desarrollar ILNBDs inteligentes. Los primeros prototipos de ILNBDs aparecieron durante los finales de los sesentas y principios de los setentas.

Dos de las más más conocidas de ese periodo fueron BASEBALL, la cual fue diseñada para contestar preguntas de los juegos de baseball de la American League de ese periodo [Green, 1961] y LUNAR. Ésta fue diseñada para responder preguntas sobre muestras de rocas lunares y tenía una exactitud de respuesta del 80% [Woods, 1972]. Ambas interfaces estaban diseñadas para el dominio específico de la base de datos que utilizaban, por lo que era muy complicado adaptarlas para ser utilizadas con otros dominios.

Otras interfaces se desarrollaron al final de los setentas, por ejemplo: PLANES (un sistema para responder preguntas incoherentes y vagas) [Waltz, 1978], LIFER/LADDER (utilizaba sistemas semánticos gramaticales para el análisis) [Hendrix, 1977] y RENDEZVOUS (desarrollada en el laboratorio de IBM, ayudaba a los usuarios a responder y formular sus consultas en caso de encontrar ambigüedad en el análisis) [Date, 1974].

Un auge de desarrollo de ILNBDs se dio en los años 80s cuando numerosos prototipos fueron implementados como CHAT-80 [Warren, 1982], el cual fue un sistema eficiente más conocido de este periodo. Utilizaba técnicas de gramática semántica para procesar las consultas en LN de los usuarios, sin embargo, su mayor problema fue que sólo podía ser utilizado para un solo dominio. Este sistema creó las bases de varios sistemas, enfocados más que nada a usarse para varios dominios, por ejemplo: TELI [Ballard, 1982], DIALOGIC [Moore, 1981], EUFID [Templeton, 1983], DATALOG [Hafner, 1984] y MASQUE [Auxerre, 1986].

La siguiente generación de ILNBDs usaban un lenguaje de representación intermedio, el cual expresaba el significado de la consulta en términos de conceptos de alto nivel. En [Androutsopoulos, 1995] se propuso un sistema llamado MASQUE/PRO que contestaba cualquier consulta escrita en inglés como LN, el cual podía usarse con diferentes manejadores de bases de datos comerciales.

Debido a que la mayoría de las BDs existentes en ese tiempo eran relacionales, los autores decidieron modificar este sistema, naciendo de esta forma MASQUE/SQL, la cual era una interfaz portable que funcionaba con bases de datos para SQL. Otro sistema portable fue TEAM [Grosz, 1983], una interfaz experimental, la cual consistía en dos componentes principales: un mapeo de expresión de LN a una representación formal y la transformación de las consultas en esta representación a lenguaje de consulta de BDs. Esto le permitía la separación del proceso lingüístico del proceso de mapeo.

Algunas de las ILNBDs más relevantes de los últimos 20 años son PRECISE, la cual es un sistema desarrollado en la universidad de Washington [Popescu, 2004]. Sus objetivos son BDs relacionales, y el lenguaje utilizado para consultar las BDs es SQL. Combina enfoques lingüísticos y matemáticos para lograr una completa independencia de información sin necesidad de soporte ni configuración. Además, es una de las primeras interfaces que utiliza un analizador intercambiable (*plug-in*), por lo que puede ser fácilmente modificada para emplear las últimas ventajas desarrolladas en el campo de los analizadores [Majhadi, 2021].

NALIX (*an interactive natural language interface for querying XML*) es una interfaz genérica, desarrollada en la universidad de Michigan [Li, 2005]. La BD para este sistema trabaja con un lenguaje de consulta '*Schema-Free-Xquery*'. Este lenguaje está diseñado principalmente para recuperar información de BDs en XML y realizar búsquedas de palabras clave. Su principal ventaja es que encuentra automáticamente todas las relaciones dadas con muchas palabras clave sin mapear una consulta en el esquema exacto de la base de datos. NaLIR es otro sistema genérico interactivo para consultar BDs relacionales, el cual acepta oraciones complejas en inglés [Li, 2014].

También se han desarrollado muchos sistemas de consulta a BDs que utilizan una interfaz de LN universal basados en el enfoque de aprendizaje automático [Bais, 2016]. Además, una ILNBD árabe llamada Seq2SQL, la cual está basada en el procesamiento del LN y los métodos de la teoría de grafos [Zhong, 2017].

Varias ILNBDs han sido desarrolladas con fines comerciales. Algunas de las más exitosas fueron INTELLECT, Q&A, English Wizard y English Query. Sin embargo, cuando las interfaces graficas fueron introducidas, estas ILNBDs fueron gradualmente descontinuadas. Su desempeño no fue muy satisfactorio debido a la existencia de diversos problemas.

3.2 Trabajos Recientes

Durante la recopilación del estado del arte, se descubrió que existen muy pocos analizadores que en sus reglas de producción incluyan información sintáctica y semántica. Por lo tanto, se realizó

una investigación de algunas de las interfaces más relevantes del estado del arte que están basadas en sintaxis.

Los sistemas basados en sintaxis utilizan un tipo de gramática que consiste en símbolos y reglas que son aplicadas a una expresión que se encuentra en LN, permitiendo analizar la estructura de las oraciones y de esta forma generar agrupaciones de palabras relacionadas sintácticamente.

FreeLing es una biblioteca de código abierto multilenguaje, personalizable y extensible para desarrolladores que les permite realizar procesamiento de LN [Atserias, 2006]. Utiliza un analizador de dependencias para relacionar unidades léxicas con modificadores de la oración [Padró, 2011]. Cuenta con diversas clases de procesamiento, realiza diversas tareas como la identificación del idioma. La versión 4.0 maneja catorce lenguajes, entre ellos español, inglés, catalán, alemán, francés, italiano, portugués, ruso, noruego, galés, gallego, croata, asturiano y esloveno [Marimon, 2018]. Una vez establecido el idioma con el que se trabajará, recibe una oración y desambigua la categoría morfológica de cada palabra. Cuenta con dos tipos de etiquetadores: uno basado en modelos ocultos de Márkov [Brants, 2000] y otro basado en *relaxation labelling* [Padró, 1997]. Éstos permiten la combinación de información estadística con reglas manuales. Una vez etiquetada la oración, utiliza WordNet para añadir información sobre los posibles sentidos de cada palabra. Posteriormente, por medio de un desambiguador, ordena por relevancia los posibles sentidos de cada palabra. Aplica un *chart parser* para enriquecer cada oración con un árbol de análisis. Una vez que tiene la oración analizada sintácticamente, por medio de un analizador de dependencias utiliza un conjunto de reglas escritas manualmente que operan en tres etapas: primero completan el árbol sintáctico superficial construido por el *chart parser*, a continuación, transforman el árbol de constituyentes a dependencias, y finalmente etiquetan la función de cada dependencia.

NaLIR, desarrollada por [Li, 2014], es una interfaz genérica de LN que permite hacer consultas a bases de datos relacionales. Acepta oraciones complejas en inglés como consultas, las cuales traduce primero a una instrucción en SQL que puede incluir funciones de agregación, anidamiento y otros tipos de combinaciones. Después evalúa dicha consulta en un sistema de administración de BDs relacional o RDBMS (*Relational Database Management System* por sus siglas en inglés) y muestra el resultado al usuario. Utiliza un analizador de LN listo para usarse para generar un árbol de análisis sintáctico. Además, cuenta con un mapeador de nodos, el cual utiliza el árbol de análisis sintáctico para identificar nodos en el árbol de análisis que puedan ser relacionados a componentes de SQL y los convierte en unidades léxicas diferentes. Después del mapeo, se supone que el sistema comprende cada nodo de la consulta, por lo que el siguiente paso es comprender correctamente la estructura del árbol desde la perspectiva de la base de datos. En caso de que sea incorrecto o presente alguna ambigüedad, se aplica un ajuste a la estructura del árbol de análisis, reformulando los nodos del árbol de análisis sintáctico para que el sistema lo pueda comprender e insertando nodos implícitos en el árbol de análisis bajo la supervisión del usuario. Una vez realizados estos dos pasos, se obtiene la interpretación exacta de la consulta que se traducirá a una instrucción en SQL con poca ambigüedad. Dado el árbol de análisis correcto y validado por el usuario, el traductor utiliza su estructura para generar la instrucción adecuada de la expresión en SQL.

En [Sujatha, 2016] se describe una ILNBD que está basada en un sistema genérico de bases de datos. Inicialmente la interfaz recibe una consulta en LN, la cual es traducida a una consulta

estructurada. Este proceso consiste en varias fases. En la primera fase, se eliminan las palabras irrelevantes (*stop words*), utilizando una lista predefinida de palabras irrelevantes. En la segunda fase, las palabras restantes se procesan para obtener la palabra raíz. En la tercera fase, las palabras generadas en la segunda fase se consideran sumamente significativas y se les asigna una etiqueta POS utilizando una herramienta de LN. En la cuarta fase, para el análisis sintáctico, la interfaz utiliza un analizador descendente (*top-down*). El análisis de la consulta dada se realiza mediante la lógica de primer orden, aplicando la notación Backus-Naur para expresar la lógica de primer orden. En la quinta fase, el análisis semántico se realiza mediante el uso de ontologías y N-gramas. La ambigüedad en el significado de las palabras se resuelve utilizando la técnica de N-gramas y ontologías que se construye sobre la base de datos del cliente. Para finalizar, en la sexta fase, la formulación de consultas candidatas se realiza mediante el algoritmo EFFECN. Este algoritmo se ocupa de dividir la consulta natural, unir las tablas y seleccionar varias columnas y varias filas según las condiciones especificadas en la consulta.

La interfaz propuesta en [Kokare, 2014] almacena las consultas formuladas por los usuarios, así como sus traducciones correspondientes a SQL. De esta forma, cuando detecta una consulta que ya fue procesada, usa automáticamente la traducción en SQL y así evitar realizar todo el proceso de traducción nuevamente. Las consultas utilizadas para la evaluación de este sistema no están especificadas, sin embargo, reportan una precisión del 91.66%, así como un tiempo total de procesamiento de 5.4803 segundos de los cuales 0.5947 son del proceso de etiquetado y 4.8522 son del tiempo de traducción a una instrucción en SQL. La interfaz utiliza un análisis sintáctico, usando el método de dependencias, para analizar las oraciones que va a traducir. Inicia extrayendo las etiquetas POS y dependencias escritas. En el análisis de dependencias, su árbol de análisis conecta las palabras según la relación que exista entre ellas. Cada nodo del árbol representa una palabra, y los hijos son las palabras que dependen del nodo padre, las etiquetas en las aristas describen la relación entre padre e hijo. Al ser un análisis de dependencias, cuenta con tres principales ventajas. En la primera, los vínculos de dependencia que se forman entre dos palabras de la oración están cerca de las relaciones semánticas. En la segunda, los árboles de dependencia contienen nodos, donde cada nodo representa una palabra, por ende, la tarea de analizar resulta más sencilla. En la tercera, el análisis de dependencias trabaja con una palabra a la vez, es decir, no espera a que se cargue la oración completa para el análisis, mejorando de esta forma el tiempo de ejecución.

NLPwin es un proyecto desarrollado por Microsoft Research con el fin de dotar de herramientas de procesamiento de LN a Microsoft. El análisis se efectúa utilizando una mezcla entre modelos basados en sintaxis y modelos basados en frases. En [Vanderwende, 2015] se explica que el proceso de análisis sintáctico se realiza en dos pasos: en el primero se genera un bosquejo sintáctico, y en el segundo se crea un retrato sintáctico, donde se construye un árbol de constituyentes. Una vez que se cuenta con el árbol de constituyentes, es posible calcular cuál es la forma lógica. El objetivo es calcular la estructura del predicado-argumento para cada cláusula y a su vez normalizar las diferentes construcciones sintácticas de lo que puede considerarse el mismo "significado". Para mejorar la semántica y la sintaxis del árbol, se utilizan dos diccionarios: "Longman Dictionary of Contemporary English" (LDOCE) y el "American Heritage Dictionary", 3ra edición.

ASP es un analizador propuesto en [Krishnamurthy, 2014], el cual fue entrenado usando un enfoque de análisis semántico-sintáctico y empleando información de CCGbank y un corpus de

oraciones de Wikipedia, utilizando el vocabulario predictivo de NELL. Este analizador produce un análisis sintáctico completo (*full parsing*) de cualquier oración y simultáneamente produce formas lógicas para parte de la oración que tienen una representación semántica dentro del vocabulario predictivo del analizador. ASP utiliza un modelo de análisis gramatical categorial combinatorio (CCG). El analizador recibe una oración etiquetada por medio de un clasificador de regresión logística, el cual predice la categoría gramatical de cada unidad léxica a partir de las características de las unidades léxicas circundantes y sus etiquetas POS. El análisis posterior está restringido sólo a considerar categorías cuya probabilidad esté dentro de un factor de α de la categoría que tenga la puntuación más alta. El analizador da como resultado un árbol CCG de análisis sintáctico junto con cero o más formas lógicas que representan la semántica de la oración. Estas formas lógicas se construyen utilizando categorías y predicados de una amplia base de conocimientos que están almacenados en un lexicón. El analizador también genera una colección de estructuras de dependencia que resumen la estructura del predicado-argumento de la oración.

La Tabla 3.1 muestra un resumen de los aspectos más importantes de las ILNBDs encontradas en el estado del arte. La tabla muestra que algunas interfaces son multilenguaje, así como el tipo de analizadores que los sistemas tienen. Sólo cuatro de los sistemas reportan el corpus de consultas con el que trabajan.

Tabla 3.1. Resumen de trabajos relacionados

Analizador /Interfaz	Año	Idioma	Tipo	Analizador	Etiquetado	Análisis Semántico	Complejidad del corpus de consultas
Freeling	2011	14 idiomas	Interfaz genérica	Analizador de Dependencias	Modelos ocultos de Márkov relaxation labelling	WordNet	-
ASP	2014	Inglés	Sistema genérico	Analizador Sintáctico-Semántico	Utiliza un clarificador de regresión logística y etiquetas PoS	Utiliza un módulo de análisis gramatical categorial combinatorio (CCG) y un lexicón	CCGbank: moderada Wikipedia: moderada
NaLiR	2014	Inglés	-	Listo para usarse, el cual genera árboles de análisis	Mapeador de nodos que identifica y relaciona componentes de la consulta con elementos de SQL	Utiliza árboles de análisis validados por el usuario	Trabaja con BDs relacionales
NLPWin	2014	Varios Idiomas	Herramienta de procesamiento de LN de Windows	Árbol de Dependencias	Utiliza etiquetas PoS	Utiliza los diccionarios, *Longman Dictionary of Contemporary English (LDOCE)* y el *American Heritage Dictionary*	-
ILNBD de Kokare	2015	Inglés	Analizador semántico de aprendizaje	Método de Dependencias	Utiliza etiquetas PoS	-	Desconocido
ILNBD de Sujatha	2016	Inglés	-	Descendente	Elimina las palabras irrelevantes, utiliza la raíz de las palabras, así como etiquetas PoS	Utiliza ontologías y N-gramas para extraer el significado de las palabras usadas en la consulta	Desconocido
Este Proyecto	2020	Inglés	Interfaz de propósito general	Analizador Sintáctico-Semántico	Utiliza etiquetas PoS	Lexicón y reglas de producción	ATIS: alto

CAPÍTULO 4

Analizador de Inglés para una ILNBD

En este capítulo se presentan los siguientes temas: (a) el diseño de la BD del lexicón de inglés, así como el algoritmo para su implementación; (b) la descripción de la arquitectura general del analizador de la ILNBD en español descrita en [Verástegui, 2020], que es la base para el desarrollo de este proyecto; (c) las reglas gramaticales utilizadas para resolver el problema de ambigüedad léxica en el postprocesamiento; y (d) por último, las reglas de producción para inglés propuestas para el AS-S.

4.1 Lexicón de Inglés

En [Cervantes, 2005] se menciona que el primer paso para comprender el significado de una consulta en LN es la identificación de las palabras que la componen, por lo que la precisión en su identificación depende de la información almacenada en el lexicón.

Para alcanzar el objetivo al **OE1,** descrito en la Sección 1.3, se optó por utilizar un corpus de palabras previamente definido que cumpliera los siguientes requisitos: contener suficiente información gramatical para el análisis de la consulta, acceso gratuito, contar con una fuente de extracción de información variado, actualizado periódicamente, y por último que permitiera la recolección de una gran cantidad de palabras del idioma.

El corpus seleccionado fue COCA, descrito en la Subsección 2.2.2. Una vez definido el corpus base para el lexicón, se procedió a diseñar e implementar un programa que permitiera la extracción de las palabras de los archivos de tres géneros: textos académicos, periódicos y revistas. Esto es debido a que cuentan con un vocabulario más adecuado para el correcto análisis de las consultas en LN, ya que son fuentes que pasan por procesos de revisión gramatical antes de su publicación.

Algoritmo 4.1 Seudocódigo de la extracción de palabras para el lexicón

```
1:    Q // Archivos del corpus COCA, artículos académicos, noticias y revistas
2:    values[] // Almacena los datos, línea por la línea de los archivos del COCA
3:    words[] // Almacena los datos de entrada
4:    while (cadena ≠ null)
5:        words.add(values)
```

```
6:    for i = 0 to words.size()-1 do
7:        delete (signos de puntuación, números, url);
8:    end for
9:    ConexiónvBDs[] // Realiza la conexión con la BD del lexicón
10:   for i = 0 to words.size()-1 do
11:       INSERT (words, lemma, tag)
12:   end for
```

La implementación del Algoritmo 4.1 permitió la extracción y almacenamiento en la BD del lexicón de inglés de 49,414 palabras diferentes de los archivos de textos académicos, periódicos y revistas del corpus COCA.

Utilizando como base las categorías gramaticales definidas para el lexicón de español, descritas en la Subsección 2.2.3, y comparando ambos idiomas, se determinó que tanto en español como en el inglés existen categorías principales semejantes, por lo cual se decidió trabajar con ocho de las categorías principales del inglés. Al igual que en el lexicón de español, se crearon otras tablas para algunas subcategorías: sustantivos plurales, verbos auxiliares y participios. Esto es con el fin de evitar alterar estructuras dentro del código del analizador y evitar problemas de configuración del lexicón de inglés. La Tabla 4.1 muestra las categorías gramaticales incluidas en el lexicón de inglés.

Tabla 4.1. Categorías gramaticales del lexicón

adjective – adjetivo	**verb** – verbo
adverb – adverbio	**auxiliary verb** – verbo auxiliar
article – artículo	**participle** – verbo en participio
conjunction – conjunción	**preposition** – preposición
noun – sustantivo	**pronoun** – pronombre
plural noun – sustantivo plural	

4.1.1 Diseño del Lexicón

Para el diseño del lexicón, se decidió que la versión para inglés seguiría la misma estructura del lexicón de español [Aguirre, 2014], nuevamente con el fin de no alterar las estructuras definidas del analizador. Utilizando como base el corpus COCA y con la implementación del Algoritmo 4.1, se realizó la extracción de los morfemas de cada palabra, así como su etiqueta correspondiente.

La BD del lexicón de inglés está implementada en PostgreSQL 12 y está constituida por once tablas cuyos nombres hacen alusión a las tradicionalmente conocidas categorías gramaticales: sustantivos, artículo, adjetivo, pronombre, verbo, adverbio, preposición, conjunción, así como verbos auxiliares, sustantivos plurales y verbos en participio. La Figura 4.1 muestra las tablas que conforman el lexicón de inglés.

El lexicón cuenta con 49,414 palabras pertenecientes al inglés americano [Davies, 2009]. Las palabras se dividen en 339 adjetivos, 1,466 adverbios, 6 artículos, 29 verbos auxiliares, 34

conjunciones, 18,820 sustantivos, 1,909 verbos en participio, 6,647 sustantivos plurales, 201 preposiciones, 72 pronombres y 19,891 verbos.

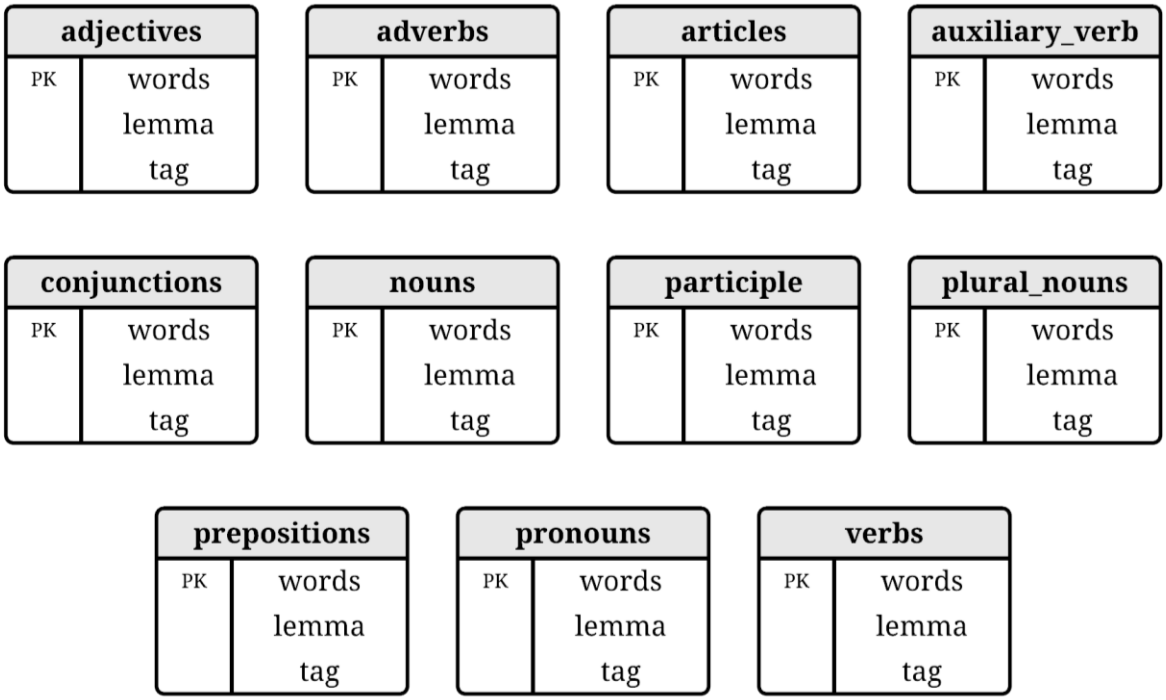


Figura 4.1. Diseño del lexicón

4.2 Arquitectura General del Analizador de Inglés

En [Verástegui, 2020] **se presenta el desarrollo de una versión completamente nueva de un analizador para una ILNBD**, el cual utiliza reglas de producción sintáctico-semánticas que permiten la generación de reducciones de palabras que facilitan el proceso de análisis de una consulta en LN.

Para alcanzar el objetivo **OE2**, descrito en la Sección 1.3, se realizó el estudio de la arquitectura del analizador de español. La Figura 4.2 muestra el resultado de la interpretación del diseño de esta arquitectura, la cual es usada como base para la adaptación e implementación de un analizador de inglés. En ella se pueden observan los diferentes módulos del analizador. Éste recibe una consulta en LN introducida por un usuario, la cual, en el caso de este proyecto, se encuentra en el idioma inglés. A partir de este punto, la consulta es analizada a través de los diferentes procesos tanto del analizador léxico como por el AS-S.

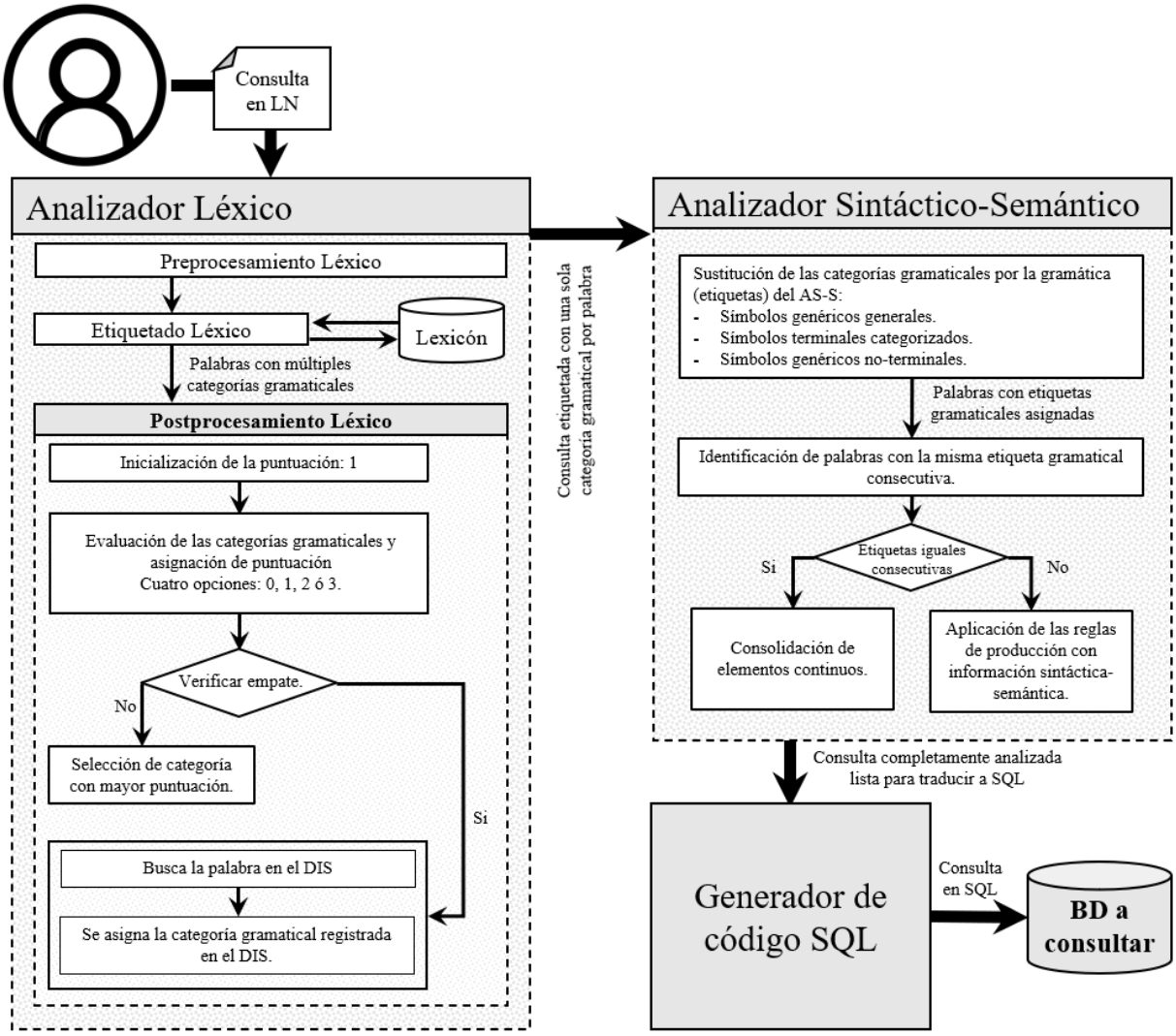


Figura 4.2. Arquitectura del analizador de inglés

El procesamiento de la consulta en LN se lleva a cabo a través de diferentes etapas en dos módulos:

1) Módulo del analizador léxico.
   i. Preprocesamiento léxico.
   ii. Etiquetado léxico.
   iii. Postprocesamiento léxico.

2) Módulo del analizador sintáctico-semántico.
   i. Determinación de categorías gramaticales (etiquetas) del AS-S.
   ii. Consolidación de etiquetado.
   iii. Aplicación de reglas de producción para inglés.

4.3 Módulo del Analizador Léxico

El analizador léxico es el primer paso en el procesamiento de la consulta en LN, por lo que cada palabra o unidad léxica se trata como un elemento independiente, el cual contiene un conjunto de columnas que se utilizan para el análisis de la consulta. En este módulo se realiza la asignación de la(s) categoría(s) gramatical(es) correspondiente(s) a cada una de las palabras que conforman la consulta.

4.3.1 Preprocesamiento Léxico

En inglés, al igual que en el español, el usuario cuenta con una gran libertad al usar los diferentes signos de puntuación (punto, coma, punto y coma, comillas, paréntesis, etc.), debido a que no cuentan con un orden estricto de aplicación dentro de una oración. Por lo tanto, al igual que la versión para español descrita en [Cervantes, 2005], el primer paso que se aborda dentro del análisis léxico es la eliminación de todo signo de puntuación detectado dentro de la consulta en LN.

Posteriormente, al igual que la versión para español, se realizó la aplicación de expresiones regulares definidas en [Verástegui, 2020], las cuales se describen en la Subsección 2.4.1. Las expresiones regulares, mostradas en la Tabla 2.4 en el Capítulo 2, facilitan la identificación de los valores de búsqueda al ser aplicadas al inicio del análisis léxico en un preprocesamiento de la consulta.

La detección temprana de los valores de búsqueda evita que éstos sean incluidos en la búsqueda de las categorías gramaticales. El preprocesamiento también asigna la etiqueta de *word* a las palabras que no fueron identificadas como valores de búsqueda, facilitando su identificación al buscarlas en el lexicón para determinar su categoría gramatical. Este proceso se muestra de forma gráfica en la Figura 4.3 con la consulta *Can I see the cost for flight 2 from SFO to LAX?*

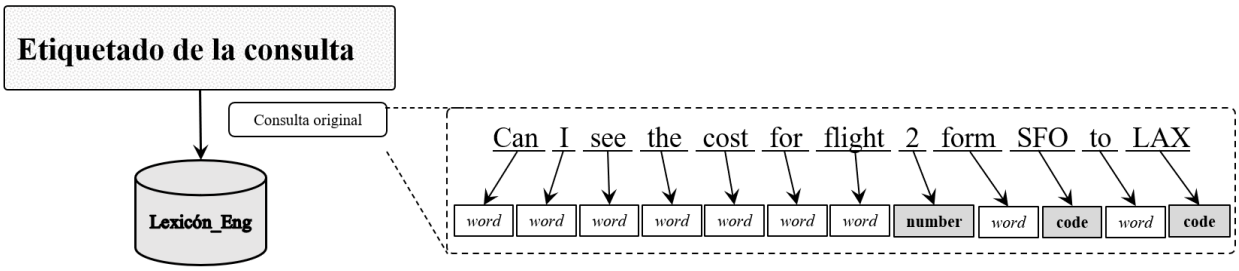


Figura 4.3. Identificación de valores de búsqueda

4.3.2 Etiquetado Léxico

El etiquetado léxico consiste en la asignación de etiquetas por medio de un proceso de extracción de la BD del lexicón de las categorías gramaticales de cada unidad léxica de la consulta. Este proceso se lleva a cabo una vez concluido el preprocesamiento léxico, usando los valores de búsqueda etiquetados e identificadas todas las palabras que se buscarán en el lexicón.

La Figura 4.4 muestra la consulta *Can I see the cost for flight 2 from SFO to LAX?* con los valores de búsqueda identificados, específicamente dos claves (*SFO*, *LAX*) y un número (*2*). El resto de las palabras fueron identificadas con la etiqueta *word*, pero aún no cuentan con una categoría gramatical asignada.

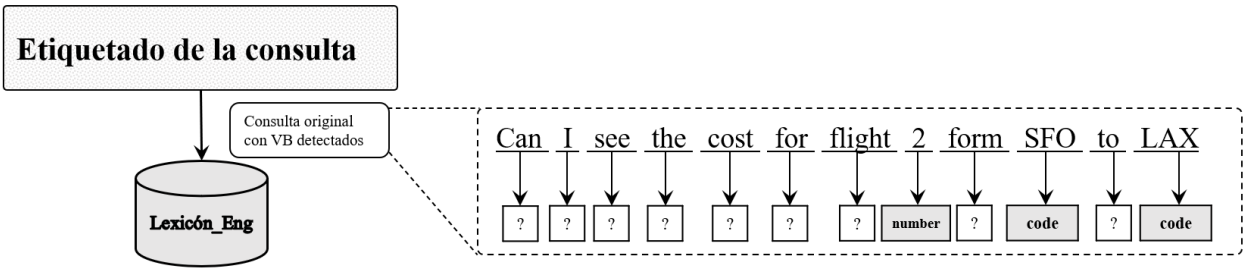


Figura 4.4. Proceso de etiquetado

Para la búsqueda de las palabras identificadas como *word*, se realizaron los cambios en el código para que el Algoritmo 2.1, definido en la Subsección 2.4.2, efectuara la extracción de las categorías gramaticales de éstas de la BD del lexicón de inglés. La Figura 4.5 muestra el resultado de este proceso con la consulta *Can I see the cost for flight 2 from SFO to LAX?*

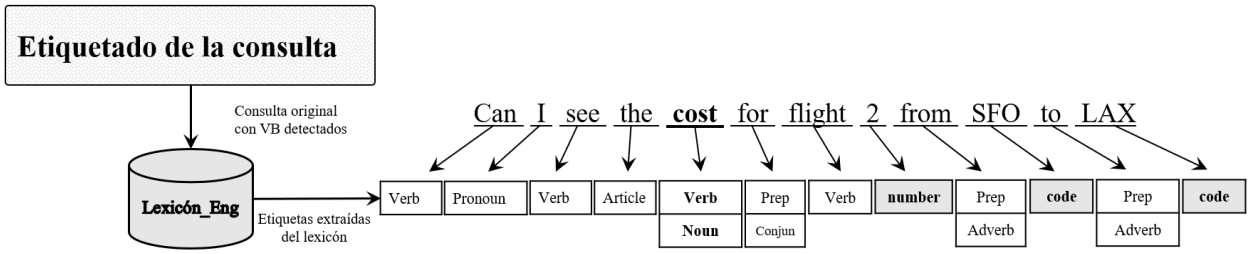


Figura 4.5. Extracción de etiquetas gramaticales

Al concluir el proceso de etiquetado, la consulta original se encuentra dividida en unidades léxicas que representan cada una de las palabras que la conforman. Éstas pueden tener asignada una o varias categorías gramaticales. La Figura 4.5 muestra un ejemplo en palabras como *cost*, *for*, *from* y *to*, las cuales cuentan con dos categorías gramaticales asignadas.

4.3.3 Postprocesamiento Léxico

El postprocesamiento está basado en una heurística propuesta en [Verástegui, 2020] por medio de la cual se resuelve la ambigüedad léxica. Es decir, cuando existen palabras que cuentan con múltiples categorías gramaticales asignadas, únicamente se aplica a aquellas palabras que presenten ambigüedad léxica. Un ejemplo se puede observar en la Figura 4.5 en la palabra ***cost***, ya que ésta tiene dos categorías gramaticales: verbo y sustantivo.

La heurística consiste en la aplicación de reglas gramaticales del idioma con el que se esté trabajando. Cuando se encuentran palabras con múltiples categorías gramaticales, se examinan tanto las palabras anteriores como posteriores de la palabra que se esté evaluando. Esto es con el fin de establecer una puntuación a cada una de las categorías gramaticales alternativas, como se define en la Subsección 2.4.3.

Para asignar la puntuación, deben aplicarse las reglas correspondientes de cada una de las categorías gramaticales. **En este proyecto se definieron las reglas basadas en las estructuras del idioma inglés** [Celce-Murcia, 1999a]**.** Las reglas para asignar una categoría gramatical para cada palabra se describen en Tabla 4.2.

Tabla 4.2. Reglas de inglés para el postprocesamiento

Adjetivos y pronombres interrogativos
Para asignar la categoría gramatical de los adjetivos y pronombres interrogativos, es necesario aplicar las siguientes reglas: – Si la palabra evaluada se encuentra en la primera posición y está seguida de una palabra con la categoría de verbo (*verb*), la categoría de pronombre interrogativo (*interrogative pronoun*) es aceptada. La puntuación asignada a esta categoría es de 3. – Si la palabra evaluada se encuentra en la primera posición y está seguida por una palabra con la categoría gramatical de sustantivo (*noun*), la categoría de adjetivo interrogativo (*interrogative adjective*) es aceptada, y se le asigna una puntuación de 3. – En caso de que la palabra no satisfaga las condiciones anteriores, la puntuación asignada para las categorías de adjetivo interrogativo (interrogative adjective) o pronombre interrogativo (interrogative pronoun) es de 0, según sea el caso.
Adjetivos y pronombres posesivos
Existen diferentes posesivos (adjetivos y pronombres). Algunos están seguidos de sustantivos (*nouns*) y otros pueden reemplazan a sustantivos (*nouns*) o pronombres (*pronouns*). Debido a lo anterior, es necesario asignar diferentes categorías para cada caso. Para determinar la categoría apropiada, es necesario aplicar las siguientes reglas: – El símbolo **poss_adj** se asigna a las siguientes palabras: *my*, *your*, *his*, *her*, *its*, *our* y *their*; siempre y cuando la palabra evaluada esté seguida de una palabra con la categoría gramatical de sustantivo (*noun*) o con la categoría de adjetivo (*adjective*). La categoría de adjetivo posesivo (*possessive adjective*) es aceptada, y se asigna el símbolo poss_adj y una puntuación de 2. – El símbolo **poss_pro** se asigna a las siguientes palabras: *mine*, *yours*, *his*, *hers*, *ours* y *theirs*; siempre y cuando la palabra en cuestión esté precedida por una palabra con la categoría de verbo (*verb*) o por una preposición (*preposition*). La categoría de pronombre posesivo (*possessive pronoun*) es aceptada, y se asigna el símbolo poss_pro y una puntuación de 2. – En caso de que la palabra no satisfaga las condiciones anteriores, la puntuación asignada para las categorías de adjetivo posesivo (*possessive adjective*) o pronombre posesivo (*possessive pronoun*) es de 0, según sea el caso.

Adjetivos y pronombres demostrativos

Los adjetivos demostrativos (*demonstrative adjective*): *this*, *that*, *these* y *those* también pueden ser pronombres demostrativos (*demonstrative pronoun*). Por otro lado, también pueden ser conjunciones (*conjunctions*). Por lo tanto, para determinar la categoría apropiada, es necesario aplicar las siguientes reglas:

- Si la palabra evaluada está seguida de una palabra con la categoría gramatical de sustantivo (*noun*) o la categoría gramatical de adjetivo (*adjective*) seguida de una palabra con la categoría gramatical de sustantivo (*noun*), la categoría de adjetivo demostrativo (*demonstrative adjective*) es aceptada con una puntuación de 3.
- Si la palabra evaluada está seguida de una palabra con la categoría gramatical de verbo (*verb*), la categoría de pronombre demostrativo (*demonstrative pronoun*) es aceptada con una puntuación de 3.
- Si la palabra evaluada es *that* y está precedida por una palabra con la categoría gramatical de verbo (*verb*) o por una palabra con la categoría gramatical de sustantivo (*noun*), o bien la palabra evaluada precede a un sustantivo (*noun*) o un nombre propio (*proper noun*), la categoría de conjunción (*conjunction*) es aceptada, y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la puntuación asignada para las categorías de adjetivo demostrativo (*demonstrative adjective*), pronombre posesivo (*possessive pronoun*) o conjunción (*conjunction*) es de 0, según sea el caso.

Pronombres y adjetivos indefinidos

Dentro de los pronombres indefinidos (*indefinite pronouns*), se encuentran las siguientes palabras: *all*, *another*, *any*, *anybody*, *anyone*, *anything*, *both*, *each*, *either*, *enough*, *everybody*, *everyone*, *everything*, *few*, *fewer*, *less*, *little*, *many*, *more*, *most*, *much*, *neither*, *nobody*, *none*, *nothing*, *other*, *several*, *some*, *somebody*, *someone*, *something* y *such*. Sin embargo, algunas de ellas también pueden ser adjetivos indefinidos (*indefinite adjectives*). Para determinar la categoría apropiada, es necesario aplicar las siguientes reglas:

- Si la palabra evaluada precede a una palabra con algunas de las siguientes categorías gramaticales: artículo (*article*), adjetivo (*adjective*), sustantivo (*noun*), pronombre (*pronoun*), verbo (*verb*) o la palabra *of*, la categoría de pronombre indefinido (*indefinite pronoun*) es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es alguna de las siguientes palabras: *all*, *another*, *any*, *both*, *each*, *either*, *enough*, *few*, *fewer*, *less*, *little*, *many*, *most*, *much*, *neither*, *other*, *several*, *some* y *such*, y está precedida por una palabra con la categoría gramatical de artículo (*article*) o preposición (*preposition*), o bien precede a una palabra con la categoría gramatical de sustantivo (*noun*), la categoría de adjetivo indefinido (*indefinite adjective*) es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es alguna de las antes mencionadas, se encuentra en la última posición de la oración y está precedida por una palabra con alguna de las siguientes categorías gramaticales: artículo (*article*), preposición (*preposition*) o verbo (*verb*), la categoría de adjetivo indefinido (*indefinite adjective*) es aceptada, y se le asigna una puntuación de 2.
- En caso de que la palabra no satisfaga las condiciones anteriores, la puntuación asignada para las categorías de pronombre indefinido (*indefinite pronoun*) o adjetivo indefinido (*indefinite adjective*) es de 0, según sea el caso.

Adverbios

Dentro de los adverbios existen algunos que pueden caer en otras categorías, por ejemplo: *all*, *any*, *enough*, *less*, *little*, *more*, *most* y *much*. Debido a lo anterior, para asignar la categoría de adverbio, es necesario aplicar las siguientes reglas:

- Si la palabra evaluada es *all* y está precedida por una palabra con alguna de las siguientes categorías gramaticales: adjetivo (*adjective*), adverbio (*adverb*), preposición (*preposition*) o conjunción (*conjunction*), la categoría de adverbio (*adverb*) es aceptada con una puntuación de 3.
- Si la palabra evaluada es alguna de las siguientes: *any*, *less*, *more*, *most* y *much*, y está precedida por una palabra con la categoría gramatical de adjetivo (*adjective*) o adverbio (*adverb*), o si la palabra evaluada precede a una palabra con la categoría gramatical de verbo (*verb*), la categoría de adverbio (*adverb*) es aceptada con una puntuación de 3.
- Si la palabra evaluada es *enough* y está precedida por a una palabra con la categoría gramatical de adjetivo (*adjective*) o adverbio (*adverb*), la categoría de adverbio (*adverb*) aceptada con una puntuación de 3.
- Si la palabra evaluada es *little* y está precedida por el artículo *a*, la categoría de adverbio (*adverb*) es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es *some* y precede a un número, la categoría de adverbio (*adverb*) es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es *in* y está precedida por una palabra con la categoría gramatical de sustantivo (*noun*), la categoría de adverbio (*adverb*) es aceptada con una puntuación de 3.
- Si la palabra evaluada está *out* y está seguida de una palabra cuya categoría gramatical sea preposición (*preposition*) o está precedida por cualquier versión del verbo *to be* (*is*, *are*, *were*, *was*, *been*, *being*, *will*), la categoría de adverbio (*adverb*) es aceptada y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de adverbio (*adverb*) no es aceptada, y se le asigna una puntuación de 0.

Conjunciones

Dentro de las conjunciones identificamos dos importantes: la conjunción copulativa *and*, a la cual se le asigna la etiqueta de **con_cop**, y la conjunción disyuntiva *or*, a la cual se le asigna la etiqueta de **con_dis**. La puntuación asignada para estas categorías (con_cop y con_dis) es de 3.

La palabra *that* tiene cuatro categorías gramaticales: adverbio (*adverb*), pronombre demostrativo (*demonstrative pronoun*), adjetivo demostrativo (*demonstrative adjective*) y conjunción (*conjunction*).

- La categoría de conjunción (*conjunction*) es aceptada, si la palabra *that* está precedida por una palabra con la categoría gramatical de verbo (*verb*) o con la categoría gramatical de sustantivo (*noun*). En este caso, se le asigna una puntuación de 3.
- Si la palabra *that* precede a una palabra con la categoría gramatical de sustantivo (*noun*) o con la categoría gramatical de nombre propio (*proper noun*), se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de conjunción (*conjuction*) no es aceptada, y se le asigna una puntuación de 0.

La palabra *than* tiene dos categorías gramaticales: conjunción (*conjunction*) y preposición (*preposition*).

- La categoría de conjunción (*conjunction*) es aceptada, si la palabra *than* está seguida de un número o precedida por una palabra con la categoría gramatical de sustantivo (*noun*). La puntuación asignada para esta categoría es de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de conjunción (*conjuction*) no es aceptada, y se le asigna una puntuación de 0.

Sustantivos

Para establecer la categoría gramatical de sustantivo, se consideran los siguientes casos:

- En caso de que la palabra evaluada no se encuentre en la primera posición, la categoría de sustantivo (*noun*) es aceptada.
- Si la palabra evaluada precede a una palabra con alguna de las siguientes categorías gramaticales: artículo (*article*), adjetivo (comparativo, descriptivo, demostrativo, indefinido, numeral, posesivo o superlativo), pronombre interrogativo, preposición (*preposition*), sustantivo (*noun*) o está precedida por un artículo, la categoría de sustantivo (*noun*) es aceptada, y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores la categoría de sustantivo (*noun*) no es aceptada, y se le asigna una puntuación de 0.

Verbos

Para definir la categoría gramatical de verbo, es necesario considerar los siguientes casos:

- Si la palabra evaluada precede a una palabra con alguna de las siguientes categorías gramaticales: artículo (*article*), verbo auxiliar (*auxiliary verb*), verb (*verb*), un nombre propio (*proper noun*), o está precedida por un adverbio (*adverb*) o por un verbo (*verb*), la categoría de verbo (*verb*) está aceptada con una puntuación de 3.
- Si la palabra en cuestión está precedida por un artículo (*article*) y si cuenta con una terminación en *-ing* (*gerundio*), la categoría de verbo (*verb*) es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es un verbo auxiliar y está precedida por una palabra con la categoría gramatical de verbo en participio, la categoría de verbo (*verb*) es aceptada, y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de verbo (*verb*) no es aceptada, y se le asigna una puntuación de 0.

Adjetivos

Para establecer la categoría gramatical de adjetivo, se deben considerar los siguientes casos:

- Si la palabra evaluada está precedida por una palabra con algunas de las siguientes categorías gramaticales: adverbio (*adverb*), artículo (*article*), verbo (*verb*) o preposición (*preposition*), la categoría de adjetivo es aceptada.
- Si la palabra evaluada está seguida de una palabra con la categoría gramatical de sustantivo (*noun*) o un nombre propio (*proper noun*), la categoría de adjetivo es aceptada, y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de adjetivo (*adjective*) no es aceptada, y se le asigna una puntuación de 0.

Preposiciones

Para definir la categoría gramatical de preposición, se consideran los siguientes casos:

- Si la palabra evaluada está seguida de una palabra con alguna de las siguientes categorías gramaticales: sustantivo (*noun*), adjetivo (*adjective*) o nombre propio (*proper noun*), la categoría de preposición es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada está precedida por una palabra con alguna de las siguientes categorías gramaticales: nombre propio (*proper noun*) o adverbio (*adverb*), la categoría de preposición es aceptada, y se le asigna una puntuación de 3.
- Si la palabra evaluada es *out* y está seguida de la palabra *of*, la categoría de preposición (*preposition*) es aceptada, y se le asigna una puntuación de 3.
- En caso de que la palabra no satisfaga las condiciones anteriores, la categoría de preposición (*preposition*) no es aceptada, y se le asigna una puntuación de 0.

Se considera importante mencionar que el problema de asignar una sola categoría gramatical para cada palabra con ambigüedad léxica es muy complejo [Jurafsky, 1996], y considerando que no es parte de los objetivos de este proyecto, se implementó una heurística para inglés cuyo funcionamiento y desempeño son aceptables.

Como se mencionó en la Subsección 2.4.3, la asignación de categorías se realiza a través de un análisis sintáctico local de las palabras más cercanas. Una vez que se tiene la consulta etiquetada, el postprocesamiento procede a la detección de palabras con múltiples categorías gramaticales. Si se detectan múltiples categorías gramaticales, se inicializan las puntuaciones de estas categorías en 1. La Figura 4.6 muestra el proceso de inicialización para la consulta: *Can I see the cost for flight 2 from SFO to LAX?* del corpus ATIS, la cual cuenta con cuatro palabras con múltiples categorías gramaticales: *cost*, *for*, *from* y *to*.

Can	I	see	the	cost	for	fligh	2	from	SFO	to	LAX
verb	pronoun	verb	article	verb 1 noun 1	preposition 1 conjunction 1	verb	number	preposition 1 adveb 1	code	preposition 1 adveb 1	code

Figura 4.6. Consulta con múltiples categorías gramaticales

Cuando las posibles categorías se encuentran inicializadas, se procede a la aplicación de las reglas gramaticales del idioma inglés, definidas en la Tabla 4.2, para la evaluación de éstas. En el siguiente ejemplo se trabaja con la palabra ***cost***. La palabra se evalúa con ambas categorías, es decir, ***cost*** como verbo y ***cost*** como sustantivo. La evaluación de las categorías de la palabra ***cost*** se lleva a cabo aplicando dos reglas: las de sustantivos y las de verbos, las cuales se muestran de forma gráfica en la Figura 4.7.

Regla para sustantivo	
Primera posición	puntuación = 0
Precedido por: artículo adjetivo pronombre interrogativo preposición sustantivo pronombre	puntuación = 3

Regla para verbo	
Precede a: artículo verbo o verbo auxiliar nombre propio	puntuación = 3
Precedido por: artículo y termina en *-ing* adverbio verbo	puntuación = 3

Figura 4.7. Reglas de evaluación de categorías gramaticales

La regla para asignar la categoría gramatical de sustantivo indica lo siguiente: si la palabra ***cost*** no está en la primera posición, la categoría de sustantivo tendrá asignada una puntuación de 3, sólo si está precedida por una palabra con alguna de las siguientes categorías: artículo, adjetivo (comparativo, descriptivo, demostrativo, indefinido, numeral, posesivo o superlativo), pronombre interrogativo, preposición y sustantivo, o si está precedida por un artículo. En caso de que la palabra ***cost*** se encuentre en la primera posición, la puntuación que se le asignará a esta categoría es de 0.

Mientras que la regla para verbos indica que si la palabra ***cost*** precede a una palabra con alguna de las siguientes categorías: artículo, verbo auxiliar, verbo, nombre propio, o está precedida por un adverbio, por un verbo o por un artículo (para este caso específico debe tener una terminación *-ing)*; a la categoría de verbo se le asignará una puntuación de 3. Como estas condiciones no se satisfacen, entonces la puntuación para la palabra ***cost*** como verbo sería de cero, mientras que para ***cost*** como sustantivo, sería de 3. La Figura 4.8 muestra la consulta con los puntajes asignados a las múltiples categorías durante el postprocesamiento.

Can	I	see	the	cost	for	fligh	2	from	SFO	to	LAX
verb	pronoun	verb	article	verb 0 **noun 3**	preposition 0 **conjunction 1**	verb	number	**preposition 3** adveb 1	code	**preposition 3** adveb 1	code

Figura 4.8. Consulta con múltiples categorías evaluada

Se selecciona la categoría con la mayor puntuación, ya que es la que tiene la mayor probabilidad de ser la correcta, por lo tanto, la palabra ***cost*** tendrá asignada la categoría de sustantivo. Al finalizar el proceso del módulo del analizador léxico, se tiene la consulta etiquetada con una sola categoría gramatical por palabra y con los valores de búsqueda identificados.

4.4 Módulo del Analizador Sintáctico-Semántico

El análisis sintáctico es el proceso de analizar un conjunto de caracteres, ya sea que se encuentren en LN, lenguaje computacional o estructuras de datos. Éstos se analizan conforme a las reglas formales de la gramática del lenguaje con el que se esté trabajando.

El principal objetivo es entender el significado exacto de una oración en LN, lo cual generalmente se realiza por medio de la división de las oraciones en sus constituyentes (palabras). Esto resulta en la generación de árboles de análisis de constituyentes, los cuales pueden contener tanto información semántica como de otro tipo que ayude al análisis de la oración [Aguirre, 2014].

La meta de este proyecto de tesis es demostrar que el enfoque para el analizador sintáctico-semántico propuesto en [Verástegui, 2020], como se define en la Sección 2.5, puede ser aplicado a idiomas europeos, **en particular la lengua inglesa**.

Lo primero para realizar la implementación de la versión del AS-S de inglés fue determinar si la gramática definida para la versión para español podía ser utilizada en la versión para inglés. Por lo tanto, se requirió identificar estructuras gramaticales similares, como se describe en la Subsección 2.5.1.

4.4.1 Asignación de Gramática (Etiquetas) del AS-S

Durante la revisión de la gramática inglesa, y como se menciona en la Subsección 2.2.1, se encontró que en inglés no existe un regulador oficial del idioma [Quirk, 1985]. No obstante, existen diferentes autoridades que instauran ciertos estándares para el idioma, autoridades como Oxford University y Cambridge University.

Utilizando la información proporcionada por dichas autoridades, se logró establecer que la gramática del analizador de español podía ser utilizada para el analizador de inglés, ya que en ambos idiomas existen estructuras gramaticales semejantes. Por lo tanto, la gramática para la versión para inglés del AS-S es la que se encuentra definida en las Tablas 2.5, 2.6 y 2.7 de la Subsección 2.5.1.

La gramática cuenta con símbolos terminales genéricos para las diferentes categorías gramaticales generales, símbolos terminales categorizados para identificar algunas subcategorías gramaticales de adjetivos, pronombres y conjunciones, además, incluye símbolos no terminales genéricos para los diferentes tipos de frases. Finalmente incluye tres símbolos no terminales para los elementos del sujeto (*Sbj*), complemento (*Com*) y oración (*Sentence*).

Con la gramática definida, se procede a la asignación del etiquetado establecido para las categorías gramaticales de la consulta. La Figura 4. 9 muestra de forma gráfica la sustitución de las categorías gramaticales por su respectiva etiqueta en la consulta *Can I see the cost for flight 2 from SFO to LAX?*

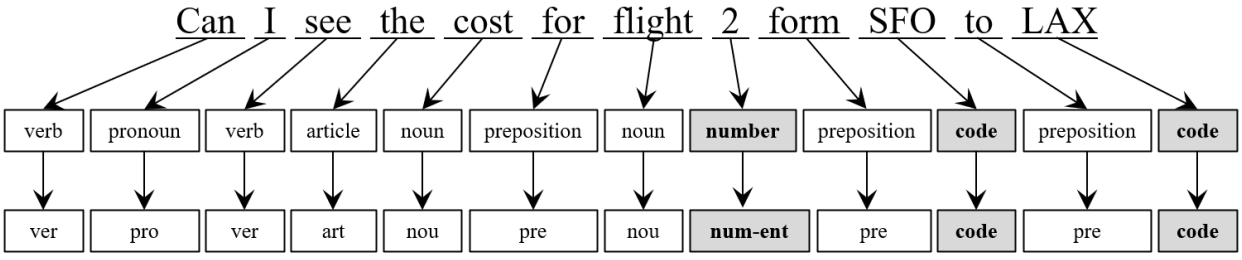


Figura 4.9. Asignación de etiquetas gramaticales

4.4.2 Consolidación de Etiquetado

Antes de iniciar el proceso de aplicación de las reglas de producción, en caso de encontrarse etiquetas iguales de forma consecutiva, se procede a realizar una consolidación de éstas. La Figura 4.10 muestra de forma gráfica la consolidación en la consulta *Give me all jets flights from CONTINENTAL AILINES*.

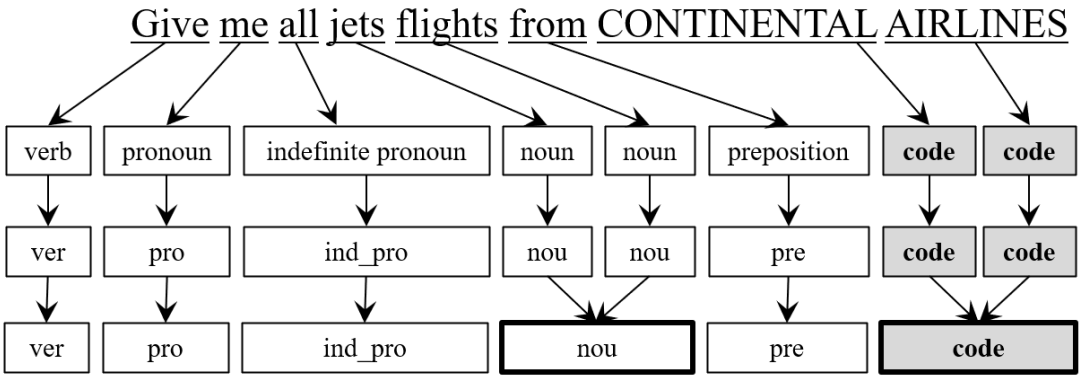


Figura 4.10. Consolidación de etiquetas gramaticales

4.4.3 Reglas de Producción para Inglés

Para definir las reglas de producción, fue necesario tener un entendimiento profundo de las unidades estructurales básicas de la lengua inglesa para determinar qué componentes de la oración en inglés son obligatorios, y qué constituyentes obligatorios u opcionales. En la Figura 4.11 se muestra el resumen definido en [Celce-Murcia, 1999b] de las 16 reglas estructurales que se utilizaron como base.

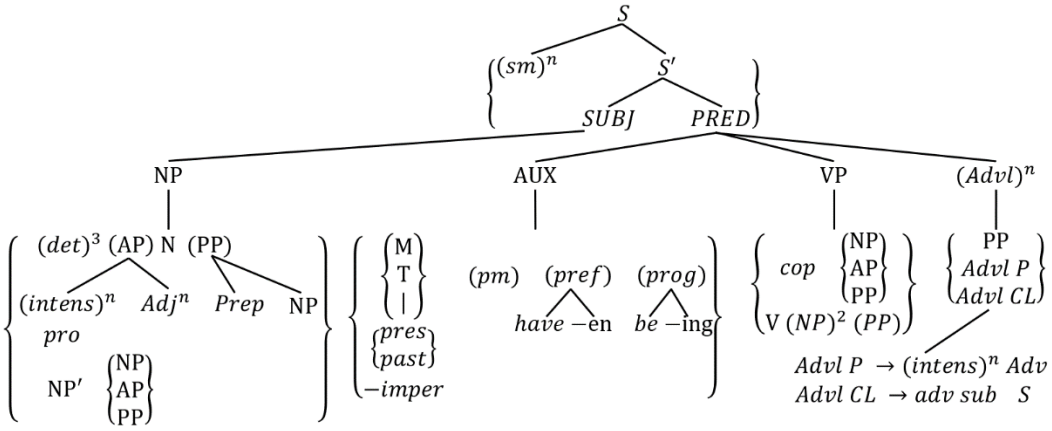


Figura 4.11. Resumen de las reglas estructurales del inglés

Utilizando las reglas estructurales de la gramática inglesa antes descritas, en la Figura 4.11 y en la Sección 2.62.5.2, así como parte de las características particulares de las categorías gramaticales del inglés, se definieron las reglas de producción del AS-S para inglés con el objetivo de realizar reducciones de las consultas en LN. Para expresar las diferentes categorías y frases, se utilizaron los símbolos que se presentaron en la Subsección 2.5.1 (Tablas 2.5, 2.6 y 2.7). Así mismo, al igual que las reglas para español descritas en [Verástegui, 2020], la nomenclatura utilizada en las reglas de producción está descrita en [World Wide Web Consortium, 1982]. La Tabla 4.3 muestra las reglas de producción para inglés del AS-S.

Tabla 4.3. Reglas de producción para inglés

ID	Regla de producción
Reglas de producción para valores de búsqueda	
SN01-V	<NouP0V> ::= <cde> | <dec> | <int> | <dat> | <hour> | <name>
SN02-V	<NouP1V> ::= <art> NouP0V
SN03-V	<NouP1V> ::= [<art>] (<dec> | <int>) <nou>
Reglas generales de producción	
SN02	<NouP0> ::= [<adj_com>|<adj_cal>|<adj_sup>|<adj_pos>|<adj_dem>|<adj_num>|<qua_adj>](<nou>| <pro>)
SN03	<NouP1> ::= ([<ind_pro>|<adv>]<art>[<adj_sup>|<aj_cal>]|<adj_sup>|<adj_cal>)<NouP0>[<ind_pro>| <adv>]
SA01	<AdjP> ::= <adj_cal> <"than">
SA02	<AdjP > ::= "as" <adj_cal> ["as"]
SA03	<AdjP > ::= <adj_sup><nou>
SA04	<AdjP > ::= <adv> <adj_cal>
SA06	<AdjP> ::= <art> (<adj_cal> | <adj_com> | <adj_sup> | <adj_num>)
SA07	<AdjP0> ::= <adj_cal> | <adj_sup> | <adj_num>
SA08	<AdjP> ::= ("more" | "less") “than” <dec> | <int>
SN09	<NouP1> ::= <AdjP> (<NouP1> | <NouP0>)
SA10	<AdjP > ::= <adj_cal><pre>
SV01	<VerP1> ::= <ver> (<NouP1> | <NouP0>)
Reglas de producción para valores de búsqueda	
SN10-V	<NouP1V> ::= [(<NouP1> | <NouP0>)] 1#<NouP0V>
SN11-V	<NouP1V> ::= (<NouP1V> | <NouP0V> | <PrePV>)(<con_cop>|<con_dis>)(<NouP1V> | <NouP0V>)
Reglas generales de producción	
SP01	<PreP> ::= <pre> (<NouP1> | <NouP0>)
Reglas de producción para valores de búsqueda	
SP01-V	<PrePV> ::= <pre> <NouP1V>
Reglas generales de producción	
SA11	<AdjP > ::= <adj> (<PreP> | <PrePV>)
SV02	<VerP1> ::= <pro_int> | <VerP1> [<ver> <qua_adj> <ver>]
SV03	<VerP1> ::= (<VerP1>) [<adv> | <nou> | <AdjP0>] [<adv> <adj_cal>]
SV04	<VerP1> ::= <VerP1> < AdjP >
SV11	<VerP3> ::= <VerP1> [<PreP>] 1#(<con_cop> <VerP1>)
SN10	<NouP2> ::= (<NouP1> | <NouP0>) "with" (<NouP1> | <NouP0>)
SN11	<NouP2> ::= (<NouP1> | <NouP0>) "than" | “that” <VerP1> <NouP0>
SN14	<NouP2> ::= (<PreP> | <VerP1>) "than" | “that” <VerP1> <NouP0>
SN15	<NouP2> ::= (<VerP1>) "than" | “that” <VerP1> <NouP0>
SN12	<NouP3> ::= (<NouP1>|<NouP0>)*(","(<NouP1>|<NouP0>))(<con_cop>|<con_dis>)(<NouP1> | <NouP0>)
SN13	<NouP3> ::= (“,” (<NouP0> | <NouP1> | <NouP1V>))
Reglas de producción para valores de búsqueda	
SP02-V	< PrePV > ::= <PrePV> (<con> | “that”) <VerP1>
SN12-V	<NouP2V> ::= (“,” (<NouP0V> | <NouP1V>) [“that”])
Reglas Terminales	
TS01	<Sbj> ::= <epsilon> (<adj_int> | <pro_int> | <art>) <NouP0> | [NouP0]
TS02	<Sbj> ::= <epsilon> (<aux_ver> (<NouP0> | NouP1 | NouP2)
TC01	<Com> ::= [<adv>] (<PreP> | <PrePV>)
TC02	<Com> ::= (<NouP1> | <NouP0> | <NouP2> | <NouP3> | <NouP1V> | <NouP2V> | <VerP1>)

TO01	<Sentence> ::= <Sbj> (<VerP1> | <VerP3>) *<Com> <epsilon>
TO03	<Sentence> ::= <Sbj> *<Com> (<VerP1>)[*<Com>] <epsilon>
TO02	<Sentence> ::= <epsilon> (<VerP1> | <VerP3>) *<Com> <epsilon>
Nota: <*epsilon*> representa una unidad léxica nula para indicar el inicio o final de una expresión en LN. **Nota:** la notación utilizada para las reglas está descrita en [World Wide Web Consortium, 1982].	

De la misma forma que la versión para español, según se describe en [Verástegui, 2020], las reglas de producción para inglés tienen prioridad al momento de ser aplicadas, es decir que algunas reglas son aplicadas antes que otras. Para tal efecto, el analizador utiliza una técnica multipasada, la cual consiste en un recorrido por cada regla según la prioridad definida.

El orden de aplicación de las reglas de producción es el que se muestra en la Tabla 4.3. Se considera importante enfatizar que **el AS-S de inglés**, al igual que su versión para español, fue diseñado para expresiones en LN para consultar BDs, por consiguiente, **sólo trabaja con oraciones del tipo interrogativo e imperativo.**

Para los objetivos de este proyecto, en el caso del AS-S de inglés, a diferencia de la versión para español, el Algoritmo 2.3 cuenta con las reglas terminales implementadas. Las reglas terminales permiten observar si las consultas son correctamente reducidas, es decir, si se reducen hasta el símbolo terminal *Sentence* para determinar si una oración es gramaticalmente correcta o no. Sin embargo, en [Verástegui, 2020] se determinó que las reglas terminales no son necesarias para realizar el análisis semántico de los valores de búsqueda.

Adicionalmente, en este proyecto se diseñaron procesos dentro de algunas reglas de producción que permiten detectar uno de los errores gramaticales más comunes: la ausencia de un verbo principal. Con esta adaptación, el AS-S de inglés logra identificar con éxito cuando una oración carece de un verbo o frase verbal principal.

El Algoritmo 2.3, definido en la Subsección 2.5.2, recibe como entrada una secuencia de símbolos que representan la categoría gramatical de cada palabra (*inputExpression*). Cada línea del algoritmo llama al método de la regla específica que menciona. El método recorre la secuencia de símbolos (terminales y no terminales) e intenta encontrar una subsecuencia de símbolos que pueda ser reducida. Si existe una subsecuencia de símbolos para reducir, ésta es reducida, de lo contrario, continúa recorriendo la secuencia de símbolos. Finalmente, el método almacena cada reducción en una matriz (Tabla 4.4).

Tabla 4.4. Secuencia de reducciones sintácticas

Paso:	**Símbolos (terminales y no terminales)**
0:	[pro_int, ver, art, nou, pre, name]
1:	SN01-V pro_int, ver, art, nou, pre, **NouP0V**
2:	SN02 pro_int, ver, art, **NouP0**, pre, **NouP0V**
3:	SN03 pro_int, ver, **NouP1**, pre, **NouP0V**
⋮	⋮
n:	Sentence

La Tabla 4.4 muestra un ejemplo de las reducciones generadas al aplicar las reglas de producción. En la tercera reducción, se puede observar que el método de la regla SN03 (línea 6 del Algoritmo 2.3) encontró la subsecuencia de símbolos *art* y *NouP0* (tercer y cuarto símbolo, línea 2 de la tabla) para ser reducidos a *NouP1* (línea 3 de la tabla). Una vez que el método termina, el algoritmo continúa con el método de la siguiente regla hasta concluir con las reducciones. Al finalizar, la matriz contiene todas las reducciones que se pudieron realizar a la expresión de entrada.

Algoritmo 4.2 Pseudocódigo de la regla SV01 para inglés

```
1:  Procedimiento SV01 (inputExpression) // Regla SV01 <VerP1> ::= <ver> (<NouP1> | <NouP0>)
2:  for i = 0 to n-1 do    // Para cada unidad léxica de la consulta Q
3:    if inputExpression [i] = ("ver") and inputExpression [i+1] = ("NouP1" or "NouP0") then
4:      updateInputExpression (i, "VerP1")
5:      reduce (inputExpression, i) // Elimina las unidades léxicas que fueron reducidas
6:      addMatrix ("SV01") // Inserta el identificador de la regla a la matriz
7:      addMatrix ("inputExpression") // Inserta la reducción a la matriz
8:    end if
9:  end for
10: return inputExpression
```

El Algoritmo 4.2 muestra el pseudocódigo de uno de los métodos de las reglas de producción para inglés, específicamente la regla <VerP1> ::= <ver> (<NouP1> | <NouP0>). Esta regla indica que, en caso de la detección de un verbo (*ver*) y una frase nominal (*NouP0* o *NouP1*), ambos símbolos se reducen a VerP1 (frase verbal del tipo 1). Una vez detectada la reducción (línea 3), *inputExpression* se actualiza (línea 4) y los símbolos reducidos son eliminados de *inputExpression* (línea 5).

Para finalizar, el identificador de la reducción e *inputExpression* actualizada son almacenados en la matriz (lineas 6 y 7). La Figura 4.12 muestra de forma tabular el proceso de reducción con la consulta *Can I see the cost for flight 2?*, del corpus ATIS.

	Can	**I**	**see**	**the**	**cost**	**for**	**flight**	**2?**
	verb	pronoun	ver	article	noun	preposition	noun	number
	ver	pro	ver	art	nou	pre	nou	num-ent
SN01-V	ver	pro	ver	art	nou	pre	nou	**NouP0V**
SN02	ver	**NouP0**	ver	art	nou	pre	nou	NouP0V
SN02	ver	NouP0	ver	art	**NouP0**	pre	nou	NouP0V
SN02	ver	NouP0	ver	art	NouP0	pre	**NouP0**	NouP0V
SN03	ver	NouP0	ver	**NouP1**		pre	NouP0	NouP0V
SV01	**VerP1**		ver	NouP1		pre	NouP0	NouP0V
SV01	VerP1		**VerP1**			pre	NouP0	NouP0V
SN10-V	VerP1		VerP1			pre	**NouP1V**	
SP01-V	VerP1		VerP1			**PreP**		
TC01	VerP1		VerP1			**Com**		
TO03	S E N T E N C E							

Figura 4.12. Aplicación de la regla para inglés <VerP1> ::= <ver> (<NouP1> | <NouP0>)

CAPÍTULO 5

Experimentación y Resultados

En este capítulo se presentan las diferentes pruebas de funcionalidad realizadas a los diferentes componentes del analizador de inglés. El objetivo de la experimentación es la validación del funcionamiento tanto del analizador léxico como del analizador sintáctico-semántico con el fin de validar las hipótesis y objetivos de este proyecto, presentadas en el Capítulo 1.

El capitulo tiene la siguiente estructura: descripción del corpus de pruebas, detalles del software y hardware, pruebas de funcionalidad del analizador léxico, pruebas de funcionalidad del AS-S y pruebas comparativas tanto del analizador léxico como del AS-S.

5.1 Descripción del Corpus de Pruebas

Las diferentes pruebas se llevaron a cabo utilizando un corpus de pruebas formado con consultas de dos corpus diferentes. El primero de ellos es el corpus del "Airline Travel Information System" (ATIS) [Linguistic Data Consortium, 1990]. Este corpus proporciona una gran cantidad de transcripciones manuales de usuarios que solicitan información de vuelos en sistemas automatizados de consultas de viajes de aerolíneas [Hemphill, 1990]. Para más información de la estructura de la BD de ATIS, consultar el Apéndice A.

El segundo corpus de consultas es el de Geoquery250, el cual consiste en 250 preguntas sobre geografía de los Estados Unidos. Las preguntas están relacionadas con ciudades, ríos, colindancias, montañas, carreteras y lagos [Machine Learning Research Group, 2004]. Para más información de la estructura de la BD de Geoquery, consultar el Apéndice B.

El corpus de pruebas está formado por 100 consultas, específicamente, se utilizaron 75 consultas de ATIS y 25 consultas de Geoquery, como se muestra en la Tabla 5.1. Se considera importante mencionar que ambos corpus cuentan con una gran cantidad de consultas que son muy similares, ya que son variaciones de otras consultas, es decir que sólo difieren en los valores de búsqueda y en la estructura de la consulta. Con el fin de ahorrar espacio, se seleccionaron 20 consultas (10 de Geoquery y 10 de ATIS) representativas de la mayoría de las consultas para mostrar los resultados en los Apéndices C, D, …, I. Las consultas seleccionadas se indican con un asterisco en la Tabla 5.1.

Tabla 5.1. Selección de muestra de un corpus de 100 consultas

Geobase	
No.	**Consulta y valores de búsqueda.**
1	How high is the highest point in America?
2*	How many people live in Hawaii?
3	How many rivers are there in Idaho?
4*	What is the smallest city in the USA?
5*	Show the rivers in the state of Florida.
6*	What are the major cities in Ohio?
7*	What is the area of Wisconsin?
8	What is the biggest city in Louisiana?
9	What is the biggest city in Oregon?
10	What is the capital of the state with the highest point?
11	What is the highest point in Montana?
12*	What is the highest point in Nevada in meters?
13	What is the highest point in the US?
14	What is the length of the Colorado river?
15	What is the lowest point in Louisiana?
16*	What is the lowest point of the state with the largest area?
17	Which state has the highest elevation?
18	What is the shortest river in the US?
19*	What is the shortest river in the USA?
20*	What rivers are in Utah?
21	What rivers run through New York?
22	What state contains the highest point in the US?
23*	Which state borders Florida?
24	Which state has the highest elevation?
25	Which state has the smallest population density?
ATIS	
No.	**Consulta y valores de búsqueda.**
26	Can I see the cost for flight 2?
27	Do you have airlines with flights in turboprop?
28	Do you have jets that fly from Denver to Dallas with discount?
29	Do you have jets that fly from Denver to Dallas?
30*	Does Delta Airlines flight 179 serve breakfast?
31	Does flight 640 serve lunch?
32	Give me a list of all the flights from Dallas to Denver.
33	Give me a list of flights from Denver to San Francisco.
34	Give me all discount flights from Philadelphia to Washington on wide-body non-pressurized aircraft.
35*	Give me all flights from Atlanta to San Francisco on Delta first class.
36	Give me all jets flights from Continental Airlines.
37	Give me all wide-body flights from Dallas to Boston.
38	Give me the airlines that have amphibian-type aircrafts.
39	Give me the application for flight with stops without discount.
40	Give me the application for flights with stops.
41	Give me the class type for flights with discount.
42	Give me the classes with discount.
43	Give me the flight codes for flights with dual carrier.
44*	Give me the flights in amphibious-type wide-body aircraft.
45	Give me the flights in turboprop type aircrafts.
46	Give me the flights in wide-body amphibian-type aircrafts.
47	Give me the flights in wide-body jets.
48*	Give me the flights of the airline Tokened Airlines that serve breakfast.
49	Give me the restrictions for flights with no stops.
50	Is breakfast served in flight 343 from Tokened Airlines?
51	Is dinner served in flight 852 departing from San Francisco?
52	Is dinner served in flight from Philadelphia to Oakland?

53	Is flight 640 turboprop?
54	Is lunch served in flight 1750?
55	Is lunch served in flight 539 of Avianca?
56	List afternoon flights from Atlanta to San Francisco.
57	List the flights with breakfast from Pittsburgh to Boston.
58	May I see flights from San Francisco to Los Angeles?
59	Premium class flights from ATL to PIT.
60	Show flights from Philadelphia to Dallas/Fort Worth.
61	Show me all American Airline's flights out from Oakland to DFW.
62	Show me all flights in jets that serve dinner.
63	Show me all the flights from Dallas to Denver with breakfast.
64	Show me all the flights in turboprop.
65*	Show me all the flights where breakfast is served in wide-body airplanes.
66	Show me all the United flights from San Francisco to Boston.
67*	Show me all the wide-body flights from Dallas to Denver that serve lunch.
68	Show me only flights in jets that serve breakfast.
69	Show me the aircrafts that serve dinner.
70	Show me the airlines that serve breakfast from San Francisco to Denver.
71	Show me the lowest round-trip fare from Oakland to Philadelphia, class M, no restrictions.
72	Show me the number of engines for aircrafts with no pressurization.
73	Show the number of engines for aircraft that do not have pressurization and are wide-body.
74	What afternoon flights are available from Washington to Boston with meals?
75	What airlines are dual carrier?
76	What flights from Dallas to San Francisco or Oakland use restriction AP/80?
77	What flights leave from SFO for LAX after noon?
78	What is the fare difference between class QW and QX?
79	What is the price of a round-trip ticket Q class on Delta flight number 317?
80	What type of aircraft is US 1750?
81*	Which aircraft types are not wide-body and are not pressurized?
82	Which aircraft types are not wide-body and are pressurize?
83	Which aircraft types are wide-body and are pressurize?
84	Which airlines are not dual carrier?
85	Which airlines do non-pressurized aircrafts?
86	Which airlines have economy flights from San Francisco to Dallas.
87*	Which airlines have economy flights with discount from San Francisco to Dallas?
88	Which airlines have flight with stops from San Francisco to Dallas?
89	Which airlines have flights that serve dinner in wide-body jets?
90	Which airlines have flights that serve dinner?
91	Which airlines have flights that serve lunch in jets?
92*	Which airlines have flights that serve lunch in wide-body jets?
93	Which airlines have flights with breakfast?
94	Which airlines have wide-body planes/aircraft?
95	Which flights are in wide-body aircraft and are not in dual carriers?
96	Which flights are in wide-body airplanes?
97	Which flights have discounted fare?
98*	Which flights have no discounted fare.
99	Which is the restrictions AP/57 or AP/80?
100	Which types of aircrafts are not wide-body?

5.2 Detalles del Hardware y Software

Para las pruebas de funcionalidad, se utilizaron las siguientes herramientas.

El equipo donde se llevaron a cabo las pruebas de funcionalidad es una laptop Lenovo Legion Y720, la cual cuenta con un procesador Intel(R) Core (TM) i7-7700HQ, CPU a 2.80GHz de 4

núcleos, una memoria RAM de 8GB y sistema operativo Microsoft Word 10 Home de 64 bits. Los algoritmos se implementaron en Java 15.0.1 con Apache NetBeans IDE 12.4. Las BDs del lexicón, ATIS y Geobase fueron implementadas en el administrador de BDs PostgreSQL 12.

5.3 Resultados de Pruebas del Analizador Léxico

Para la evaluación del funcionamiento de los procesos del analizador de inglés, se diseñaron diferentes experimentos que cubren todos los aspectos que se deseaban evaluar. En cada subsección se definen los experimentos realizados en esa etapa.

5.3.1 Preprocesamiento

Experimento 1: se busca verificar si el preprocesamiento efectúa la correcta identificación de los valores de búsqueda. Dentro del corpus de 100 consultas de prueba, únicamente 58 consultas cuentan con valores de búsqueda. Por lo tanto, el objetivo de este experimento es comprobar que en esas 58 consultas se asignan de forma correcta las siguientes etiquetas: texto en general (*word*), nombres propios (*name*), números decimales y enteros (*number*), así como fechas (*date*) y horas (*hour*).

5.3.2 Pruebas del Preprocesamiento

La Tabla 5.2 muestra un fragmento del corpus de consultas con los valores de búsqueda identificados en cada una de ellas, los cuales se encuentran resaltados en negrita en la consulta. Todas las demás palabras fueron marcadas con la etiqueta *word*, indicando con esto que serán buscadas en el lexicón para extraer su(s) categoría(s) gramatical(es). Con el objetivo de facilitar la visualización, se omitió la etiqueta *word*, y sólo se dejaron los valores de búsqueda. Para mayor detalle de los resultados de las pruebas del preprocesamiento léxico, consultar el Apéndice C.

Tabla 5.2. Pruebas del preprocesamiento léxico

No.	Consulta y valores de búsqueda.
1	How many people live in **Hawaii**? **name**
2	What is the shortest river in the **USA**? **code**
3	Does **Delta Airlines** flight **179** serve breakfast? **code** **code** **number**
4	Give me the flights of the airline **Tokened Airlines** that serve breakfast. **name** **name**
5	Which airlines have economy flights with discount from **San Francisco** to **Dallas**? **name**, **name** **name**

Como resultado del **Experimento 1**, la Tabla 5.3 muestra un resumen de los resultados de las pruebas realizadas. En la Tabla 5.2 se puede observar que el preprocesamiento realiza la correcta identificación y etiquetado de los valores de búsqueda en las 5 consultas de muestra, y lo mismo

ocurre con las 58 consultas que los contienen. En base a los resultados obtenidos del **Experimento 1**, se concluye que el preprocesamiento funciona correctamente.

Tabla 5.3. Resumen de los resultados de las pruebas del preprocesamiento

Número total de consultas.	100
Consultas sin valores de búsqueda.	42
Consultas con valores de búsqueda.	58
Consultas en que el preprocesamiento identificó los valores de búsqueda.	58

5.3.3 Etiquetado Léxico

Experimento 2: se busca comprobar el funcionamiento de la implementación del Algoritmo 2.1, y que el etiquetado léxico extrae todas las categorías gramaticales de las palabras que conforman las consultas del corpus de pruebas de la BD del lexicón. El objetivo del experimento es comprobar que todas las consultas se encuentren etiquetadas y que cada palabra tenga asignadas todas las categorías que le corresponden.

5.3.4 Pruebas del Etiquetado Léxico

En la Tabla 5.4 se muestra un fragmento del etiquetado del corpus de pruebas. Además, se presentan todas las categorías gramaticales extraídas del lexicón para cada una de las palabras que forman las consultas. Es importante señalar que en esta parte del analizador léxico, algunas palabras pueden incluir más de una categoría gramatical asignada. Para mayor detalle de los resultados de las pruebas del preprocesamiento léxico, consultar el Apéndice D.

Tabla 5.4. Pruebas del etiquetado léxico

No.	Consulta y categorías gramaticales.
1	**How many people live in Hawaii?** adverb | indefinite adjective | noun | verb | adverb | name | adverb | indefinite pronoun | | preposition |
2	**What is the shortest river in the USA?** interrogative adjective | verb | article | superlative adjective | noun | adverb | article | code | interrogative pronoun | | preposition |
3	**Does Delta Airlines flight 179 serve breakfast?** auxiliary verb | code | code | noun | num-ent | verb | noun |
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | conjunction | verb | noun | | determiner | | pronoun |
5	**Which airlines have economy flights with discount from San Francisco to Dallas?** interrogative adjective | noun | verb | qualifying adjective | noun | preposition | noun | preposition | name | name | preposition | name interrogative pronoun |

Como resultado del **Experimento 2**, como se muestra en la Tabla 5.4, el etiquetado léxico funciona correctamente, ya que extrajo de la BD del lexicón de inglés todas las categorías gramaticales de cada una de las palabras que conforman las consultas del corpus de pruebas.

5.3.5 Postprocesamiento

Experimento 3: se busca verificar el funcionamiento de la implementación del Algoritmo 2.2. Es decir, comprobar que las reglas definidas para la heurística, definidas en la Subsección 4.3.3, permiten la asignación de puntuaciones a las categorías gramaticales de palabras que presentan ambigüedad léxica. El objetivo del experimento es comprobar la asignación de puntuaciones para resolver la ambigüedad léxica existente en algunas palabras.

5.3.6 Pruebas del Postprocesamiento

En la Tabla 5.5 se muestra un fragmento de las puntuaciones asignadas a las categorías gramaticales de las palabras que presentan ambigüedad léxica de algunas consultas del corpus de pruebas. Se asignó aquella categoría con el mayor puntaje (resaltada en negrita) como la categoría gramatical. Para mayor detalle de los resultados de las pruebas del preprocesamiento léxico, consultar el Apéndice E.

Tabla 5.5. Pruebas del postprocesamiento léxico

No.	Consulta y categorías gramaticales.
1	**How many people live in Hawaii?** adverb | indefinite adjective 0 | noun | verb | adverb 2 | name | | **indefinite pronoun 1** | | **preposition 3** |
2	**What is the shortest river in the USA?** interrogative adjective 0 | verb | article | superlative adjective | noun | adverb 1 | article | code | **interrogative pronoun 3** | | **preposition 2** |
3	**Does Delta Airlines flight 179 serves breakfast?** *No contiene palabras con múltiples categorías gramaticales.*
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | **conjunction 3** | verb | noun | | determiner 0 | | pronoun 1 |
5	**Which airlines have economy flights with discount from San Francisco to Dallas?** **interrogative adjective 3** | noun | verb | qualifying adjective | noun | preposition | noun | preposition | name | name | preposition | name interrogative pronoun 0 |

De las 100 consultas evaluadas, 92 presentaban ambigüedad léxica en algunas de sus palabras, es decir, contaban con múltiples categorías gramaticales asignadas. Como resultado del **Experimento 3**, como se muestra en la Tabla 5.5, las reglas para inglés diseñadas para la heurística permiten la evaluación y asignación de diferentes puntuaciones a las múltiples categorías de las palabras que presentan ambigüedad léxica.

Se comprobó que el postprocesamiento léxico funciona correctamente al resolver la ambigüedad léxica de las 92 consultas con este tipo de problema. Específicamente, asigna como la categoría correcta aquélla con la mayor puntuación, ya que es la que tiene la mayor probabilidad de ser la correcta. La Tabla 5.6 muestra un resumen de estos resultados.

Tabla 5.6. Resumen de los resultados de las pruebas del postprocesamiento

Número total de consultas.	100
Consultas con ambigüedad léxica.	92

Consultas con ambigüedad léxica resuelta correctamente.	92

En base a los resultados obtenidos del Experimento 3 y los datos mostrados en la Tabla 5.6, se puede concluir que el postprocesamiento léxico funciona para lo que fue diseñado al conseguir que todas las palabras de las consultas tengan una sola categoría gramatical asignada. Por lo tanto, se concluye que el postprocesamiento funciona correctamente.

5.3.7 Pruebas de Funcionalidad del Analizador Léxico

Las principales características del analizador léxico son las siguientes: (a) permite la eliminación de todos los caracteres de las oraciones considerados como no relevantes (signos de puntuación), (b) permite la detección temprana de los valores de búsqueda, lo cual reduce el tiempo de procesamiento, y (c) permite dar solución a la ambigüedad léxica en las consultas que lo requieren, es decir, asignar una sola categoría gramatical por palabra de la consulta en LN.

Experimento 4: se busca realizar una evaluación completa del analizador léxico con el fin de comprobar si el etiquetado es el correcto. Se registraron los tiempos de procesamiento para cada una de las consultas del corpus de pruebas con el fin de evaluar el tiempo de procesamiento del analizador léxico. Cada consulta se procesó 20 veces, por lo que los tiempos que se muestran en la tabla son el tiempo promedio de las repeticiones. Se calcularon tiempos promedio, porque existen variaciones en el tiempo de ejecución en cada corrida.

En la Tabla 5.7 se muestra un fragmento de los resultados obtenidos del Experimento 4, donde se presentan las categorías asignadas a cada palabra del corpus de pruebas, así como el tiempo que le tomó al analizador léxico procesar cada consulta del corpus de pruebas. Para mayor detalle de los resultados de las pruebas del preprocesamiento léxico, consultar el Apéndice F.

Tabla 5.7. Pruebas de funcionalidad del analizador léxico

No.	Consulta y categorías gramaticales.	Tiempo
1	**How many people live in Hawaii?** adverb | indefinite adjective | noun | verb | preposition | name |	0.148
2	**What is the shortest river in the USA?** interrogative pronoun | verb | article | superlative adjective | noun | preposition | article | code |	0.138
3	**Does Delta Airlines flight 179 serve breakfast?** auxiliary verb | code | code | noun | num-ent | verb | noun |	0.082
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | conjunction | verb | noun |	0.152
5	**Which airlines have economy flights with discount from San Francisco to Dallas?** interrogative adjective | noun | verb | qualifying adjective | noun | preposition | noun | preposition | name | name | preposition | name |	0.147

Como resultado del **Experimento 4**, como se muestra en la Tabla 5.7 y en base a todas las pruebas mostradas en la Sección 5.3, se concluye que el **analizador léxico** cuenta con la capacidad de eliminar los signos de puntuación, identificar los valores de búsqueda, y etiquetar correctamente todas las palabras de las consultas que se le introdujeron.

De las 100 consultas evaluadas, el analizador léxico logró procesar correctamente 100 consultas. Por lo tanto, se concluye que el analizador léxico cumple las funciones para las cuales

fue diseñado, funcionando de forma correcta, debido a que cuenta con un desempeño aceptable. **Al analizador léxico le tomó en promedio 162 milisegundos procesar cada una de las 100 consultas del corpus de pruebas.**

La Tabla 5.8 muestra un resumen de los resultados de los experimentos realizados al analizador léxico.

Tabla 5.8. Resumen de los resultados generales del analizador léxico

Número total de consultas.	100
Consultas sin valores de búsqueda.	42
Consultas con valores de búsqueda.	58
Consultas en que el preprocesamiento identificó los valores de búsqueda.	58
Consultas con ambigüedad léxica.	92
Consultas con ambigüedad léxica resuelta correctamente.	92
Consultas etiquetadas correctamente con una sola categoría por palabra.	100
Tiempo promedio de procesamiento del analizador léxico.	0.162

5.4 Resultados de Pruebas del AS-S

Para la evaluación del funcionamiento de los procesos del analizador sintáctico-semántico de inglés, se diseñaron diferentes experimentos que cubren todos los aspectos que se deseaban evaluar. En cada subsección se definen los experimentos realizados en esa etapa.

5.4.1 Pruebas de Funcionalidad del AS-S

Experimento 5: se busca evaluar el funcionamiento de la implementación del Algoritmo 2.3 y los métodos de las reglas de producción para inglés, definidas en la Subsección 4.4.3, para la generación de reducciones. El objetivo del experimento es comprobar que los métodos de las reglas de producción realizan reducciones que llegan al símbolo *Sentence*, así como la detección cuando una oración es gramaticalmente incorrecta.

La Tabla 5.9 muestra un fragmento del corpus de pruebas con las reducciones generadas por los métodos de las reglas de producción. La primera columna muestra el identificador de la consulta, y la segunda columna muestra todas las reducciones realizadas a las 100 consultas del corpus de prueba. Dentro de la segunda columna del lado derecho, se encuentra el identificador de la regla aplicada, por ejemplo, ***SNO1-V***, seguida de la secuencia de símbolos reducidos que recibe *inputExpression* en cada recorrido. Las reducciones se encuentran resaltadas en negrita.

Por último, en la tercera columna se encuentra el tiempo promedio que le tomó al AS-S procesar cada una de las 100 consultas del corpus de pruebas. Para mayor detalle de los resultados de las pruebas de funcionalidad del AS-S, consultar el Apéndice G.

Tabla 5.9. Reducciones aplicando reglas de producción

No.	Consulta y categorías gramaticales.	Time
1	**How many people live in Hawaii?** EXPRESION ORIGINAL [adv, adj_ind, nou, ver, pre, name] SN01-V [adv, adj_ind, nou, ver, pre, **NouP0V**] SN02 [adv, adj_ind, **NouP0**, ver, pre, NouP0V] SA04 [**SAdj1**, NouP0, ver, pre, NouP0V] SN09 [**NouP1**, ver, pre, NouP0V] SN10-V [NouP1, ver, pre, **NouP1V**] SP01-V [NouP1, ver, **PrePV]** SV03 [NouP1, **VerP1**, PrePV] TS01 [**Suj**, VerP1, PrePV] TC01 [Suj, VerP1, **Com**] TO01 **[S E N T E N C E**]	**0.004**
2	**What is the shortest river in the USA?** EXPRESION ORIGINAL [pro_int, ver, art, adj_sup, nou, pre, art, code] SN01-V [pro_int, ver, art, adj_sup, nou, pre, art, **NouP0V**] SN02-V [pro_int, ver, art, adj_sup, nou, pre, **NouP1V**] SN02 [pro_int, ver, art, **NouP0**, pre, NouP1V] SN03 [pro_int, ver, **NouP1**, pre, NouP1V] SV01 [pro_int, **VerP1**, pre, NouP1V] SP01-V [pro_int, VerP1, **PrePV**] SV02 [**VerP1**, PrePV] TC01 [VerP1, **Com**] TO02 [**S E N T E N C E**]	**0.002**
3	**Give me the flights of the airline Tokened Airlines that serve breakfast.** EXPRESION ORIGINAL [ver, pro, art, nou, pre, art, nou, name, that, ver, nou] SN01-V [ver, pro, art, nou, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, nou, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, SNom0, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, SNom0, SNom0V, that, ver, SNom0] SN03 [ver, SNom0, SNom1, pre, art, SNom0, SNom0V, that, ver, SNom0] SN03 [ver, SNom0, SNom1, pre, SNom1, SNom0V, that, ver, SNom0] SV01 [SVer1, SNom1, pre, SNom1, SNom0V, that, ver, SNom0] SV01 [SVer1, SNom1, pre, SNom1, SNom0V, that, SVer1] SN10-V [SVer1, SNom1, pre, SNom1V, that, SVer1] SP01-V [SVer1, SNom1, SPreV, that, SVer1] SP02-V [SVer1, SNom1, SPreV] TC01 [SVer1, SNom1, Com] TC02 [SVer1, Com, Com] TO02_1 [S E N T E N C E]	**0.039**

Con los resultados obtenidos del Experimento 5, se comprobó que el AS-S de inglés aplica las reglas de producción para generar agrupaciones y reducciones en las 100 consultas del corpus de pruebas. Con estas pruebas se demuestra que el AS-S de inglés funciona correctamente, ya que, además de generar reducciones, logró identificar 99 consultas correctas y una incorrecta, específicamente, la consulta 59, la cual carece de un verbo principal dentro de su estructura. **Al AS-S le tomó en promedio 21 milisegundos procesar cada una de las 100 consulta del corpus de pruebas.**

5.5 Pruebas Comparativas

Una vez concluida la implementación de los algoritmos y las pruebas de funcionalidad de los módulos del analizador léxico y del AS-S, se llevó a cabo una comparación de ambos analizadores

versus Wolfram y CLAWS 7, utilizando el mismo corpus de pruebas con el que se ha estado trabajando a lo largo de este proyecto.

5.5.1 Pruebas Comparativas del Analizador Léxico

Experimento 6: se busca comparar el proceso de etiquetado del analizador léxico de inglés implementado en este proyecto contra el analizador de Wolfram (específicamente el etiquetador) y con el etiquetador CLAWS C7. El objetivo del experimento es comprobar el desempeño del analizador léxico implementado en este proyecto comparado contra dos analizadores comerciales.

5.5.2 Pruebas de Funcionalidad del AS-S

La Tabla 5.10 muestra un fragmento del corpus de pruebas con el etiquetado de las consultas realizado por CLAWS (CLS), Wolfram (WLF) y el Analizador Léxico de inglés (AL_Eng). La nomenclatura de CLAWS y Wolfram se adaptó a la gramática descrita en la Subsección 2.5.1. Para mayor detalle de los resultados de las pruebas comparativas del analizador léxico, consultar el Apéndice H.

Tabla 5.10. Pruebas comparativas del analizador léxico

1	**How many people live in Hawaii?** CLS: adv | det | nou | ver | pre | PN WLF: adv | adj | nou | ver | pre | PN ALEng: adv | adj_ind | nou | ver | pre | name
2	**What is the shortest river in the USA?** CLS: Q_det | ver | art | adj_sup | nou | pre | art | PN WLF: pro_int | ver | det | adj | nou | pre | det | PN ALEng: pro_int | ver | art | adj_sup | nou | pre | art | code |
3	**Does Delta Airlines flight 179 serves breakfast?** CLS: ver | nou | nou | nou | num | ver | nou WLF: PN | PN | PN | ver | num | num | nou ALEng: aux_ver | code | code | nou | num-ent | ver | nou
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** CLS: ver | pro | art | nou | pre | art | nou | adj | nou | con | ver | nou WLF: ver | pro | det | nou | pre | det | nou | ver | nou | Q_det | ver | nou ALEng: ver | pro | art | nou | pre | art | nou | name | name | con | ver | nou
5	**Which airlines have economy flights with discount from San Francisco to Dallas?** CLS: Q_det | nou | ver | nou | nou | pre | nou | pre | PN | PN | pre | PN WLF: Q_det | nou | ver | nou | nou | pre | nou | pre | PN | PN | pre | PN | ALEng: adj_int | nou | ver | adj_cal | nou | pre | nou | pre | name | name | pre | name |

Como resultado del Experimento 6, mostrado en la Tabla 5.10, se puede observar que el analizador léxico logra diferenciar los diferentes tipos de determinantes (artículos, adjetivos demostrativos, adjetivos posesivos y adjetivos indefinidos) del inglés, a diferencia de Wolfram y CLAWS 7. Ambos realizan el etiquetado de éstos como determinantes sin diferenciar entre ellos.

CLAWS identifica la palabra *list* (lista) al inicio de la consulta, en algunas ocasiones, como sustantivo en lugar de verbo (forma imperativa). Esto se puede observar en la consulta "*List*

afternoon flights from Atlanta to San Fransisco" (Lista los vuelos de la tarde de Atlanta a San Francisco). En esta consulta, tanto Wolfram como el analizador léxico realizan la identificación de la palabra *list* de forma correcta. El analizador léxico por su parte logra la identificación de los verbos que se encuentran en la primera posición, tomando en cuenta que las consultas pueden ser oraciones imperativas.

El analizador léxico logra la identificación de los valores de búsqueda de forma eficiente, a diferencia de CLAWS y Wolfram que no fueron diseñados para su identificación. Por ejemplo, en la consulta *What flights from Dallas to San Francisco or Okland use restrictions AP/80?*, CLAWS no logra identificar la clave *AP/80*, por lo que la etiqueta como *uclassified word* (palabra sin clasificación). Wolfram por su parte la identifica como *proper noun* (nombre propio).

Se detectaron dos errores principales en el etiquetado de Wolfram. En el primero de ellos, los verbos auxiliares que se encuentran al inicio de las consultas son identificados como nombres propios, y en el segundo cuando se tiene un número seguido de un verbo. Este último no se clasifica de forma correcta, ya que generalmente se le asignó la categoría de número (ver la consulta 3 de la Tabla 5.10). La Tabla 5.11 resume los resultados de las pruebas comparativas.

Tabla 5.11. Resumen de resultados de las pruebas comparativas del AL de inglés

Número total de consultas.	100
Consultas correctamente identificadas por CLAWS.	97
Consultas correctamente identificadas por Wolfram.	92
Consultas correctamente identificadas por AL de inglés.	99
Palabras sin clasificación por CLAWS.	2
Palabras sin clasificación por Wolfram.	0
Palabras sin clasificación por AL de inglés.	0
Palabras clasificadas incorrectamente por CLAWS.	1
Palabras clasificadas incorrectamente por Wolfram.	10
Palabras clasificadas incorrectamente por AL de inglés.	1
Tiempo promedio de procesamiento de una consulta del AL de inglés.	161 mls

En base a los resultados del analizador léxico de inglés, se detectó que sólo presentó problemas al identificar correctamente una palabra. Esto es debido a que se encuentra en la primera posición y tanto ella como la palabra que la precede presentan ambigüedad léxica, por lo cual no logra procesarla de forma correcta.

El analizador léxico realizó el etiquetado de todas las palabras de las consultas con las categorías gramaticales correspondientes. Le tomó en promedio procesar cada una de las 100 consulta del corpus de pruebas 161 milisegundos.

5.5.3 Pruebas Comparativas del AS-S

Con relación a las pruebas comparativas del AS-S de inglés y la función *TextStructure* de Wolfram, los resultados de Wolfram se expresaron utilizando la gramática definida para el AS-S de inglés, ya que, como se muestra en la Tabla 2.6 de la Sección 2.8, la estructura utilizada por Wolfram

difiere de la utilizada en este proyecto. Adicionalmente, algunos símbolos difirieren de los definidos en la gramática del AS-S. La Tabla 5.12 muestra estos símbolos.

Tabla 5.12. Símbolos genéricos para Wolfram

Categoría gramatical	Símbolos genéricos
determiner	det
proper noun	PN
possessive modifier	pos_mod
interrogative determiner	Q_det
interrogative noun phrase	QNouP
interrogative adverb	QAdv
quantifier phrase	QuanP
interrogative adjective phrase	QAdjP

Experimento 7: se busca comparar el proceso de reducciones de las consultas del corpus de pruebas del AS-S versus la función *TextStructure* de Wolfram que también genera reducciones gramaticales. El objetivo del experimento es evaluar el desempeño del AS-S implementado en este proyecto comparado contra un analizador comercial.

La Tabla 5.13 muestra un fragmento del corpus de pruebas con las reducciones generadas por Wolfram (Wlf) y el Analizador Sintáctico-Semántico (AS-S). La tabla muestra las reducciones generadas por Wolfram expresadas utilizando la gramática definida en la Subsección 2.5.1. Esto es con el fin de tener consistencia en la expresión de los resultados, ya que la estructura utilizada por Wolfram es diferente de la de AS-S, como se muestra en la Figura 2.6. Para mayor detalle de los resultados de las pruebas de funcionalidad del AS-S, consultar el Apéndice I.

Tabla 5.13. Pruebas comparativas del AS-S de inglés

1	**How many people live in Hawaii?**
AS-S de inglés	**Wolfram**
[adv, adj_ind, nou, ver, pre, name] [adv, adj_ind, nou, ver, pre, **NouP0V**] [adv, adj_ind, **NouP0**, ver, pre, NouP0V] [**SAdj1**, NouP0, ver, pre, NouP0V] [**NouP1**, ver, pre, NouP0V] [NouP1, ver, pre, **NouP1V**] [NouP1, ver, **PrePV**] [NouP1, **VerP1**, PrePV] [**Suj**, VerP1, PrePV] [Suj, VerP1, **Com**] [**S E N T E N C E**]	[adv, adj, nou, ver, pre, PN] [**QAdjP**, nou, ver, pre, PN] [QAdjP, nou, ver, pre, **NouP**] [QNouP, ver, pre, **NouP**] [QNouP, ver, **PreP**] [**C L A U S E**]
2	**What is the lowest point of the state with the largest area?**
AS-S de inglés	**Wolfram**
[pro_int, ver, art, adj_sup, nou, pre, art, nou, pre, art, adj_sup, nou] [pro_int, ver, art, **NouP0**, pre, art, nou, pre, art, adj_sup, nou] [pro_int, ver, art, NouP0, pre, art, **NouP0**, pre, art, adj_sup, nou] [pro_int, ver, art, NouP0, pre, art, NouP0, pre, art, **NouP0**] [pro_int, ver, **NouP1**, pre, art, NouP0, pre, art, NouP0] [pro_int, ver, NouP1, pre, **NouP1**, pre, art, NouP0] [pro_int, ver, NouP1, pre, NouP1, pre, **NouP1**] [pro_int, **VerP1**, pre, NouP1, pre, NouP1]	[pro_int, ver, det, adj, nou, pre, det, nou, pre, det, adj, nou] [**QNouP**, ver, det, adj, nou, pre, det, nou, pre, det, adj, nou] [QNouP, ver, **NouP**, pre, det, nou, pre, det, adj, nou] [QNouP, ver, NouP, pre, **NouP**, pre, det, adj, nou] [QNouP, ver, NouP, pre, NouP, pre, **NouP**] [QNouP, ver, NouP, pre, NouP, **PreP**] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**]

[pro_int, VerP1, **PreP**, pre, NouP1] [pro_int, VerP1, PreP, **PreP**] [**VerP1**, PreP, PreP] [VerP1, **Com**, PreP] [VerP1, Com, Com] [**S E N T E N C E**]	[QNouP, ver, **NouP**] [QNouP, **VerP**] [**C L A U S E**]
3	**Give me all flights from Atlanta to San Francisco on Delta first class.**
AS-S de inglés	**Wolfram**
[ver, pro, ind_pro, nou, pre, name, pre, name, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, **NouP0V**, pre, name, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, NouP0V, pre, **NouP0V**, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, **NouP0V**, adj_cal, nou] [ver, **NouP0**, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] [ver, NouP0, ind_pro, **NouP0**, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] [ver, NouP0, ind_pro, NouP0, pre, NouP0V, pre, NouP0V, pre, NouP0V, **NouP0**] [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] [**VerP1**, NouP1, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] [VerP1, NouP1, pre, **NouP1V**, pre, NouP0V, pre, NouP0V, NouP0] [VerP1, NouP1, pre, NouP1V, pre, **NouP1V**, pre, NouP0V, NouP0] [VerP1, NouP1, pre, NouP1V, pre, NouP1V, pre, **NouP1V**, NouP0] [VerP1, NouP1, **PrePV**, pre, NouP1V, pre, NouP1V, NouP0] [VerP1, NouP1, PrePV, **PrePV**, pre, NouP1V, NouP0] [VerP1, NouP1, PrePV, PrePV, **PrePV**, NouP0] [VerP1, NouP1, **Com**, PrePV, PrePV, NouP0] [VerP1, NouP1, Com, **Com**, PrePV, NouP0] [VerP1, NouP1, Com, Com, **Com**, NouP0] [VerP1, **Com**, Com, Com, Com, NouP0] [VerP1, Com, Com, Com, Com, **Com**] [**S E N TE N C E**]	[ver, pro, det, nou, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, **NouP**, det, nou, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, **NouP**, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, NouP, pre, **NouP**, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, NouP, pre, NouP, pre, **NouP**, pre, PN, adj, nou] [ver, NouP, NouP, pre, NouP, pre, NouP, pre, **NouP**] [ver, NouP, NouP, pre, NouP, pre, NouP, **PreP**] [ver, NouP, NouP, pre, NouP, pre, **NouP**] [ver, NouP, NouP, pre, NouP, **PreP**] [ver, NouP, NouP, pre, **NouP**] [ver, NouP, NouP, **PreP**] [ver, NouP, **NouP**] [**VerP**] [**S E N TE N C E**]

CAPÍTULO 6

Conclusiones y Trabajos Futuros

Las interfaces de LN a base de datos (ILNBDs) son herramientas que permiten a usuarios no especializados tener acceso a información recopilada en BDs a través de consultas formuladas en su LN nativo. Las ILNBDs realizan la interpretación de las consultas formuladas por los usuarios en LN por medio de un analizador que procesa la consulta. Para el procesamiento de LN, existen diferentes enfoques de analizadores para ILNBDs, sin embargo, uno de los enfoques más prometedores es un analizador sintáctico que integre semántica, mejorando de esta manera el desempeño.

En [Verástegui, 2020] se presenta un analizador sintáctico-semántico (AS-S) que integra información sintáctico-semántica del idioma español en sus reglas de producción, permitiéndole identificar las relaciones estructurales internas de las consultas en LN, así como resolver varios de los problemas presentes dentro del procesamiento de éstas. El enfoque del AS-S descrito en [Verástegui, 2020] puede ser aplicado a otros leguajes europeos como el inglés, italiano, francés y portugués.

En este capítulo se presentan las conclusiones de este proyecto cuyo principal objetivo fue comprobar que los métodos definidos en [Verástegui, 2020] para un AS-S de español pueden ser implementados para trabajar con consultas en LN en inglés. Esto permite ampliar la funcionalidad de la ILNBD desarrollada en el Instituto Tecnológico de Cd. Madero (ITCM) al lograr que trabaje con consultas en LN en español e inglés. En ese proyecto se realizaron cuatro contribuciones principales, las cuales se desglosan en el Capítulo 4.

La **primera contribución** es la definición de un lexicón de inglés con suficiente información que permite la implementación del analizador para la interfaz. La **segunda contribución** es la adaptación de los procesos y la implementación de un analizador léxico de inglés. La **tercera contribución** es la definición de las reglas de producción del AS-S basadas en las estructuras gramaticales de la lengua inglesa, permitiéndole trabajar con consultas en LN en inglés. Por último, la **cuarta contribución** fue la implementación de un AS-S para el idioma inglés.

Para finalizar, en la Sección 6.2 se proponen algunos temas de investigación para trabajos futuros para mejorar el desempeño y resolver problemas que no se abordaron en este proyecto.

6.1 Conclusiones

6.1.1 Lexicón de Inglés

Como parte de este proyecto, se definió una primera versión del lexicón de inglés por medio del Algoritmo 4.1 descrito en el Capítulo 4. Éste utiliza como base un corpus de palabras de inglés existente que cuenta con una gran cantidad de lexemas, así como con reconocimiento dentro del estado del arte. El corpus COCA fue seleccionado para este fin.

Se realizó una verificación superficial para confirmar que las etiquetas estuvieran asignadas de forma correcta utilizando el diccionario [Macmillan Dictionary, 2021]. Se considera importante mencionar que es necesaria una verificación más profunda de las etiquetas asignadas a las palabras del lexicón, ya que se detectó que **algunos de los etiquetados asignados por CLAWS a las palabras no corresponden a la categoría correcta.**

El lexicón de inglés cuenta con 49,414 palabras que se dividen en 11 categorías gramaticales diferentes. Las palabras almacenadas en la base de datos pertenecen al inglés americano. Se considera que el lexicón cumple la función para la que fue diseñado, ya que cuenta y proporciona la información necesaria para la realización de los procesos, tanto del analizador léxico como del analizador sintáctico-semántico. Por lo tanto, **el objetivo específico OE1 de este proyecto se considera alcanzado.**

6.1.2 Analizador Léxico

Durante el desarrollo del **objetivo específico OE2**, se observó que el analizador léxico, en las pruebas efectuadas para entender el funcionamiento del preprocesamiento léxico, realizaba **la correcta identificación de los valores de búsqueda, sin importar si la consulta en LN se encontraba escrita en español o en inglés**, exceptuando un grupo de palabras en inglés que eran etiquetadas erróneamente como valores de búsqueda, específicamente: *may* (como verbo "poder"), *I* (yo) y adjetivos compuestos como *amphibian-type* (tipo anfibio).

Para solucionar este problema, se efectuaron las modificaciones necesarias en el preprocesamiento para su correcta identificación. En las pruebas de funcionalidad realizadas al preprocesamiento léxico (Tabla 5.3), descritas en el Capítulo 5, se observa la correcta identificación de los valores de búsqueda en el corpus de pruebas, concluyendo que el preprocesamiento léxico puede ser aplicado de forma correcta tanto en consultas en LN en español como en inglés.

El Algoritmo 2.1 descrito en el Capítulo 2 permite realizar el etiquetado léxico de consultas en inglés. La evaluación experimental (Tabla 5.4) permite comprobar que el Algoritmo 2.1, diseñado para la versión para español del analizador léxico, efectúa la extracción del lexicón de inglés de todas las categorías gramaticales de las palabras que conforman las consultas del corpus de pruebas en inglés, incluidas aquellas palabras con múltiples categorías gramaticales.

Para resolver la ambigüedad léxica (múltiples categorías gramaticales por palabra), al igual que la versión para español, se implementó el Algoritmo 2.2. Éste corresponde a un postprocesamiento léxico que permite por medio de una heurística asignar una sola categoría gramatical por palabra. Es importante mencionar que resolver la ambigüedad léxica es un problema muy complejo y no forma parte de los objetivos de este proyecto. Por lo tanto, al igual que la versión para español, sólo se definió e implementó una heurística basada en las estructuras descritas en [Verástegui, 2020] adaptada a las particularidades del inglés y cuyo desempeño es aceptable.

En las pruebas de funcionalidad realizadas al postprocesamiento léxico (Tabla 5.5) descritas en el Capítulo 5, se observó que los métodos desarrollados para la versión para español de éste pueden ser adaptados al idioma inglés. Esto permite resolver la ambigüedad léxica existente en 92 consultas del corpus de pruebas al seleccionar la categoría con la mayor probabilidad de ser la correcta por medio de la asignación de puntuaciones, seleccionando la de mayor puntuación. Al concluir el postprocesamiento, cada consulta se encuentra etiquetada con una sola categoría gramatical por palabra.

6.1.3 Pruebas Comparativas del Analizador Léxico

Se realizaron pruebas comparativas (Tabla 5.11) del analizador léxico de inglés con el etiquetador CLAWS C7 y la función de análisis de estructuras de Wolfram, específicamente la sección de etiquetado de la consulta. Las pruebas mostraron que tanto Wolfram como CLAWS C7 identifican artículos, adjetivos demostrativos, posesivos e indefinidos como determinantes, mientras que el analizador léxico hace la distinción entre ellos.

El analizador léxico, a diferencia de CLAWS C7 y Wolfram, fue diseñado para la identificación de valores de búsqueda. Por ejemplo, en la consulta *What flights from Dallas to San Francisco or Okland use restrictions AP/80?*, el analizador léxico logró identificar como valores de búsqueda Dallas, San Francisco, Oakland y AP/80, específicamente, nombres (*name)* y clave (*code*). Mientras que CLAWS los identifica como nombres propios (*proper noun*) y a *AP/80* como *uclassified word* (palabra sin clasificación). Wolfram por su parte los identifica a todos como *proper nouns* (nombres propios).

Con todas las pruebas efectuadas a los diferentes componentes del analizador léxico, se concluye que realiza las funciones para las que fue diseñado y que los métodos diseñados para el idioma español pueden ser aplicados al idioma inglés. Además, **el analizador léxico posee una eficiencia notable, ya que procesa en promedio cada una de las 100 consultas del corpus de pruebas en 162 milisegundos.**

6.1.4 Analizador Sintáctico-Semántico

El **objetivo específico OE3 se considera cumplido** y **la hipótesis H2 se considera probada** al definir 39 reglas de producción (Tabla 4.3) para el AS-S de inglés, utilizando las reglas estructurales de la gramática inglesa (Figura 4.11).

De las 39 reglas, 7 son reglas terminales que permiten la generación de oraciones para identificar si una oración es gramaticalmente correcta o no. En este proyecto se agregó al AS-S de inglés una función dentro de las reglas de producción terminales (TO01, TO02 y TO03), la cual le permite identificar cuando una oración es gramaticalmente incorrecta al no contar con un verbo principal. En caso de la detección de una oración con esta característica, el AS-S etiqueta la consulta como "The query is grammatically incorrect. It doesn't have a main verb*"* (la oración es gramaticalmente incorrecta, ya que no cuenta con un verbo principal).

La aplicación de las reglas de producción se logró por medio de la implementación del Algoritmo 2.3 descrito en el Capítulo 2. Cada regla de producción está implementada como un método, como se muestra en el Algoritmo 4.1, permitiendo cambiar el orden de aplicación sin alterar el código fuente de las reglas. Al igual que la versión para español, el AS-S fue diseñado para expresiones en LN para consulta a bases de datos, por lo tanto, sólo funciona con oraciones interrogativas e imperativas.

En las pruebas de funcionalidad al AS-S (Tabla 5.9) descritas en el Capítulo 5, se observó que las reglas de producción definidas permiten hacer agrupaciones de componentes léxicos, generando reducciones para el análisis de las consultas en inglés.

De las 100 consultas del corpus de pruebas, el AS-S logró evaluar las 100 consultas, identificando 99 consultas gramaticalmente correctas y una gramaticalmente incorrecta (consulta 59), ya que no cuenta con un verbo principal dentro de su estructura.

6.1.5 Pruebas Comparativas del AS-S

Se realizaron pruebas comparativas (Tabla 5.13) del analizador sintáctico semántico de inglés contra la función *TextStructure* de Wolfram. Las pruebas mostraron que tanto el AS-S como Wolfram identifican la consulta 59 como incorrecta, ya que carece de un verbo principal. Como muestra la Figura 6.1, el AS-S indica que la consulta es incorrecta, mientras que Wolfram ya no realiza más reducciones.

Premium class flights from ATL to PIT.

AS-S de inglés	**Wolfram**
[adj_cal, nou, pre, code, pre, code]	[PN, nou, nou, pre, PN, pre, PN]
[adj_cal, nou, pre, **NouP0V**, pre, code]	[**NouP**, pre, PN, pre, PN]
[adj_cal, nou, pre, NouP0V, pre, NouP0V]	[NouP, pre, **NouP**, pre, PN]
[NouP0, pre, NouP0V, pre, **NouP0V**]	[NouP, pre, NouP, pre, **NouP**]
[NouP0, pre, **NouP1V**, pre, NouP0V]	[NouP, pre, NouP, **PreP**]
[NouP0, pre, NouP1V, pre, **NouP1V**]	[NouP, pre, **NouP**]
[NouP0, **PrePV**, pre, NouP1V]	[NouP, **PreP**]
[NouP0, PrePV, **PrePV**]	[**NouP**]
[**Suj**, PrePV, PrePV]	
[Suj, Com, PrePV]	
[Suj, Com, **Com**]	
[**The query is grammatically incorrect. It doesn't have a main verb.**]	

Figura 6.1. Análisis de la consulta *Premium class flight from ATL to PIT*

Wolfram realiza agrupaciones excesivas y comete algunos errores al momento de identificar los tipos de oraciones, por ejemplo, en la consulta *Show me all the flights from Dallas to Denver with breakfast*. Como muestra la Figura 6.2, en la línea 1 Wolfram no identifica ningún verbo en la oración, sin embargo, en las reducciones en la línea 13 genera una frase verbal agrupando tres símbolos: dos de frase nominal y uno de frase preposicional, sin tener ningún verbo.

Show me all the flights from Dallas to Denver with breakfast.

AS-S de inglés

1 [ver, pro, ind_pro, art, nou, pre, name, pre, name, pre, nou]
2 [ver, pro, ind_pro, art, nou, pre, **NouP0V**, pre, name, pre, nou]
3 [ver, pro, ind_pro, art, nou, pre, NouP0V, pre, **NouP0V**, pre, nou]
4 [ver, **NouP0**, ind_pro, art, nou, pre, NouP0V, pre, NouP0V, pre, nou]
5 [ver, NouP0, ind_pro, art, **NouP0**, pre, NouP0V, pre, NouP0V, pre, nou]
6 [ver, NouP0, ind_pro, art, NouP0, pre, NouP0V, pre, NouP0V, pre,**NouP0**]
7 [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, pre, NouP0]
8 [**SVer1**, NouP1, pre, NouP0V, pre, NouP0V, pre, NouP0]
9 [SVer1, NouP1, pre, **NouP1V**, pre, NouP0V, pre, NouP0]
10 [SVer1, NouP1, pre, NouP1V, pre, **NouP1V**, pre, NouP0]
11 [SVer1, NouP1, pre, NouP1V, pre, NouP1V, **PreP**]
12 [SVer1, NouP1, **PrePV**, pre, NouP1V, PreP]
13 [SVer1, NouP1, PrePV, **PrePV**, PreP]
14 [SVer1, NouP1, **Com**, PrePV, PreP]
15 [SVer1, NouP1, Com, **Com**, PreP]
16 [SVer1, NouP1, Com, Com, **Com**]
17 [SVer1, **Com**, Com, Com, Com]
18 [**S E N T E N C E**]

Wolfram

1 [PN, pro, det, det, nou, pre, PN, pre, PN, pre, nou]
2 [**NouP**, pro, det, det, nou, pre, PN, pre, PN, pre, nou]
3 [NouP, **NouP**, det, det, nou, pre, PN, pre, PN, pre, nou]
4 [NouP, NouP, **NouP**, pre, PN, pre, PN, pre, nou]
5 [NouP, NouP, NouP, pre, **NouP**, pre, PN, pre, nou]
6 [NouP, NouP, NouP, pre, NouP, pre, **NouP**, pre, nou]
7 [NouP, NouP, NouP, pre, NouP, pre, NouP, pre, **NouP**]
8 [NouP, NouP, NouP, **PreP**, pre, NouP, pre, NouP]
9 [NouP, NouP, NouP, PreP, pre, NouP, **PreP**]
10 [NouP, NouP, **NouP**, pre, NouP, PreP]
11 [NouP, NouP, NouP, pre, **NouP**]
12 [NouP, NouP, NouP, **PreP**]
13 [NouP, **VerP**]
14 [**S E N T E N C E**]

Figura 6.2. Análisis de consulta *Show me all the flights from Dallas to Denver with breakfast*

Con las pruebas realizadas al analizador sintáctico-semántico, se concluye que los métodos diseñados e implementados para la versión para español pueden ser adaptados e implementados para el idioma inglés. Se considera que **el AS-S de inglés cumple las funciones para las que fue diseñado y cuenta con un desempeño notable al procesar en promedio cada una de las 100 consultas del corpus de pruebas en 21 milisegundos**. **Consecuentemente, el objetivo general OG y el objetivo específico OE4 de este proyecto se consideran alcanzados y completos.**

Con los resultados obtenidos a lo largo de los diferentes experimentos realizados en este proyecto de tesis, **se probó la hipótesis H1** al efectuar la implementación de los procesos desarrollados para un analizador de español, descritos en [Verástegui, 2020], para la creación de un analizador de inglés.

6.2 Trabajos Futuros

Considerando que en este trabajo sólo se abordó la implementación de un analizador léxico y un analizador sintáctico-semántico de inglés para una ILNBD, es necesario llevar a cabo trabajos que aborden los siguientes aspectos a fin de mejorar el desempeño del analizador y en general de la interfaz:

1. Realizar una verificación más profunda de las etiquetas asignadas a las palabras del lexicón, ya que se detectó que algunos de los etiquetados asignado por CLAWS a las palabras no corresponden a la categoría correcta.

2. Diseñar e implementar un nuevo lexicón que contenga más información sobre las particularidades del idioma inglés, específicamente los tipos de categorías gramaticales, por ejemplo: adjetivos descriptivos, pronombres interrogativos, adjetivos interrogativos, etc.

3. Continuar con la implementación del proceso de traducción de la consulta en LN en inglés a SQL. Esto es con el fin de determinar si los métodos desarrollados para la versión para español adaptados al idioma inglés permiten a la ILNBD contestar consultas en inglés.

6.3 Productos Académicos

Del presente trabajo de investigación se derivan los siguientes productos académicos:

- Se presentó el artículo "Analizador Sintáctico de Inglés para una Interfaz de Lenguaje Natural a Bases de Datos" en el Congreso Internacional de Investigación de Academia Journals: Desafíos y Perspectivas en un Mundo Cambiante, Tabasco, septiembre 2021.

  El artículo está incluido en las siguientes publicaciones: (1) en el portal de Internet academiajournals.com, con ISSN 1946-5351 en línea, vol. 13, no. 8, 2021 e indexación en la base de datos Fuente Académica Plus de EBSCOHOST, Massachusetts, Estados Unidos y (2) en el libro digital ebook titulado Diseminación de Resultados de Investigación Universitaria - Tabasco 2021, con ISBN 978-1-939982-68-1 en línea.

APÉNDICES

APÉNDICE A. Descripción de la Base de Datos ATIS

Air Travel Information System (ATIS) es una base de datos (BD) relacional que almacena información de *Official Airline Guide* (Guia Oficial de aerolíneas) [OAG, 1990]. Incluye información sobre vuelos, costos, aerolíneas, ciudades, aeropuertos y servicios terrestres. ATIS es usada como punto de referencia para comprender el lenguaje hablado y escrito.

Este apéndice presenta el esquema utilizado, el cual fue definido en [Verástegui, 2020]. Éste cuenta con 27 tablas y 125 columnas, como se muestra en la Figura A.1.

Para proporcionar un mayor entendimiento del contenido de la BD de ATIS, en las siguientes paginas se presentan las descripciones del contenido de cada tabla y de cada columna que contiene.

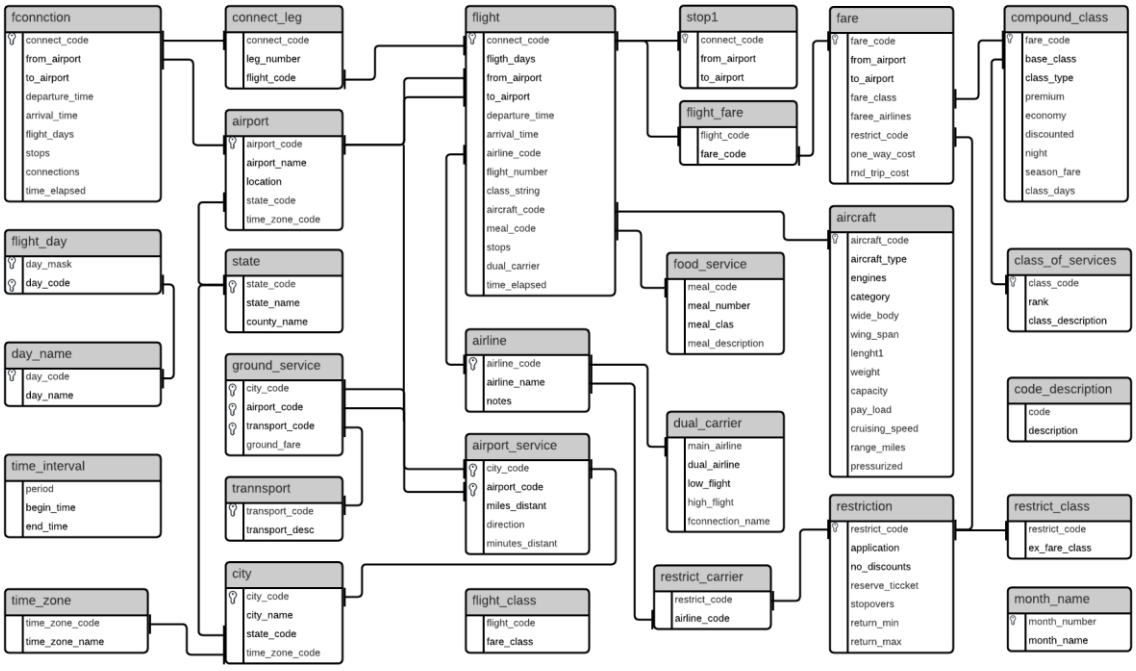


Figura A.1. Esquema de la base de datos ATIS

En la parte superior se muestra el nombre de cada tabla y su descripción. La primera columna contine el nombre de las columnas de cada tabla, la segunda columna muestra el tipo de

dato de cada columna, y la tercera contiene una breve descripción de las columnas. **Nota:** la BD ATIS cuenta con diferentes versiones con leves diferencias. La que se describe en este apéndice cuenta con dos tablas adicionales (*month_name* y *time_interval*) y algunas llaves primarias y llaves foráneas, las cuales pueden no estar definidas en otras versiones.

Tabla A.1. Descripción del esquema de la base de datos ATIS

Tabla: *aircraft*	**Descripción:** Aircraft (Aeronave)	
Columna	**Tipo**	**Descripción**
aircraft_code	Texto	código de avión
aircraft_type	Texto	tipo de avión
engines	Numérico	numero de motores
category	Texto	categoría
wide_body	Texto	de cuerpo ancho
wing_span	Numérico	extensión de alas
lenght1	Numérico	longitud
weight	Numérico	peso
capacity	Numérico	capacidad
pay_load	Numérico	carga
cruising_speed	Numérico	velocidad de crucero
range_miles	Numérico	alcance de vuelo
pressurized	Texto	presurización
LLAVE PRIMARIA:	aircraft_code	

Tabla: *airline*	**Descripción:** Airline (Aerolínea)	
Columna	**Tipo**	**Descripción**
airline_code	Texto	código de aerolínea
ariline_name	Texto	nombre de aerolínea
notes	Texto	notas
LLAVE PRIMARIA:	airline_code	
LLAVE FORANEA:	airline (airline_code) – restrict_carrier (airline_code)	

Tabla: *airport*	**Descripción:** Airport (Aeropuerto)	
Columna	**Tipo**	**Descripción**
airport_code	Texto	código de aeropuerto
airport_name	Texto	nombre de aeropuerto
location	Texto	ubicación
state_code	Texto	código de estado
time_zone_code	Texto	código de zona horaria
LLAVE PRIMARIA:	airport_code	
LLAVE FORANEA:	airport (state_code) – state (state_code)	
	airport (time_zone_code) – time_zone (time_zone_code)	

Tabla: *airport_service*	**Descripción:** AirportService (Servicio de Aeropuerto)	
Columna	**Tipo**	**Descripción**
city_code	Texto	código de ciudad

airport_code	Texto	código de aeropuerto
miles_distant	Numérico	distancia en millas
direction	Texto	dirección
minutes_distant	Numérico	distancia en minutos
LLAVE PRIMARIA:	city_code airport_code	
LLAVE FORANEA:	airport_service (airport_code) – airport (airport_code) airport_service (city_code) . city (city_code) airport_service (airport_code) – ground_service (airport_code) airport_service (city_code) – ground_service (city_code)	

Tabla: *city*	**Descripción:** City (Ciudad)	
Columna	**Tipo**	**Descripción**
city_code	Texto	código de ciudad
city_name	Texto	nombre de ciudad
state_code	Texto	código de estado
time_zone_code	Texto	código de zona horaria
LLAVE PRIMARIA:	city_code	
LLAVE FORANEA:	city (state_code) – state (state_code) city (time_zone_code) – time_zone (time_zone_code)	

Tabla: *class_of_service*	**Descripción:** Class of service (Clase de servicio)	
Columna	**Tipo**	**Descripción**
class_code	Texto	código de clase de servicio
rank	Numérico	rango
class_description	Texto	descripción de clase de servicio
LLAVE PRIMARIA:	class_code	

Tabla: *code_description*	**Descripción:** code description (descripcion de código)	
Columna	**Tipo**	**Descripción**
code	Texto	código
description	Texto	descripción
LLAVE PRIMARIA:	code	

Tabla: *compound_class*	**Descripción:** compound class (clase compuesta)	
Columna	**Tipo**	**Descripción**
fare_class	Texto	tarifa de clase
base_class	Texto	clase base
class_type	Texto	tipo de clase
premium	Texto	primera clase
economy	Texto	clase económica
discounted	Texto	descuento
night	Texto	vuelo nocturno
season_fare	Texto	tarida de temporada
class_days	Texto	días
LLAVE PRIMARIA:	fare_class	

LLAVE FORANEA:	compound_class (base_class) – class_of_service (code_class)

Tabla: *connect_leg*	**Descripción:** connect_leg (conexión)	
Columna	**Tipo**	**Descripción**
connect_code	Numérico	código de conexión
leg_number	Numérico	numero de conexión
flight_code	Numérico	código de vuelo
LLAVE PRIMARIA:	connect_code	
	connect_leg.flight_code	
LLAVE FORANEA:	connect_leg (connect_code) – fconnection (connect_code)	
	connect_leg (flight_code) – flight (flight_code)	

Tabla: *day_name*	**Descripción:** Days (Días)	
Columna	**Tipo**	**Descripción**
day_code	Numérico	código de día
day_name	Texto	nombre del día
LLAVE PRIMARIA:	day_code	

Tabla: *dual_carrier*	**Descripción:** dual carrier (compañía dual)	
Columna	**Tipo**	**Descripción**
main_airline	Texto	código de aerolínea principal
dual_carrier	Texto	código de compañía dual
low_flight	Numérico	vuelo inferior
high_flight	Numérico	vuelo superior
fconnection_name	Texto	nombre de conexión
LLAVE PRIMARIA:	main_airline	
	dual_airline	
	low_flight	
LLAVE FORANEA:	dual_carrier (dual_airline) -airline (airline_code)	
	dual_carrier (main_airline) -airline (airline_code)	

Tabla: *fare*	**Descripción:** fare (tarifa)	
Columna	**Tipo**	**Descripción**
fare_code	Texto	código de tarifa
from_airport	Texto	aeropuerto de origen
to_ariport	Texto	aeropuerto de destino
fare_class	Texto	clase de tarifa
fare_airline	Texto	código de aeropuerto
restrict_code	Texto	código de restricción
one_way_cost	Numérico	tarifa de vuelo sencillo
rnd_trip_cost	Numérico	tarifa de vuelo redondo
LLAVE PRIMARIA:	fare_code	
LLAVE FORANEA:	fare (restrict_code) – restruction (restrict_code)	
	fare (fare_class) – compound_class (fare_class)	

Tabla: *fconnection*	**Descripción:** Connection (Conexión)	
Columna	**Tipo**	**Descripción**
connect_code	Numérico	código de conexión
from_airport	Texto	aeropuerto de origen
to_ariport	Texto	aeropuerto de destino
departure_time	Numérico	hora de salida
arrival_time	Numérico	hora de llegada
flights_days	Texto	días de vuelo
stops	Numérico	escalas
connections	Numérico	conexiones
LLAVE PRIMARIA:	connect_code	
LLAVE FORANEA:	fconnection (to_ariport) – airport (airport_code)	
	fconnection (from_airport) – airport (airport_code)	

Tabla: *flight_class*	**Descripción:** Flight (Vuelo)	
Columna	**Tipo**	**Descripción**
flight_code	Numérico	código de vuelo
flights_days	Texto	días de vuelo
from_airport	Texto	aeropuerto de origen
to_ariport	Texto	aeropuerto de destino
departure_time	Numérico	hora de salida
arrival_time	Numérico	hora de llegada
airline_code	Texto	código de aerolínea
flight_number	Numérico	numero de vuelo
class_string	Texto	clase de código
aircraft_code	Texto	código de avión
meal_code	Texto	código de comida
stops	Numérico	escalas
dual_carrier	Texto	código de compañía dual
time_elapsed	Numérico	tiempo de vuelo
LLAVE PRIMARIA:	flight_code	
LLAVE FORANEA:	flight (aircraft_code) – aircraft (aircraft_code)	
	flight (airline_code) – airline (airline_code)	
	flight (from_airport) – airport (airport_code)	
	flight (to_ariport) – airport (airport_code)	
LINK INFORMAL:	flight (flights_days) – flight_day (day_mask)	

Tabla: *flight_day*	**Descripción:** Flight days (Días de Vuelo)	
Columna	**Tipo**	**Descripción**
day_mask	Texto	mascara de día
day_code	Numérico	código de día
LLAVE PRIMARIA:	day_mask	
	day_code	
LLAVE FORANEA:	flight_day (day_code) – day_name (day_code)	

Tabla: *flight_fare*	**Descripción:** Flight fare (Tarifa de vuelo)	
Columna	**Tipo**	**Descripción**
flight_code	Numérico	código de vuelo
fare_code	Texto	código de tarifa
LLAVE PRIMARIA:	flight_code fare_code	
LLAVE FORANEA:	flight_fare (fare_code) – fare (fare_code) flight_fare (flight_code) – flight (flight_code)	

Tabla: *food_service*	**Descripción:** Food service (Servicio de comida)	
Columna	**Tipo**	**Descripción**
meal_code	Texto	código de comida
meal_number	Numérico	numero de comida
meal_class	Texto	clase de comida
meal_description	Texto	descripción de comida
LLAVE PRIMARIA:	meal_code meal_number meal_class	
LLAVE FORANEA:	food_service (meal_code) – flight (meal_code)	

Tabla: *ground_service*	**Descripción:** ground service (Transporte terrestre)	
Columna	**Tipo**	**Descripción**
city_code	Texto	código de ciudad
airport_code	Texto	código de aeropuerto
transport_code	Texto	código de transporte
ground_fare	Numérico	tarifa terrestre
LLAVE PRIMARIA:	city_code airport_code transport_code	
LLAVE FORANEA:	ground_service (transport_code) – transport (transport_code)	

Tabla: *month_name*	**Descripción:** month name (nombre del mes)	
Columna	**Tipo**	**Descripción**
month_number	Numérico	número del mes
month_name	Texto	nombre del mes
LLAVE PRIMARIA:	month_number	

Tabla: *restricción*	**Descripción:** restriction (restriccion)	
Columna	**Tipo**	**Descripción**
restrict_code	Texto	código de restricción
application	Texto	aplicación
no_discounts	Texto	sin descuento

reserve_ticket	Numérico	boleto en reserve
stopover	Texto	escalas
return_min	Numérico	permanencia mínima
return_max	Texto	permanencia máxima
LLAVE PRIMARIA:	restrict_code	

Tabla: *restrict_carrier*	**Descripción:** restrict carrier (restricciones de aerolínea)	
Columna	**Tipo**	**Descripción**
restrict_code	Texto	código de restricción
airline_code	Texto	código de aerolínea
LLAVE PRIMARIA:	restrict_code airline_code	
LLAVE FORANEA:	restrict_carrier (restrict_code) – restriction (restrict_code)	

Tabla: *restrict_class*	**Descripción:** restrict class (restricciones de aerolínea)	
Columna	**Tipo**	**Descripción**
restrict_code	Texto	código de restricción
ex_fare_class	Texto	clase de tarifa
LLAVE PRIMARIA:	ex_fare_class	
LLAVE FORANEA:	restric_class (restrict_code) – restriction (restrict_code)	

Tabla: *state*	**Descripción:** state (esatdo)	
Columna	**Tipo**	**Descripción**
state_code	Texto	código de estado
state_name	Texto	nombre del estado
country_name	Texto	nombre del país
LLAVE PRIMARIA:	state_code	

Tabla: *stop*	**Descripción:** Stop (Escala)	
Columna	**Tipo**	**Descripción**
flight_code	Numérico	código de vuelo
stop_number	Numérico	numero de escala
stop_flight	Numérico	escala de vuelo
LLAVE PRIMARIA:	flight_code stop_number	
LLAVE FORANEA:	stop (flight_code) – flight (flight_code) stop (stop_flight) – flight (stop_flight)	

Tabla: *interval_time*	**Descripción:** Interval time (Intervalo de tiempo)	
Columna	**Tipo**	**Descripción**
period	Texto	periodo
begin_time	Numérico	tiempo inicial
end_time	Numérico	tiempo final
LLAVE PRIMARIA:	period begin_time	

Tabla: *time_zone*	**Descripción:** Time zone (Zona horaria)	
Columna	**Tipo**	**Descripción**
time_zone_code	Texto	código de zona horaria
time_zone_name	Texto	nombre de zona horaria
LLAVE PRIMARIA:	time_zone_code	

Tabla: *transport*	**Descripción:** Transport (Transporte)	
Columna	**Tipo**	**Descripción**
transport_code	Texto	código de transporte
transport_desc	Texto	descripción de transporte
LLAVE PRIMARIA:	transport_code	

APÉNDICE B. Descripción de la Base de Datos Geobase

Geobase es una base de datos (BD) con información geográfica de los Estados Unidos. Fue implementada como una aplicación de ejemplo en la versión comercial para PC de TurboProlog 2.0 [Borland International, 1988].

Incluye información sobre población, área, ciudades capitales, estados colindantes, ríos principales, ciudades, principales, y puntos más altos y bajos junto con su altura. Así mismo contiene las longitudes de los ríos y el número de población de las ciudades. La principal razón para utilizar esta BD es que es utilizada como punto de referencia para evaluaciones de varias ILNBD de los 90's, ya que cuenta con una estructura simple y se encuentra disponible al público.

En este apéndice se muestra la BD cuyas tablas fueron definidas según los predicados de la versión original Geobase[5]:

1. state(name, abbreviation, capital, population, area, state_number, city1, city2, city3, city4)
2. city(state, state_abbreviation, name, population)
3. river(name, length, [states through which it flows])
4. border(state, state_abbreviation, [states that border it])
5. highlow(state, state_abbreviation, highest_point, highest_elevation, lowest_point, lowest_elevation)
6. mountain(state, state_abbreviation, name, height)
7. road(number, [states it passes through])
8. lake(name, area, [states it is in])

El esquema utilizado fue definido en [Verástegui, 2020], el cual cuenta con 11 tablas y 41 columnas, como se muestra en la Figura B.1.

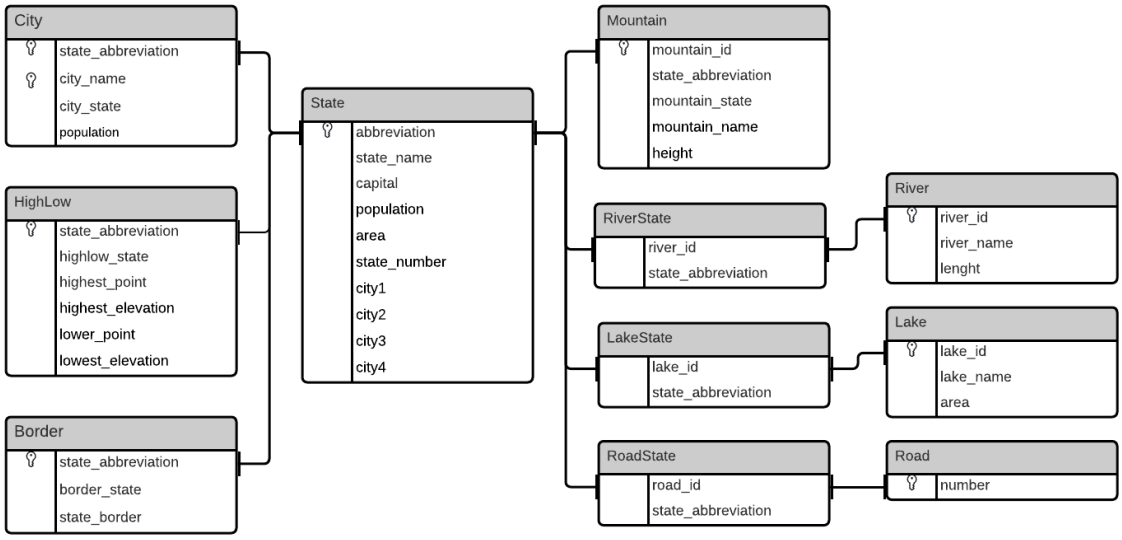


Figura B.1. Esquema de la base de datos Geobase

[5] https://www.cs.utexas.edu/ftp/mooney/nl-ilp-data/geosystem/geobase

Para proporcionar un mayor entendimiento del contenido de la BD Geobase, a continuación se presentan las descripciones del contenido de cada tabla y de cada columna que contiene. En la parte superior se muestra el nombre de cada tabla y su descripción. La primera columna muestra el nombre de las columnas de cada tabla, la segunda columna contiene el tipo de dato de cada columna, y la tercera muestra una breve descripción de las columnas.

Tabla B.1. Descripción del esquema de la base de datos Geobase

Tabla: *st|ate*	**Descripción:** State (Estado)	
Columna	**Tipo**	**Descripción**
abbreviation	Texto	abreviatura
state_name	Texto	estado
capital	Texto	capital
population	Numérico	población
area	Numérico	área
state_number	Numérico	numero de estado
city1	Texto	ciudad uno
city2	Texto	ciudad dos
city3	Texto	ciudad tres
city4	Texto	ciudad cuatro
LLAVE PRIMARIA:	abbreviation	

Tabla: *city*	**Descripción:** City (Ciudad)	
Columna	**Tipo**	**Descripción**
state_abbrebiation	Text	abreviatura de estado
city_name	Text	nombre de ciudad
city_state	Text	nombre de estado
population	Numérico	población
LLAVE PRIMARIA:	state_abbrebiation city_name	

Tabla: *HighLow*	**Descripción:** HighLow (Punto de estado)	
Columna	**Tipo**	**Descripción**
state_abbrebiation	Texto	abreviatura de estado
Highlow_state	Texto	nombre de estado
Highest_point	Texto	punto más alto
Highesr_elevation	Numérico	elevación más alta
Lowest_point	Texto	punto más bajo
Lowest_elevation	Numérico	elevación más baja
LLAVE PRIMARIA:	state_abbreviation	

Tabla: *Border*	**Descripción:** Border (Colindancia)	
Columna	**Tipo**	**Descripción**
state_abbrebiation	Texto	abreviatura de estado
border_state	Texto	nombre de estado
state_border	Texto	colindancia

Tabla: *Mountain*	**Descripción:** Mountain (Montaña)	
Columna	**Tipo**	**Descripción**
mountain_id	Numérico	identificador de montaña
state_abbrebiation	Texto	abreviatura de estado
mountain_state	Texto	nombre de estado
mountain_name	Texto	nombre de montaña
height	Numérico	altura
LLAVE PRIMARIA:	mountain_id	

Tabla: *River*	**Descripción:** River (Río)	
Columna	**Tipo**	**Descripción**
rive_id	Numérico	identificador de río
river_name	Texto	nombre de río
lenght	Numérico	longitud
LLAVE PRIMARIA:	rive_id	

Tabla: *RiverState*	**Descripción:** RiverState (Río Estado)	
Columna	**Tipo**	**Descripción**
rive_id	Numérico	identificador de río
state_abbrebiation	Texto	abreviatura de estado

Tabla: *Lake*	**Descripción:** Lake (Lago)	
Columna	**Tipo**	**Descripción**
lake_id	Numérico	identificador de lago
lake_name	Texto	nombre de lago
area	Numérico	área
LLAVE PRIMARIA:	lake_id	

Tabla: *LakeState*	**Descripción:** LakeState (Lago_Estado)	
Columna	**Tipo**	**Descripción**
lake_id	Numérico	identificador de lago
state_abbrebiation	Texto	abreviatura de estado

Tabla: *Road*	**Descripción:** Road (Carretera)	
Columna	**Tipo**	**Descripción**
number	Numérico	Carretera
LLAVE PRIMARIA:	lake_id	

Tabla: *RoadState*	**Descripción:** RoadState (Carretera_Estado)	
Columna	**Tipo**	**Descripción**
Road_id	Numérico	Carretera
state_abbrebiation	Texto	abreviatura de estado

APÉNDICE C. Resultados de las Pruebas del Preprocesamiento Léxico

La Tabla C.1 muestra un fragmento del corpus de prueba y los resultados del Experimento 1, descrito en la Subsección 5.3.1 del Capitulo 5. En ella se muestran los valores de búsqueda identificados en las consultas, los cuales se encuentran resaltados en negrita en la consulta. Todas las demás palabras fueron marcadas con la etiqueta *word*, indicando que serán las que se buscarán en el lexicón para extraer su(s) categoría(s) gramatical(es) (con el objetivo de facilitar la visualización, se omitió la etiqueta *word* y sólo se dejaron los valores de búsqueda).

Tabla C.1. Pruebas del preprocesamiento léxico

Geobase	
No.	**Consulta y valores de búsqueda.**
1	How many people live in **Hawaii**? **name**
2	What is the smallest city in the **USA**? **code**
3	Show the rivers in the state of **Florida**. **name**
4	What are the major cities in **Ohio**? **name**
5	What is the area of **Wisconsin**? **name**
6	What is the highest point in **Nevada** in meters? **name**
7	What is the lowest point of the state with the largest area? *No tiene valores de búsqueda.*
8	What is the shortest river in the **USA**? **code**
9	What rivers are in **Utah**? **name**
10	Which state borders **Florida**? **name**
ATIS	
No.	**Consulta y valores de búsqueda.**
1	Does **Delta Airlines** flight **179** serve breakfast? **code code number**
2	Give me all flights from **Atlanta** to **San Francisco** on **Delta** first class. **name name**, **name name**
3	Give me the flights in amphibious-type wide-body aircraft. *No tiene valores de búsqueda.*
4	Give me the flights of the airline **Tokened Airlines** that serve breakfast. **name name**
5	Show me all the flights where breakfast is served in wide-body airplanes. *No tiene valores de búsqueda.*
6	Show me all the wide-body flights from **Dallas** to **Denver** that serve lunch. **name name**
7	Which aircraft types are not wide-body and are not pressurized? *No tiene valores de búsqueda.*
8	Which airlines have economy flights with discount from **San Francisco** to **Dallas**? **name**, **name name**
9	Which airlines have flights that serve lunch in wide-body jets? *No tiene valores de búsqueda.*
10	Which flights have no discounted fare. *No tiene valores de búsqueda.*

APÉNDICE D. Resultados de las Pruebas del Etiquetado Léxico

La Tabla D.1 muestra los resultados del Experimento 2 (Subsección 5.3.3) referente al etiquetado del corpus de pruebas. Además, se presentan todas las categorías gramaticales extraídas del lexicón para cada una de las palabras que forman las consultas. Es importante señalar que, en este paso del analizador léxico, algunas palabras pueden incluir más de una categoría gramatical asignada.

Tabla D.1. Pruebas del etiquetado léxico

Geobase	
No.	**Consulta y categorías gramaticales.**
1	**How many people live in Hawaii?** adverb | indefinite adjective | noun | verb | adverb | name | adverb | indefinite pronoun | | preposition |
2	**Name the rivers in Arkansas.** verb | article | noun | preposition | name | | advrb |
3	**Show the rivers in the state of Florida.** verb | article | noun | preposition | article | noun | preposition | name | | adverb | | verb |
4	**What are the major cities in Ohio?** interrogative adjective | verb | article | noun | noun | adverb | name | interrogative pronoun | | preposition |
5	**What is the area of Wisconsin?** interrogative adjective | verb | article | noun | preposition | name | interrogative pronoun |
6	**What is the highest point in Nevada in meters?** interrogative adjevtive | verb | definite article | qualifying adjective | noun | preposition | name | preposition | noun | interrogative pronoun | | verb | advern | | advern |
7	**What is the lowest point of the state with the largest area?** Int adjective | verb | article | superlative adjective | verb | Prep | article | noun | Prep | article | superlative adjective | noun | Int pronoun | | noun | | verb |
8	**What is the shortest river in the USA?** interrogative adjective | verb | article | superlative adjective | noun | adverb | article | code | interrogative pronoun | | | | | preposition | | |
9	**What state contains the highest point in the US?** interrogative adjective (Int Adj) | verb | verb | article | adjective | verb | preposition (Prep) | article | code interrogative pronoun (Int Pro) | noun | | noun | adverb |
10	**Which state borders Florida?** interrogative adjective | noun | noun | name | interrogative pronoun | verb | verb |
ATIS	
No.	**Consulta y categorías gramaticales.**
1	**Does Delta Airlines flight 179 serve breakfast?** auxiliar verb | code | code | noun | num-ent | verb | noun |
2	**Give me all flights from Atlanta to San Francisco on Delta first class.** verb | pronoun | indefinite adjective | noun | Prep | name | preposition | name | name | adverb | name | adjective | noun | indefinite pronoun | | preposition | | adverb |
3	**Give me the flights in amphibious-type wide-body aircraft.** verb | pronoun | article | noun | preposition | qualifying adjective | qualifying adjective | noun | | adverb |
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | conjunction | verb | noun | | determiner | | pronoun |
5	**Show me all the flights where breakfast is served in wide-body airplanes.** verb | pronoun | indefinite adjective | article | noun | adverb | noun | verb | verb | adverb | qualifying adjective | noun | | indefinite pronoun | |conjunction | | preposition | | adverb |

6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.** verb | pronoun | indefinite adjective | article | adjective | noun | Prep | name | Prep | name | conjunction | verb | noun | indefinite pronoun | | pronoun | | adverb |
7	**Which aircraft types are not wide-body and are not pressurized?** interrogative adjective | noun | noun | verb | adverb | qualifying adjective | con_cop | verb | adverb | qualifying adjective | interrogative pronoun |
8	**Which airlines have economy flights with discount from San Francisco to Dallas?** interrogative adjective | noun | verb | qualifying adjective | noun | Prep | noun | preposition | name | name | preposition | name interrogative pronoun |
9	**Which airlines have flights that serve lunch in wide-body jets?** interrogative adjective | noun | verb | noun | conjunction | verb | noun | adverb | qualifying adjective | noun | interrogative pronoun | | pronoun | | preposition |
10	**Which flights have no discounted fare.** interrogative adjective | noun | verb | denial | verb | noun | interrogative pronoun |

APÉNDICE E. Resultados de las Pruebas del Postprocesamiento Léxico

En la Tabla E.1 se muestran los resultados del Experimento 3 (Subsección 5.3.5), específicamente, las puntuaciones asignadas a las categorías gramaticales de las palabras que presentan ambigüedad léxica. Finalmente, se asignará aquella categoría con el mayor puntaje (resaltada en negrita) como la categoría gramatical.

Tabla E.1. Pruebas del postprocesamiento léxico

Geobase	
No.	**Consulta y categorías gramaticales.**
1	**How many people live in Hawaii?** adverb | indefinite adjective 0 | noun | verb | adverb 2 | name | | **indefinite pronoun 1** | | **preposition 3** |
2	**Name the rivers in Arkansas.** verb | article | noun | **preposition 3** | name | | advrb 1 |
3	**Show the rivers in the state of Florida.** verb | article | noun | **preposition 2** | article | **noun 3** | preposition | name | | adverb 1 | | verb 0 |
4	**What are the major cities in Ohio?** interrogative adjective 0 | verb | article | noun | noun | adverb 1 | name | **interrogative pronoun 3** | | **preposition 3** |
5	**What is the area of Wisconsin?** interrogative adjective 0 | verb | article | noun | preposition | name | **interrogative pronoun 3** |
6	**What is the highest point in Nevada in meters?** interrogative adjevtive 0 | verb | definite article | qualifying adjective | **noun 3** | **preposition 3** | name | **preposition 3** | noun | **interrogative pronoun 3** | | verb 0 | adverb 1 | | advern 1 |
7	**What is the lowest point of the state with the largest area?** Int Adj 0 | verb | article | superlative adjective | verb 0 | Prep | article | **noun 3** | Prep | article | superlative adjective | noun | **Int Pron 3** | | **noun 3** | | verb 0 |
8	**What is the shortest river in the USA?** interrogative adjective 0 | verb | article | superlative adjective | noun | adverb 1 | article | code | **interrogative pronoun 3** | | **preposition 2** |
9	**What state** contains the highest **point in** the US**?** Int Adj 0 | **verb 3** | verb 0 | **preposition 2** | **Int Pro 3** | noun 0 | **noun 3** | adverb 1 |ode
10	**Which state borders Florida?** interrogative adjective 0 | **noun 3** | noun 2 | name | **interrogative pronoun 3** | verb 0 | **verb 3** |
ATIS	
No.	**Consulta y categorías gramaticales.**
1	**Does Delta Airlines flight 179 serves breakfast?** *No contiene palabras con múltiples categorías gramaticales.*
2	**Give me all flights from Atlanta to San Francisco on Delta first class.** verb | pronoun | indefinite adjective 0 | noun | Prep | name | Prep | name | name | adverb 1 | name | adjective | noun | **indefinite pronoun 3** | | **preposition 3** | | adverb 1 |
3	**Give me the flights in amphibious-type wide-body aircraft.** verb | pronoun | article | noun | **preposition 3** | qualifying adjective | qualifying adjective | noun | | adverb 1 |
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | **conjunction 3** | verb | noun | | determiner 0 | | pronoun 1 |
5	**Show me all the flights where breakfast is served in wide-body airplanes.** verb | pronoun | indefinite adjective 0 | article |noun | adverb 0 | noun | verb | verb | adverb 1 | qualifying adjective | noun | **indefinite pronoun 3** | |**conjunction 1** | | **Prep 3** | | adverb 1 |
6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.** verb | pronoun | indefinite adjective 0 | article | adjective | noun | Prep | name | Prep | name | **conjunction 3** | verb | noun

	| **indefinite pronoun 3** | | pronoun 0 | | adverb 1 |
7	**Which aircraft types are not wide-body and are not pressurized?** **interrogative adjective 3** | noun | noun | verb | adverb | qualifying adjective | con_cop | verb | adverb | qualifying adjective | interrogative pronoun 0 |
8	**Which airlines have economy flights with discount from San Francisco to Dallas?** **interrogative adjective 3** | noun | verb | qualifying adjective | noun | Prep | noun | Prep | name | name | preposition | name interrogative pronoun 0 |
9	**Which airlines have flights that serve lunch in wide-body jets?** **interrogative adjective 3** | noun | verb | noun | **conjunction 3** | verb | noun | adverb 1 | qualifying adjective | noun | interrogative pronoun 0 | | pronoun 1 | | **preposition 3** |
10	**Which flights have no discounted fare.** **interrogative adjective 3** | noun | verb | denial | verb | noun | interrogative pronoun 0 |

APÉNDICE F. Resultados de las Pruebas de Funcionalidad del Analizador Léxico

En la Tabla F.1 se muestran los resultados del Experimento 4 (Subsección 5.3.7), concretamente, las categorías asignadas a cada palabra del corpus de pruebas, así como el tiempo promedio que le tomó al analizador léxico procesar cada consulta.

Tabla F.1. Pruebas funcionalidad del analizador léxico

Geobase		
No.	**Consulta y categorías gramaticales.**	**Tiempo**
1	**How many rivers are there in Idaho?** adverb | indefinite adjective | noun | verb | adverb | preposition | name |	0.138
2	**Name the rivers in Arkansas.** verb | article | noun | preposition | name |	0.09
3	**Show the rivers in the state of Florida.** verb | article | noun | preposition | article | noun | preposition | name |	0.235
4	**What are the major cities in Ohio?** interrogative pronoun | verb | article | noun | noun | preposition | name |	0.127
5	**What is the area of Wisconsin?** interrogative pronoun | verb | article | noun | preposition | name |	0.094
6	**What is the highest point in Nevada in meters?** interrogative pronoun | verb | definite article | qualifying adjective | noun | preposition | name | preposition | noun |	0.118
7	**What is the highest point in the US?** interrogative pronoun | verb | article | qualifying adjective | noun | preposition | article | code |	0.116
8	**What is the shortest river in the USA?** interrogative pronoun | verb | article | superlative adjective | noun | preposition | article | code |	0.138
9	**What state contains the highest point in the US?** interrogative pronoun | noun | verb | article | qualifying adjective | noun | preposition | article | code |	0.125
10	**Which state borders Florida?** interrogative pronoun | noun | verb | name |	0.082
ATIS		
No.	**Consulta**	**Tiempo**
1	**Does Delta Airlines flight 179 serve breakfast?** auxiliar verb | code | code | noun | num-ent | verb | noun |	0.082
2	**Give me all flights from Atlanta to San Francisco on Delta first class.** verb | pronoun | indefinite pronoun | noun | Prep | name | Prep | name | name | Prep | name | qualifying Adj | noun |	0.139
3	**Give me the flights in amphibious-type wide-body aircraft.** verb | pronoun | article | noun | preposition | qualifying adjective | qualifying adjective | noun |	0.123
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** verb | pronoun | article | noun | preposition | article | noun | name | name | conjunction | verb | noun |	0.152
5	**Show me all the flights where breakfast is served in wide-body airplanes.** verb | pronoun | indefinite pronoun | article | noun | adverb | noun | verb | verb |Prep | qualifying adjective | noun |	0.154
6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.** verb | pronoun | indefinite pronoun | article | qualifying Adj | noun | Prep | name | Prep | name | that | verb | noun |	0.16
7	**Which aircraft types are not wide-body and are not pressurized?** interrogative adjective | noun | noun | verb | adverb | qualifying Adj | con_cop | verb | adverb | qualifying adjective	0.155
8	**Which airlines have economy flights with discount from San Francisco to Dallas?** interrogative adjective | noun | verb | qualifying adjective | noun |Prep | noun | Prep | name | name | Prep | name |	0.147
9	**Which airlines have flights that serve lunch in wide-body jets?** interrogative adjective | noun | verb | noun | conjunction | verb | noun | preposition | qualifying adjective | noun |	0.125
10	**Which flights have no discounted fare.** interrogative adjective | noun | verb | denial | verb | noun |	0.11

APÉNDICE G. Resultados de las Pruebas de Funcionalidad del AS-S en Inglés

La Tabla G.1 muestra los resultados del Experimento 5, descrito en la Subsección 5.4.1 del Capítulo 5. La tabla muestra las reducciones generadas por lo métodos de las reglas de producción, donde la primera columna muestra el identificador de la consulta, y la segunda columna muestra todas las reducciones realizadas a las 100 consultas del corpus de prueba. Dentro de la segunda columna del lado derecho se encuentra el identificador de la regla aplicada, por ejemplo, ***SNO1-V***, seguida de la secuencia de símbolos reducidos que recibe *inputExpression* en cada recorrido. Las reducciones se encuentran resaltadas en negrita. Por último, en la tercera columna se encuentra el tiempo promedio que le tomó al AS-S procesar cada consulta 20 veces.

Tabla G.1. Pruebas de funcionalidad del AS-S en inglés

Geobase		
No.	**Consulta y categorías gramaticales.**	**Time**
1	**How many people live in Hawaii?** EXPRESION ORIGINAL [adv, adj_ind, nou, ver, pre, name] SN01-V [adv, adj_ind, nou, ver, pre, **NouP0V**] SN02 [adv, adj_ind, **NouP0**, ver, pre, NouP0V] SA04 [**SAdj1**, NouP0, ver, pre, NouP0V] SN09 [**NouP1**, ver, pre, NouP0V] SN10-V [NouP1, ver, pre, **NouP1V**] SP01-V [NouP1, ver, **PrePV]** SV03 [NouP1, **VerP1**, PrePV] TS01 [**Suj**, VerP1, PrePV] TC01 [Suj, VerP1, **Com**] TO01 **[S E N T E N C E**]	**0.004**
2	**Name the rivers in Arkansas.** EXPRESION ORIGINAL [ver, art, nou, pre, name] SN01-V [ver, art, nou, pre, **NouP0V**] SN02 [ver, art, **NouP0**, pre, NouP0V] SN03 [ver, **NouP1**, pre, NouP0V] SV01 [**VerP1,** pre, NouP0V] SN10-V [VerP1, pre, **NouP1V**] SP01-V [VerP1, **PrePV**] TC01 [VerP1, **Com**] TO02 [**S E N T E N C E**]	**0.0**
3	**Show the rivers in the state of Florida.** EXPRESION ORIGINAL [ver, art, nou, pre, art, nou, pre, name] SN01-V [ver, art, nou, pre, art, nou, pre, **NouP0V**] SN02 [ver, art, **NouP0**, pre, art, nou, pre, NouP0V] SN02 [ver, art, NouP0, pre, art, **NouP0**, pre, NouP0V] SN03 [ver, **NouP1**, pre, art, NouP0, pre, NouP0V] SN03 [ver, NouP1, pre, **NouP1**, pre, NouP0V] SV01 [**VerP1**, pre, NouP1, pre, NouP0V] SN10-V [VerP1, pre, NouP1, pre, **NouP1V**] SP01 [VerP1, **PreP**, pre, NouP1V] SP01-V [VerP1, PreP, **PrePV**] TC01 [VerP1, **Com**, PrePV] TC01 [VerP1, Com, **Com**] TO02 [**S E N T E N C E**]	**0.002**
4	**What are the major cities in Ohio?** EXPRESION ORIGINAL [pro_int, ver, art, nou, pre, name] SN01-V [pro_int, ver, art, nou, pre, **NouP0V**] SN02 [pro_int, ver, art, **NouP0**, pre, NouP0V]	**0.001**

	SN03 [pro_int, ver, **NouP1**, pre, NouP0V] SV01 [pro_int, **VerP1**, pre, NouP0V] SN10-V [pro_int, VerP1, pre, **NouP1V**] SP01-V [pro_int, VerP1, **PrePV**] SV02 [**VerP1**, PrePV] TC01 [VerP1, **Com**] TO02 [**S E N T E N C E**]	
5	**What is the area of Wisconsin?** EXPRESION ORIGINAL [pro_int, ver, art, nou, pre, name] SN01-V [pro_int, ver, art, nou, pre, **NouP0V**] SN02 [pro_int, ver, art, **NouP0**, pre, NouP0V] SN03 [pro_int, ver, **NouP1**, pre, NouP0V] SV01 [pro_int, **VerP1**, pre, NouP0V] SN10-V [pro_int, VerP1, pre, **NouP1V**] SP01-V [pro_int, VerP1, **PrePV**] SV02 [**VerP1**, PrePV] TC01 [VerP1, **Com**] TO02 [**S E N T E N C E**]	**0.003**
6	**What is the highest point in Nevada in meters?** EXPRESION ORIGINAL [pro_int, ver, art, adj_cal, nou, pre, name, pre, nou] SN01-V [pro_int, ver, art, adj_cal, nou, pre, **NouP0V**, pre, nou] SN02 [pro_int, ver, art, **NouP0**, pre, NouP0V, pre, nou] SN02 [pro_int, ver, art, NouP0, pre, NouP0V, pre, **NouP0**] SN03 [pro_int, ver, **NouP1**, pre, NouP0V, pre, NouP0] SV01 [pro_int, **VerP1**, pre, NouP0V, pre, NouP0] SN10-V [pro_int, VerP1, pre, **NouP1V**, pre, NouP0] SP01 [pro_int, VerP1, pre, NouP1V, **PreP**] SP01-V [pro_int, VerP1, **PrePV**, PreP] SV02 [**VerP1**, PrePV, PreP] TC01 [VerP1, **Com**, PreP] TC01 [VerP1, Com, **Com**] TO02 [**S E N T E N C E**]	**0.002**
7	**What is the lowest point of the state with the largest area?** EXPRESION ORIGINAL [pro_int, ver, art, adj_sup, nou, pre, art, nou, pre, art, adj_sup, nou] SN02 [pro_int, ver, art, **NouP0**, pre, art, nou, pre, art, adj_sup, nou] SN02 [pro_int, ver, art, NouP0, pre, art, **NouP0**, pre, art, adj_sup, nou] SN02 [pro_int, ver, art, NouP0, pre, art, NouP0, pre, art, **NouP0**] SN03 [pro_int, ver, **NouP1**, pre, art, NouP0, pre, art, NouP0] SN03 [pro_int, ver, NouP1, pre, **NouP1**, pre, art, NouP0] SN03 [pro_int, ver, NouP1, pre, NouP1, pre, **NouP1**] SV01 [pro_int, **VerP1**, pre, NouP1, pre, NouP1] SP01 [pro_int, VerP1, **PreP**, pre, NouP1] SP01 [pro_int, VerP1, PreP, **PreP**] SV02 [**VerP1**, PreP, PreP] TC01 [VerP1, **Com**, PreP] TC01 [VerP1, Com, **Com**] TO02 [**S E N T E N C E**]	**0.028**
8	**What is the shortest river in the USA?** EXPRESION ORIGINAL [pro_int, ver, art, adj_sup, nou, pre, art, code] SN01-V [pro_int, ver, art, adj_sup, nou, pre, art, **NouP0V**] SN02-V [pro_int, ver, art, adj_sup, nou, pre, **NouP1V**] SN02 [pro_int, ver, art, **NouP0**, pre, NouP1V] SN03 [pro_int, ver, **NouP1**, pre, NouP1V] SV01 [pro_int, **VerP1**, pre, NouP1V] SP01-V [pro_int, VerP1, **PrePV**] SV02 [**VerP1**, PrePV] TC01 [VerP1, **Com**] TO02 [**S E N T E N C E**]	**0.002**
9	**What state contains the highest point in the US?** EXPRESION ORIGINAL [pro_int, nou, ver, art, adj_cal, nou, pre, art, code] SN01-V [pro_int, nou, ver, art, adj_cal, nou, pre, art, **NouP0V**] SN02-V [pro_int, nou, ver, art, adj_cal, nou, pre, **NouP1V**] SN02 [pro_int, **NouP0**, ver, art, adj_cal, nou, pre, NouP1V] SN02 [pro_int, NouP0, ver, art, **NouP0**, pre, NouP1V]	**0.003**

	SN03 [pro_int, NouP0, ver, **NouP1**, pre, NouP1V] SV01 [pro_int, NouP0, **VerP1**, pre, NouP1V] SP01-V [pro_int, NouP0, VerP1, **PrePV**] TS01 [**Suj**, VerP1, PrePV] TC01 [Suj, VerP1, **Com**] TO01 [**S E N T E N C E]**	
10	**Which state borders Florida?** EXPRESION ORIGINAL [pro_int, nou, ver, name] SN01-V [pro_int, nou, ver, **NouP0V**] SN02 [pro_int, **NouP0**, ver, NouP0V] SN10-V [pro_int, NouP0, ver, **NouP1V**] SV03 [pro_int, NouP0, **VerP1**, NouP1V] TS01 [**Suj**, VerP1, NouP1V] TC02 [Suj, VerP1, **Com**] TO01 [**S E N T E N C E**]	**0.001**
ATIS		
No.	**Consulta**	**Time**
1	**Does Delta Airlines flight 179 serve breakfast?** EXPRESION ORIGINAL [aux_ver, code, nou, num-ent, ver, nou] SN01-V [aux_ver, **NouP0V**, nou, num-ent, ver, nou] SN01-V [aux_ver, NouP0V, nou, **NouP0V**, ver, nou] SN02 [aux_ver, NouP0V, **NouP0**, NouP0V, ver, nou] SN02 [aux_ver, NouP0V, NouP0, NouP0V, ver, **NouP0**] SV01 [aux_ver, NouP0V, NouP0, NouP0V, **VerP1**] SN10-V [aux_ver, **NouP1V**, NouP0, NouP0V, VerP1] SN10-V [aux_ver, NouP1V, **NouP1V**, VerP1] TS02 [**Suj**, NouP1V, VerP1] TC02 [Suj, **Com**, VerP1] TO03 [**S E N T E N C E**]	**0.0**
2	**Give me all flights from Atlanta to San Francisco on Delta first class.** EXPRESION ORIGINAL [ver, pro, ind_pro, nou, pre, name, pre, name, pre, name, adj_cal, nou] SN01-V [ver, pro, ind_pro, nou, pre, **NouP0V**, pre, name, pre, name, adj_cal, nou] SN01-V [ver, pro, ind_pro, nou, pre, NouP0V, pre, **NouP0V**, pre, name, adj_cal, nou] SN01-V [ver, pro, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, **NouP0V**, adj_cal, nou] SN02 [ver, **NouP0**, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] SN02 [ver, NouP0, ind_pro, **NouP0**, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] SN02 [ver, NouP0, ind_pro, NouP0, pre, NouP0V, pre, NouP0V, pre, **NouP0**] SN03 [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] SV01 [**VerP1**, NouP1, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] SN10-V [VerP1, NouP1, pre, **NouP1V**, pre, NouP0V, pre, NouP0V, NouP0] SN10-V [VerP1, NouP1, pre, NouP1V, pre, **NouP1V**, pre, NouP0V, NouP0] SN10-V [VerP1, NouP1, pre, NouP1V, pre, NouP1V, pre, **NouP1V**, NouP0] SP01-V [VerP1, NouP1, **PrePV**, pre, NouP1V, pre, NouP1V, NouP0] SP01-V [VerP1, NouP1, PrePV, **PrePV**, pre, NouP1V, NouP0] SP01-V [VerP1, NouP1, PrePV, PrePV, **PrePV**, NouP0] TC01 [VerP1, NouP1, **Com**, PrePV, PrePV, NouP0] TC01 [VerP1, NouP1, Com, **Com**, PrePV, NouP0] TC01 [VerP1, NouP1, Com, Com, **Com**, NouP0] TC02 [VerP1, **Com**, Com, Com, Com, NouP0] TC02 [VerP1, Com, Com, Com, Com, **Com**] TO02 [**S E N TE N C E]**	**0.083**
3	**Give me the flights in amphibious-type wide-body aircraft.** EXPRESION ORIGINAL [ver, pro, art, nou, pre, adj_cal, nou] SN02 [ver, **NouP0**, art, nou, pre, adj_cal, nou] SN02 [ver, NouP0, art, **NouP0**, pre, adj_cal, nou] SN02 [ver, NouP0, art, NouP0, pre, **NouP0**] SN03 [ver, NouP0, **NouP1**, pre, NouP0] SV01 [**VerP1**, NouP1, pre, NouP0] SP01 [VerP1, NouP1, **PreP**] TC01 [VerP1, NouP1, **Com**] TC02 [VerP1, **Com**, Com] TO02 [**S E N T E N C E**]	**0.004**

4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** EXPRESION ORIGINAL [ver, pro, art, nou, pre, art, nou, name, that, ver, nou] SN01-V [ver, pro, art, nou, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, nou, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, nou, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, SNom0, SNom0V, that, ver, nou] SN02 [ver, SNom0, art, SNom0, pre, art, SNom0, SNom0V, that, ver, SNom0] SN03 [ver, SNom0, SNom1, pre, art, SNom0, SNom0V, that, ver, SNom0] SN03 [ver, SNom0, SNom1, pre, SNom1, SNom0V, that, ver, SNom0] SV01 [SVer1, SNom1, pre, SNom1, SNom0V, that, ver, SNom0] SV01 [SVer1, SNom1, pre, SNom1, SNom0V, that, SVer1] SN10-V [SVer1, SNom1, pre, SNom1V, that, SVer1] SP01-V [SVer1, SNom1, SPreV, that, SVer1] SP02-V [SVer1, SNom1, SPreV] TC01 [SVer1, SNom1, Com] TC02 [SVer1, Com, Com] TO02_1 [S E N T E N C E]	**0.039**
5	**Show me all the flights where breakfast is served in wide-body airplanes.** EXPRESION ORIGINAL [ver, pro, ind_pro, art, nou, adv, nou, ver, pre, adj_cal, nou] SN02 [ver, **NouP0**, ind_pro, art, nou, adv, nou, ver, pre, adj_cal, nou] SN02 [ver, NouP0, ind_pro, art, **NouP0**, adv, nou, ver, pre, adj_cal, nou] SN02 [ver, NouP0, ind_pro, art, NouP0, adv, **NouP0**, ver, pre, adj_cal, nou] SN02 [ver, NouP0, ind_pro, art, NouP0, adv, NouP0, ver, pre, **NouP0**] SN03 [ver, NouP0, **NouP1**, adv, NouP0, ver, pre, NouP0] SN03 [ver, NouP0, NouP1, **NouP1**, ver, pre, NouP0] SV01 [**VerP1**, NouP1, NouP1, ver, pre, NouP0] SP01 [VerP1, NouP1, NouP1, ver, **PreP**] SV03 [VerP1, NouP1, NouP1, **VerP1**, PreP] TC01 [VerP1, NouP1, NouP1, VerP1, **Com**] TC02 [VerP1, Com, NouP1, VerP1, **Com**] TC02 [VerP1, **Com**, Com, VerP1, Com] TO03 [**S E N T E N C E**]	**0.024**
6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.** EXPRESION ORIGINAL [ver, pro, ind_pro, art, adj_cal, nou, pre, name, pre, name, that, ver, nou] SN01-V [ver, pro, ind_pro, art, adj_cal, nou, pre, **NouP0V**, pre, name, that, ver, nou] SN01-V [ver, pro, ind_pro, art, adj_cal, nou, pre, NouP0V, pre, **NouP0V**, that, ver, nou] SN02 [ver, **NouP0**, ind_pro, art, adj_cal, nou, pre, NouP0V, pre, NouP0V, that, ver, nou] SN02 [ver, NouP0, ind_pro, art, **NouP0**, pre, NouP0V, pre, NouP0V, that, ver, nou] SN02 [ver, NouP0, ind_pro, art, NouP0, pre, NouP0V, pre, NouP0V, that, ver, **NouP0**] SN03 [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, that, ver, NouP0] SV01 [**VerP1**, NouP1, pre, NouP0V, pre, NouP0V, that, ver, NouP0] SV01 [VerP1, NouP1, pre, NouP0V, pre, NouP0V, that, **VerP1**] SN10-V [VerP1, NouP1, pre, **NouP1V**, pre, NouP0V, that, VerP1] SN10-V [VerP1, NouP1, pre, NouP1V, pre, **NouP1V**, that, VerP1] SP01-V [VerP1, NouP1, **PrePV**, pre, NouP1V, that, VerP1] SP01-V [VerP1, NouP1, PrePV, **PrePV,** that, VerP1] SP02-V [VerP1, NouP1, PrePV, **PrePV**] TC01 [VerP1, NouP1, **Com**, PrePV] TC01 [VerP1, NouP1, Com, **Com**] TC02 [VerP1, **Com**, Com, Com] TO02 [**S E N T E N C E**]	**0.066**
7	**Which aircraft types are not wide-body and are not pressurized?** EXPRESION ORIGINAL [adj_int, nou, ver, adv, adj_cal, con_cop, ver, adv, adj_cal] SN02 [adj_int, **NouP0**, ver, adv, adj_cal, con_cop, ver, adv, adj_cal] SA07 [adj_int, NouP0, ver, adv, **SAdj0**, con_cop, ver, adv, adj_cal] SA07 [adj_int, NouP0, ver, adv, SAdj0, con_cop, ver, adv, **SAdj0**] SV03 [adj_int, NouP0, **VerP1**, con_cop, ver, adv, SAdj0] SV03 [adj_int, NouP0, VerP1, con_cop, **VerP1**] SV11 [adj_int, NouP0, **VerP3**] TS01 [**Suj**, VerP3] TO01 [**S E N T E N C E**]	**0.002**

8	Which airlines have economy flights with discount from San Francisco to Dallas? EXPRESION ORIGINAL [adj_int, nou, ver, adj_cal, nou, pre, nou, pre, name, pre, name] SN01-V [adj_int, nou, ver, adj_cal, nou, pre, nou, pre, **NouP0V**, pre, name] SN01-V [adj_int, nou, ver, adj_cal, nou, pre, nou, pre, NouP0V, pre, **NouP0V**] SN02 [adj_int, **NouP0**, ver, adj_cal, nou, pre, nou, pre, NouP0V, pre, NouP0V] SN02 [adj_int, NouP0, ver, **NouP0**, pre, nou, pre, NouP0V, pre, NouP0V] SN02 [adj_int, NouP0, ver, NouP0, pre, **NouP0**, pre, NouP0V, pre, NouP0V] SV01 [adj_int, NouP0, **VerP1**, pre, NouP0, pre, NouP0V, pre, NouP0V] SN10-V [adj_int, NouP0, VerP1, pre, NouP0, pre, **NouP1V**, pre, NouP0V] SN10-V [adj_int, NouP0, VerP1, pre, NouP0, pre, NouP1V, pre, **NouP1V**] SP01 [adj_int, NouP0, VerP1, **PreP**, pre, NouP1V, pre, NouP1V] SP01-V [adj_int, NouP0, VerP1, PreP, **PrePV**, pre, NouP1V] SP01-V [adj_int, NouP0, VerP1, PreP, PrePV, **PrePV**] TS01 [**Suj**, VerP1, PreP, PrePV, PrePV] TC01 [Suj, VerP1, **Com**, PrePV, PrePV] TC01 [Suj, VerP1, Com, **Com**, PrePV] TC01 [Suj, VerP1, Com, Com, **Com**] TO01 [**S E N T E N C E**]	0.053
9	**Which airlines have flights that serve lunch in wide-body jets?** EXPRESION ORIGINAL [adj_int, nou, ver, nou, that, ver, nou, pre, adj_cal, nou] SN02 [adj_int, **NouP0**, ver, nou, that, ver, nou, pre, adj_cal, nou] SN02 [adj_int, NouP0, ver, **NouP0**, that, ver, nou, pre, adj_cal, nou] SN02 [adj_int, NouP0, ver, NouP0, that, ver, **NouP0**, pre, adj_cal, nou] SN02 [adj_int, NouP0, ver, NouP0, that, ver, NouP0, pre, **NouP0**] SV01 [adj_int, NouP0, **VerP1**, that, ver, NouP0, pre, NouP0] SV01 [adj_int, NouP0, VerP1, that, **VerP1**, pre, NouP0] SP01 [adj_int, NouP0, VerP1, that, VerP1, **PreP**] SN15 [adj_int, NouP0, **VerP3**, PreP] TS01 [**Suj**, VerP3, PreP] TC01 [Suj, VerP3, **Com**] TO01 [**S E N T E N C E**]	**0.011**
10	**Which flights have no discounted fare.** EXPRESION ORIGINAL [adj_int, nou, ver, qua_adj, ver, nou] SN02 [adj_int, **NouP0**, ver, qua_adj, ver, nou] SN02 [adj_int, NouP0, ver, qua_adj, ver, **NouP0**] SV01 [adj_int, NouP0, ver, qua_adj, **VerP1**] SV02 [adj_int, NouP0, **VerP1**] TS01 [**Suj**, VerP1] TO01 [**S E N T E N C E**]	**0.005**

APÉNDICE H. Resultados de las Pruebas Comparativas del Analizador Léxico

La Tabla H.1 muestra los resultados obtenidos del Experimento 6, descrito en la Subsección 5.4.1 del Capítulo 5. En ella se muestra el etiquetado de las consultas realizado por CLAWS (CLS), Wolfram (WLF) y el Analizador Léxico en inglés (AL_Eng). La nomenclatura de CLAWS y Wolfram se adaptó a la gramática descrita en la Subsección 2.5.1.

Tabla H.1. Pruebas comparativas del analizador léxico

Geobase	
No.	**Consulta**
1	**How many people live in Hawaii?** CLS: adv | det | nou | ver | pre | PN WLF: adv | adj | nou | ver | pre | PN ALEng: adv | adj_ind | nou | ver | pre | name
2	**Name the rivers in Arkansas.** CLS: ver | art | nou | pre | PN WLF: ver | det | nou | pre | PN ALEng: ver | art | nou | pre | name
3	**Show the rivers in the state of Florida.** CLS: ver | art | nou | pre | art | nou | pre | PN WLF: ver | det | nou | pre | det | nou | pre | PN ALEng: ver | art | nou | pre | art | nou | pre | name
4	**What are the major cities in Ohio?** CLS: Q_det | ver | art | adj | nou | pre | PN WLF: pro_int | ver | det | adj | nou | pre | PN ALEng: pro_int | ver | art | nou | nou | pre | name
5	**What is the area of Wisconsin?** CLS: Q_det | ver | art | nou | pre | PN WLF: pro_int | ver | det | nou | pre | PN ALEng: pro_int | ver | art | nou | pre | name
6	**What is the highest point in Nevada in meters?** CLS: Q_det | ver | art | adj_sup | nou | pre | PN | pre | nou WLF: pro_int | ver | det | adj | nou | pre | PN | pre | nou ALEng: pro_int | ver | art | adj_cal | nou | pre | name | pre | nou
7	**What is the lowest point of the state with the largest area?** CLS: Q_det | ver | art | adj_sup | nou | pre | art | nou | pre | art | adj_sup | nou WLF: pro_int | ver | det | adj | nou | pre | det | nou | pre | det | adj | nou ALEng: pro_int | ver | art | adj_sup | nou | pre | art | nou | pre | art | adj_sup | nou
8	**What is the shortest river in the USA?** CLS: Q_det | ver | art | adj_sup | nou | pre | art | PN WLF: pro_int | ver | det | adj | nou | pre | det | PN ALEng: pro_int | ver | art | adj_sup | nou | pre | art | code |
9	**What state contains the highest point in the US?** CLS: Q_det | nou | ver | art | adj_sup | nou | pre | PN WLF: pro_int | ver | det | adj | adj | nou | pre | det | PN ALEng: pro_int | nou | ver | art | adj_cal | nou | pre | art | code |
10	**Which state borders Florida?** CLS: Q_det | nou | ver | PN WLF: Q_det | nou | ver | PN ALEng: pro_int | nou | ver | name

ATIS	
No.	**Consulta**
1	**Does Delta Airlines flight 179 serves breakfast?** CLS: ver | nou | nou | nou | num | ver | nou WLF: PN | PN | PN | ver | num | num | nou ALEng: aux_ver | code | code | nou | num-ent | ver | nou
2	**Give me all flights from Atlanta to San Francisco on Delta first class.** CLS: ver | pro | det | nou | pre | PN | pre | PN | PN | pre | nou | num | nou WLF: ver | pro | det | nou | pre | PN | pre | PN | PN | pre | PN | adj | nou ALEng: ver | pro | ind_pro | nou | pre | name | pre | name | name | pre | name | adj_cal | nou |
3	**Give me the flights in amphibious-type wide-body aircraft.** CLS: ver | pro | art | nou | pre | adj | adj | nou WLF: ver | pro | det | nou | pre | adj | adj | nou ALEng: ver | pro | art | nou | pre | adj_cal | adj_cal | nou
4	**Give me the flights of the airline Tokened Airlines that serve breakfast.** CLS: ver | pro | art | nou | pre | art | nou | adj | nou | con | ver | nou WLF: ver | pro | det | nou | pre | det | nou | ver | nou | Q_det | ver | nou ALEng: ver | pro | art | nou | pre | art | nou | name | name | con | ver | nou
5	**Show me all the flights where breakfast is served in wide-body airplanes.** CLS: ver | pro | det | art | nou | adv | nou | ver | ver | pre | adj | nou WLF: ver | pro | det | det | nou | QAdv | nou | ver | ver | pre | adj | nou ALEng: ver | pro | ind_pro | art | nou | adv | nou | ver | ver | pre | adj_cal | nou
6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.** CLS: ver | pro | det | art | adj | nou | pre | PN | pre | PN | con | ver | nou WLF: ver | pro | adv | det | adj | nou | pre | PN | pre | PN | Q_det | ver | nou ALEng: ver | pro | ind_pro | art | adj_cal | nou | pre | name | pre | name | that | ver | nou |
7	**Which aircraft types are not wide-body and are not pressurized?** CLS: Q_det | nou | nou | ver | not | adj | con | ver | not | ver WLF: Q_det | adj | nou | ver | adv | adj | con | ver | adv | ver ALEng: adj_int | nou | nou | ver | adv | adj_cal | con_cop | ver | adv | adj_cal |
8	**Which airlines have economy flights with discount from San Francisco to Dallas?** CLS: Q_det | nou | ver | nou | nou | pre | nou | pre | PN | PN | pre | PN WLF: Q_det | nou | ver | nou | nou | pre | nou | pre | PN | PN | pre | PN | ALEng: adj_int | nou | ver | adj_cal | nou | pre | nou | pre | name | name | pre | name |
9	**Which airlines have flights that serve lunch in wide-body jets?** CLS: Q_det | nou | ver | nou | con | ver | nou | pre | adj | nou WLF: Q_det | nou | ver | nou | Q_det | ver | nou | pre | adj | nou ALEng: adj_int | nou | ver | nou | that | ver | nou | pre | adj_cal | nou
10	**Which flights have no discounted fare.** CLS: Q_det | nou | ver | art | adj | nou WLF: adj | nou | ver | det | adj | nou ALEng: adj_int | nou | ver | qua_adj | ver | nou |

APÉNDICE I. Resultados de las Pruebas Comparativas del AS-S

La Tabla I.1 muestra los resultados obtenidos del Experimento 7, descrito en la Subsección 5.5.3 del Capítulo 5. La tabla muestra las reducciones generadas por Wolfram (Wlf) y el Analizador Sintáctico-Semántico (AS-S). La tabla muestra las reducciones generadas por Wolfram adaptadas a la gramática definida en la Subsección 2.5.1 (debido a que la estructura utilizada por Wolfram es diferente a la del AS-S) como se muestra en la Figura 2.6.

Tabla I.1. Pruebas comparativas del AS-S en inglés

Geobase	
No.	**Consulta**
1	**How many people live in Hawaii?**
AS-S en inglés	**Wolfram**
[adv, adj_ind, nou, ver, pre, name] [adv, adj_ind, nou, ver, pre, **NouP0V**] [adv, adj_ind, **NouP0**, ver, pre, NouP0V] [**SAdj1**, NouP0, ver, pre, NouP0V] [**NouP1**, ver, pre, NouP0V] [NouP1, ver, pre, **NouP1V**] [NouP1, ver, **PrePV**] [NouP1, **VerP1**, PrePV] [**Suj**, VerP1, PrePV] [Suj, VerP1, **Com**] [**S E N T E N C E**]	[adv, adj, nou, ver, pre, PN] [**QAdjP**, nou, ver, pre, PN] [QAdjP, nou, ver, pre, **NouP**] [QNouP, ver, pre, **NouP**] [QNouP, ver, **PreP**] [**C L A U S E**]
2	**Name the rivers in Arkansas.**
AS-S en inglés	**Wolfram**
[ver, art, nou, pre, name] [ver, art, nou, pre, **NouP0V**] [ver, art, **NouP0**, pre, NouP0V] [ver, **NouP1**, pre, NouP0V] [**VerP1**, pre, NouP0V] [VerP1, pre, **NouP1V**] [VerP1, **PrePV**] [VerP1, **Com**] [**S E N T E N C E**]	[ver, det, nou, pre, PN] [ver, **NouP**, pre, PN] [ver, NouP, pre, **NouP**] [ver, NouP, **PreP**] [ver, **NouP**] [**VerP**] [**S E N T E N C E**]
3	**Show the rivers in the state of Florida.**
AS-S en inglés	**Wolfram**
[ver, art, nou, pre, art, nou, pre, name] [ver, art, nou, pre, art, nou, pre, **NouP0V**] [ver, art, **NouP0**, pre, art, nou, pre, NouP0V] [ver, art, NouP0, pre, art, **NouP0**, pre, NouP0V] [ver, **NouP1**, pre, art, NouP0, pre, NouP0V] [ver, NouP1, pre, **NouP1**, pre, NouP0V] [**VerP1**, pre, NouP1, pre, NouP0V] [VerP1, pre, NouP1, pre, **NouP1V**] [VerP1, **PreP**, pre, NouP1V] [VerP1, PreP, **PrePV**] [VerP1, **Com**, PrePV] [VerP1, Com, **Com**] [**S E N T E N C E**]	[ver, det, nou, pre, det, nou, pre, PN] [ver, **NouP**, pre, det, nou, pre, PN] [ver, NouP, pre, **NouP**, pre, PN] [ver, NouP, pre, NouP, pre, **NouP**] [ver, NouP, pre, NouP, **PreP**] [ver, NouP, pre, **NouP**] [ver, NouP, **PreP**] [**VerP**] [**S E N T E N C E**]
4	**What are the major cities in Ohio?**
AS-S en inglés	**Wolfram**
[pro_int, ver, art, nou, pre, name] [pro_int, ver, art, nou, pre, **NouP0V**] [pro_int, ver, art, **NouP0**, pre, NouP0V] [pro_int, ver, **NouP1**, pre, NouP0V] [pro_int, **VerP1**, pre, NouP0V] [pro_int, VerP1, pre, **NouP1V**] [pro_int, VerP1, **PrePV**] [**VerP1**, PrePV] [VerP1, **Com**]	[pro_int, ver, det, nou, pre, PN] [**QNouP**, ver, det, nou, pre, PN] [QNouP, ver, **NouP**, pre, PN] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**] [QNouP, ver, **NouP**] [**C L A U S E**]

[**S E N T E N C E**]	
5	**What is the area of Wisconsin?**
AS-S en inglés	**Wolfram**
[pro_int, ver, art, nou, pre, name] [pro_int, ver, art, nou, pre, **NouP0V**] [pro_int, ver, art, **NouP0**, pre, NouP0V] [pro_int, ver, **NouP1**, pre, NouP0V] [pro_int, **VerP1**, pre, NouP0V] [pro_int, VerP1, pre, **NouP1V**] [pro_int, VerP1, **PrePV**] [**VerP1**, PrePV] [VerP1, **Com**] [**S E N T E N C E**]	[pro_int, ver, det, nou, pre, PN] [**QNouP**, ver, det, nou, pre, PN] [QNouP, ver, **NouP**, pre, PN] [QNouP, ver, NouP, pre, NouP] [QNouP, ver, NouP, **PreP**] [QNouP, ver, **NouP**] [QNouP, **VerP**] [**C L A U S E**]
6	**What is the highest point in Nevada in meters?**
AS-S en inglés	**Wolfram**
[pro_int, ver, art, adj_cal, nou, pre, name, pre, nou] [pro_int, ver, art, adj_cal, nou, pre, **NouP0V**, pre, nou] [pro_int, ver, art, **NouP0**, pre, NouP0V, pre, nou] [pro_int, ver, art, NouP0, pre, NouP0V, pre, **NouP0**] [pro_int, ver, **NouP1**, pre, NouP0V, pre, NouP0] [pro_int, **VerP1**, pre, NouP0V, pre, NouP0] [pro_int, VerP1, pre, **NouP1V**, pre, NouP0] [pro_int, VerP1, pre, NouP1V, **PreP**] [pro_int, VerP1, **PrePV**, PreP] [**VerP1**, PrePV, PreP] [VerP1, **Com**, PreP] [VerP1, Com, **Com**] [**S E N T E N C E**]	[pro_int, ver, det, adj, nou, pre, PN, pre, nou] [**QNouP**, ver, det, adj, nou, pre, PN, pre, nou] [QNouP, ver, **NouP**, pre, PN, pre, nou] [QNouP, ver, NouP, pre, **NouP**, pre, nou] [QNouP, ver, NouP, pre, NouP, pre, **NouP**] [QNouP, ver, NouP, pre, NouP, **PreP**] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**] [QNouP, ver, **NouP**] [**C L A U S E**]
7	**What is the lowest point of the state with the largest area?**
AS-S en inglés	**Wolfram**
[pro_int, ver, art, adj_sup, nou, pre, art, nou, pre, art, adj_sup, nou] [pro_int, ver, art, **NouP0**, pre, art, nou, pre, art, adj_sup, nou] [pro_int, ver, art, NouP0, pre, art, **NouP0**, pre, art, adj_sup, nou] [pro_int, ver, art, NouP0, pre, art, NouP0, pre, art, **NouP0**] [pro_int, ver, **NouP1**, pre, art, NouP0, pre, art, NouP0] [pro_int, ver, NouP1, pre, **NouP1**, pre, art, NouP0] [pro_int, ver, NouP1, pre, NouP1, pre, **NouP1**] [pro_int, **VerP1**, pre, NouP1, pre, NouP1] [pro_int, VerP1, **PreP**, pre, NouP1] [pro_int, VerP1, PreP, **PreP**] [**VerP1**, PreP, PreP] [VerP1, **Com**, PreP] [VerP1, Com, Com] [**S E N T E N C E**]	[pro_int, ver, det, adj, nou, pre, det, nou, pre, det, adj, nou] [**QNouP**, ver, det, adj, nou, pre, det, nou, pre, det, adj, nou] [QNouP, ver, **NouP**, pre, det, nou, pre, det, adj, nou] [QNouP, ver, NouP, pre, **NouP**, pre, det, adj, nou] [QNouP, ver, NouP, pre, NouP, pre, **NouP**] [QNouP, ver, NouP, pre, NouP, **PreP**] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**] [QNouP, ver, **NouP**] [QNouP, **VerP**] [**C L A U S E**]
8	**What is the shortest river in the USA?**
AS-S en inglés	**Wolfram**
[pro_int, ver, art, adj_sup, nou, pre, art, code] [pro_int, ver, art, adj_sup, nou, pre, art, **NouP0V**] [pro_int, ver, art, adj_sup, nou, pre, **NouP1V**] [pro_int, ver, art, **NouP0**, pre, NouP1V] [pro_int, ver, **NouP1**, pre, NouP1V] [pro_int, **VerP1**, pre, NouP1V] [pro_int, VerP1, **PrePV**] [**VerP1**, PrePV] [VerP1, **Com**] [**S E N T E N C E**]	[pro_int, ver, det, adj, nou, pre, det, PN] [**QNouP**, ver, det, adj, nou, pre, det, PN] [QNouP, ver, **NouP**, pre, det, PN] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**] [QNouP, ver, **NouP**] [**C L A U S E**]
9	**What state contains the highest point in the US?**
AS-S en inglés	**Wolfram**
[pro_int, nou, ver, art, adj_cal, nou, pre, art, code] [pro_int, nou, ver, art, adj_cal, nou, pre, art, **NouP0V**] [pro_int, nou, ver, art, adj_cal, nou, pre, **NouP1V**] [pro_int, **NouP0**, ver, art, adj_cal, nou, pre, NouP1V] [pro_int, NouP0, ver, art, **NouP0**, pre, NouP1V] [pro_int, NouP0, ver, **NouP1**, pre, NouP1V] [pro_int, NouP0, **VerP1**, pre, NouP1V] [pro_int, NouP0, VerP1, **PrePV**] [**Suj**, VerP1, PrePV] [Suj, VerP1, **Com**]	[pro_int, ver, det, adj, adj, nou, pre, det, PN] [**QNouP**, ver, det, adj, adj, pre, det, PN] [QNouP, ver, **NouP**, pre, det, PN] [QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, **PreP**] [QNouP, **VerP**] [**C L A U S E**]

[**S E N T E N C E**]	
10 **Which state borders Florida?**	
AS-S en inglés	**Wolfram**
[pro_int, nou, ver, name] [pro_int, nou, ver, **NouP0V**] [pro_int, **NouP0**, ver, NouP0V] [pro_int, NouP0, ver, **NouP1V**] [pro_int, NouP0, **VerP1**, NouP1V] [**Suj**, VerP1, NouP1V] [Suj, VerP1, **Com**] [**S E N T E N C E**]	[Q_det, nou, ver, PN] [**QNouP**, ver, PN] [QNouP, ver, **NouP**] [QNouP, **VerP**] [**C L A U S E**]
ATIS	
1 **Does Delta Airlines flight 179 serve breakfast?**	
AS-S en inglés	**Wolfram**
[aux_ver, code, nou, num-ent, ver, nou] [aux_ver, **NouP0V**, nou, num-ent, ver, nou] [aux_ver, NouP0V, nou, **NouP0V**, ver, nou] [aux_ver, NouP0V, NouP0, **NouP0V**, ver, nou] [aux_ver, NouP0V, NouP0, NouP0V, ver, **NouP0**] [aux_ver, NouP0V, NouP0, NouP0V, **VerP1**] [aux_ver, **NouP1V**, NouP0, NouP0V, VerP1] [aux_ver, NouP1V, **NouP1V**, VerP1] [**Suj**, NouP1V, VerP1] [Suj, **Com**, VerP1] [**S E N T E N C E**]	[PN, PN, PN, ver, num, num, nou] [PN, **NouP**, ver, num, num, nou] [PN, NouP, ver, **QuanP**, nou] [PN, NouP, ver, **NouP**] [PN, NouP, **VerP**] [**C L A U S E**]
2 **Give me all flights from Atlanta to San Francisco on Delta first class.**	
AS-S en inglés	**Wolfram**
[ver, pro, ind_pro, nou, pre, name, pre, name, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, **NouP0V**, pre, name, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, NouP0V, pre, **NouP0V**, pre, name, adj_cal, nou] [ver, pro, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, **NouP0V**, adj_cal, nou] [ver, **NouP0**, ind_pro, nou, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] [ver, NouP0, ind_pro, **NouP0**, pre, NouP0V, pre, NouP0V, pre, NouP0V, adj_cal, nou] [ver, NouP0, ind_pro, NouP0, pre, NouP0V, pre, NouP0V, pre, NouP0V, **NouP0**] [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] [**VerP1**, NouP1, pre, NouP0V, pre, NouP0V, pre, NouP0V, NouP0] [VerP1, NouP1, pre, **NouP1V**, pre, NouP0V, pre, NouP0V, NouP0] [VerP1, NouP1, pre, NouP1V, pre, **NouP1V**, pre, NouP0V, NouP0] [VerP1, NouP1, pre, NouP1V, pre, NouP1V, pre, **NouP1V**, NouP0] [VerP1, NouP1, **PrePV**, pre, NouP1V, pre, NouP1V, NouP0] [VerP1, NouP1, PrePV, **PrePV**, pre, NouP1V, NouP0] [VerP1, NouP1, PrePV, PrePV, **PrePV**, NouP0] [VerP1, NouP1, **Com**, PrePV, PrePV, NouP0] [VerP1, NouP1, Com, **Com**, PrePV, NouP0] [VerP1, NouP1, Com, Com, **Com**, NouP0] [VerP1, **Com**, Com, Com, Com, NouP0] [VerP1, Com, Com, Com, Com, **Com**] [**S E N TE N C E**]	[ver, pro, det, nou, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, **NouP**, det, nou, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, **NouP**, pre, PN, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, NouP, pre, **NouP**, pre, PN, PN, pre, PN, adj, nou] [ver, NouP, NouP, pre, NouP, pre, **NouP**, pre, PN, adj, nou] [ver, NouP, NouP, pre, NouP, pre, NouP, pre, **NouP**] [ver, NouP, NouP, pre, NouP, pre, NouP, **PreP**] [ver, NouP, NouP, pre, NouP, pre, **NouP**] [ver, NouP, NouP, pre, NouP, **PreP**] [ver, NouP, NouP, pre, **NouP**] [ver, NouP, NouP, **PreP**] [ver, NouP, **NouP**] [**VerP**] [**S E N TE N C E**]
3 **Give me the flights in amphibious-type wide-body aircraft.**	
AS-S en inglés	**Wolfram**
[ver, pro, art, nou, pre, adj_cal, nou] [ver, NouP0, art, nou, pre, adj_cal, nou] [ver, NouP0, art, NouP0, pre, adj_cal, nou] [ver, NouP0, art, NouP0, pre, NouP0] [ver, NouP0, NouP1, pre, NouP0] [VerP1, NouP1, pre, NouP0] [VerP1, NouP1, PreP] [VerP1, NouP1, Com] [VerP1, Com, Com] [S E N T E N C E]	[ver, pro, det, nou, pre, adj, adj, nou] [ver, NouP, det, nou, pre, adj, adj, nou] [ver, NouP, NouP, pre, adj, adj, nou] [ver, NouP, NouP, pre, NouP] [ver, NouP, NouP, PreP] [ver, NouP, NouP] [VerP] [S E N T E N C E]
4 **Give me the flights of the airline Tokened Airlines that serve breakfast.**	
AS-S en inglés	**Wolfram**
[ver, pro, art, nou, pre, art, nou, code, that, ver, nou] [ver, pro, art, nou, pre, art, nou, **NouP0V**, that, ver, nou] [ver, **NouP0**, art, nou, pre, art, nou, NouP0V, that, ver, nou] [ver, NouP0, art, **NouP0**, pre, art, nou, NouP0V, that, ver, nou] [ver, NouP0, art, NouP0, pre, art, **NouP0**, NouP0V, that, ver, nou] [ver, NouP0, art, NouP0, pre, art, NouP0, NouP0V, that, ver, **NouP0**] [ver, NouP0, **NouP1**, pre, art, NouP0, NouP0V, that, ver, NouP0] [ver, NouP0, NouP1, pre, **NouP1**, NouP0V, that, ver, NouP0]	[ver, pro, det, nou, pre, det, nou, ver, nou, Q_det, ver, nou] [ver, **NouP**, det, nou, pre, det, nou, ver, nou, Q_det, ver, nou] [ver, NouP, **NouP**, pre, det, nou, ver, nou, Q_det, ver, nou] [ver, NouP, NouP, pre, **NouP**, ver, nou, Q_det, ver, nou] [ver, NouP, NouP, pre, NouP, ver, **NouP**, Q_det, ver, nou] [ver, NouP, NouP, pre, NouP, ver, NouP, **QNouP**, ver, nou] [ver, NouP, NouP, pre, NouP, ver, NouP, QNouP, ver, **NouP**] [ver, NouP, NouP, **PreP**, ver, NouP, QNouP, ver, NouP]

[**VerP1**, NouP1, pre, NouP1, NouP0V, that, ver, NouP0] [VerP1, NouP1, pre, NouP1, NouP0V, that, **VerP1**] [VerP1, NouP1, pre, **NouP1V**, that, VerP1] [VerP1, NouP1, **PrePV**, that, VerP1] [VerP1, NouP1, **PrePV**] [VerP1, NouP1, **Com**] [VerP1, **Com**, Com] [**S E N T E N C E**]	[ver, NouP, NouP, PreP, ver, NouP, QNouP, **VerP**] [ver, NouP, **NouP**, ver, NouP, QNouP, VerP] [ver, NouP, NouP, ver, NouP, **CLAUSE**] [**VerP**, ver, NouP, CLAUSE] [VerP, ver, **NouP**] [**CLAUSE**, ver, NouP] [CLAUSE, **VerP**] [**S E N T E N C E]**
5	**Show me all the flights where breakfast is served in wide-body airplanes.**
AS-S en inglés	**Wolfram**
[ver, pro, ind_pro, art, nou, adv, nou, ver, pre, adj_cal, nou] [ver, **NouP0**, ind_pro, art, nou, adv, nou, ver, pre, adj_cal, nou] [ver, NouP0, ind_pro, art, **NouP0**, adv, nou, ver, pre, adj_cal, nou] [ver, NouP0, ind_pro, art, NouP0, adv, **NouP0**, ver, pre, adj_cal, nou] [ver, NouP0, ind_pro, art, NouP0, adv, NouP0, ver, pre, **NouP0**] [ver, NouP0, **NouP1**, adv, NouP0, ver, pre, NouP0] [ver, NouP0, NouP1, **NouP1**, ver, pre, NouP0] [**VerP1**, NouP1, NouP1, ver, pre, NouP0] [VerP1, NouP1, NouP1, ver, **PreP**] [VerP1, NouP1, NouP1, **VerP1**, PreP] [VerP1, NouP1, NouP1, VerP1, **Com**] [VerP1, **Com**, NouP1, VerP1, Com] [VerP1, Com, **Com**, VerP1, Com] [**S E N T E N C E**]	[ver, pro, det, det, nou, QAdv, nou, ver, ver, pre, adj, nou] [ver, **NouP**, det, det, nou, QAdv, nou, ver, ver, pre, adj, nou] [ver, NouP, **NouP**, QAdv, nou, ver, ver, pre, adj, nou] [ver, NouP, NouP, **QAdvP**, nou, ver, ver, pre, adj, nou] [ver, NouP, NouP, QAdvP, **NouP**, ver, ver, pre, adj, nou] [ver, NouP, NouP, QAdvP, NouP, ver, ver, pre, **NouP**] [ver, **CLAUSE**, QAdvP, NouP, ver, ver, pre, NouP] [ver, CLAUSE, QAdvP, NouP, ver, ver, **PreP**] [ver, CLAUSE, QAdvP, NouP, ver, **VerP**] [ver, CLAUSE, QAdvP, NouP, **VerP**] [ver, CLAUSE, **CLAUSE**] [**VerP**] [**S E N T E N C E**]
6	**Show me all the wide-body flights from Dallas to Denver that serve lunch.**
AS-S en inglés	**Wolfram**
[ver, pro, ind_pro, art, adj_cal, nou, pre, name, pre, name, that, ver, nou] [ver, pro, ind_pro, art, adj_cal, nou, pre, **NouP0V**, pre, name, that, ver, nou] [ver, pro, ind_pro, art, adj_cal, nou, pre, NouP0V, pre, **NouP0V**, that, ver, nou] [ver, **NouP0**, ind_pro, art, adj_cal, nou, pre, NouP0V, pre, NouP0V, that, ver, nou] [ver, NouP0, ind_pro, art, **NouP0**, pre, NouP0V, pre, NouP0V, that, ver, nou] [ver, NouP0, ind_pro, art, NouP0, pre, NouP0V, pre, NouP0V, that, ver, **NouP0**] [ver, NouP0, **NouP1**, pre, NouP0V, pre, NouP0V, that, ver, NouP0] [**VerP1**, NouP1, pre, NouP0V, pre, NouP0V, that, ver, NouP0] [VerP1, NouP1, pre, NouP0V, pre, NouP0V, that, **VerP1**] [VerP1, NouP1, pre, **NouP1V**, pre, NouP0V, that, VerP1] [VerP1, NouP1, pre, NouP1V, pre, **NouP1V**, that, VerP1] [VerP1, NouP1, **PrePV**, pre, NouP1V, that, VerP1] [VerP1, NouP1, PrePV, **PrePV**, that, VerP1] [VerP1, NouP1, PrePV, **PrePV**] [VerP1, NouP1, **Com**, PrePV] [VerP1, NouP1, Com, **Com**] [VerP1, **Com**, Com, Com] [**S E N T E N C E**]	[ver, pro, adv, det, adj, nou, pre, PN, pre, PN, Q_det, ver, nou] [ver, **NouP**, adv, det, adj, nou, pre, PN, pre, PN, Q_det, ver, nou] [ver, NouP, adv, **NouP**, pre, PN, pre, PN, Q_det, ver, nou] [ver, NouP, adv, NouP, pre, **NouP**, pre, PN, Q_det, ver, nou] [ver, NouP, adv, NouP, pre, NouP, pre, **NouP**, Q_det, ver, nou] [ver, NouP, adv, NouP, pre, NouP, pre, NouP, **QNouP**, ver, nou] [ver, NouP, adv, NouP, pre, NouP, pre, NouP, QNouP, ver, **NouP**] [ver, NouP, adv, NouP, pre, NouP, pre, NouP, QNouP, **VerP**] [ver, NouP, adv, NouP, **PreP**, pre, NouP, QNouP, VerP] [ver, NouP, adv, NouP, PreP, pre, NouP, **CALUSE**] [ver, NouP, **AdvP**, pre, NouP, CALUSE] [ver, NouP, AdvP, pre, **NouP**] [ver, NouP, AdvP, **PreP**] [**VerP**] [**S E N T E N C E**]
7	**Which aircraft types are not wide-body and are not pressurized?**
AS-S en inglés	**Wolfram**
[adj_int, nou, ver, adv, adj_cal, con_cop, ver, adv, adj_cal] [adj_int, **NouP0**, ver, adv, adj_cal, con_cop, ver, adv, adj_cal] [adj_int, NouP0, ver, adv, **SAdj0**, con_cop, ver, adv, adj_cal] [adj_int, NouP0, ver, adv, SAdj0, con_cop, ver, adv, **SAdj0**] [adj_int, NouP0, **VerP1**, con_cop, ver, adv, SAdj0] [adj_int, NouP0, VerP1, con_cop, **VerP1**] [adj_int, NouP0, **VerP3**] [Suj, **VerP3**] [**S E N T E N C E**]	[Q_det, adj, nou, ver, adv, adj, con, ver, adv, ver] [**QNouP**, ver, adv, adj, con, ver, adv, ver] [QNouP, ver, adv, **AdjP**, con, ver, adv, ver] [QNouP, ver, adv, AdjP, con, ver, adv, **VerP**] [**QNouP**, VerP, con, ver, adv, VerP] [QNouP, **VerP**, con, VerP] [QNouP, **VerP**] [**C L A U S E**]
8	**Which airlines have economy flights with discount from San Francisco to Dallas?**
AS-S en inglés	**Wolfram**
[adj_int, nou, ver, adj_cal, nou, pre, nou, pre, name, pre, name] [adj_int, nou, ver, adj_cal, nou, pre, nou, pre, **NouP0V**, pre, name] [adj_int, nou, ver, adj_cal, nou, pre, nou, pre, NouP0V, pre, **NouP0V**] [adj_int, **NouP0**, ver, adj_cal, nou, pre, nou, pre, NouP0V, pre, NouP0V] [adj_int, NouP0, ver, **NouP0**, pre, nou, pre, NouP0V, pre, NouP0V] [adj_int, NouP0, ver, NouP0, pre, **NouP0**, pre, NouP0V, pre, NouP0V] [adj_int, NouP0, **VerP1**, pre, NouP0, pre, NouP0V, pre, NouP0V] [adj_int, NouP0, VerP1, pre, NouP0, pre, **NouP1V**, pre, NouP0V] [adj_int, NouP0, VerP1, pre, NouP0, pre, NouP1V, pre, **NouP1V**] [adj_int, NouP0, VerP1, **PreP**, pre, NouP1V, pre, NouP1V] [adj_int, NouP0, VerP1, PreP, **PrePV**, pre, NouP1V] [adj_int, NouP0, VerP1, PreP, PrePV, **PrePV**] [**Suj**, VerP1, PreP, PrePV, PrePV] [Suj, VerP1, **Com**, PrePV, PrePV] [Suj, VerP1, Com, **Com**, PrePV] [Suj, VerP1, Com, Com, **Com**] [**S E N T E N C E**]	[Q_det, nou, ver, nou, nou, pre, nou, pre, PN, PN, pre, PN] [**QNouP**, ver, nou, nou, pre, nou, pre, PN, PN, pre, PN] [QNouP, ver, **NouP**, pre, nou, pre, PN, PN, pre, PN] [QNouP, ver, NouP, pre, **NouP**, pre, PN, PN, pre, PN] [QNouP, ver, NouP, pre, NouP, pre, **NouP**, pre, PN] [QNouP, ver, NouP, pre, NouP, pre, NouP, pre, **NouP**] [QNouP, ver, NouP, PreP, pre, NouP, pre, **NouP**] [QNouP, ver, NouP, PreP, pre, NouP, **PreP**] [QNouP, ver, NouP, PreP, pre, **NouP**] [QNouP, ver, NouP, PreP, **PreP**] [QNouP, **VerP**] [**C L A U S E**]

9	**Which airlines have flights that serve lunch in wide-body jets?**
AS-S en inglés	**Wolfram**
[adj_int, nou, ver, nou, that, ver, nou, pre, adj_cal, nou] [adj_int, **NouP0**, ver, nou, that, ver, nou, pre, adj_cal, nou] [adj_int, NouP0, ver, **NouP0**, that, ver, nou, pre, adj_cal, nou] [adj_int, NouP0, ver, NouP0, that, ver, **NouP0**, pre, adj_cal, nou] [adj_int, NouP0, ver, NouP0, that, ver, NouP0, pre, **NouP0**] [adj_int, NouP0, **VerP1**, that, ver, NouP0, pre, NouP0] [adj_int, NouP0, VerP1, that, **VerP1**, pre, NouP0] [adj_int, NouP0, VerP1, that, VerP1, **PreP**] [adj_int, NouP0, **VerP3**, PreP] [**Suj**, VerP3, PreP] [Suj, VerP3, **Com**] [**S E N T E N C E**]	[Q_det, nou, ver, nou, Q_det, ver, nou, pre, adj, nou] [**QNouP**, ver, nou, Q_det, ver, nou, pre, adj, nou] [QNouP, ver, **NouP**, Q_det, ver, nou, pre, adj, nou] [QNouP, ver, NouP, **QNouP**, ver, nou, pre, adj, nou] [QNouP, ver, NouP, QNouP, ver, **NouP**, pre, adj, nou] [QNouP, ver, NouP, QNouP, ver, NouP, pre, **NouP**] [QNouP, ver, NouP, QNouP, ver, NouP, **PreP**] [QNouP, ver, NouP, QNouP, ver, **NouP**] [QNouP, ver, NouP, QNouP, **VerP**] [QNouP, ver, NouP, **CLAUSE**] [QNouP, ver, **NouP**] [QNouP, **VerP**] [**C L A S U E**]
10	**Which flights have no discounted fare.**
AS-S en inglés	**Wolfram**
[adj_int, nou, ver, qua_adj, ver, nou] [adj_int, **NouP0**, ver, qua_adj, ver, nou] [adj_int, NouP0, ver, qua_adj, ver, **NouP0**] [adj_int, NouP0, ver, qua_adj, **VerP1**] [adj_int, NouP0, **VerP1**] [**Suj**, VerP1] [**S E N T E N C E**]	[adj, nou, ver, det, adj, nou] [**NouP**, ver, det, adj, nou] [NouP, ver, **NouP**] [NouP, **VerP**] [**S E N T E N C E**]

REFERENCIAS

[Aguirre, 2014] M.A. Aguirre, *Modelo Semánticamente Enriquecido de Bases de Datos para su Explotación por Interfaces de Lenguaje Natural*, tesis doctoral, Div. de Estudios de Posgrado e Investugación, Inst. Tecnológico de Tijuana, Tijuana, México, jul. 2014.

[Androutsopoulos, 1995] I. Androutsopoulos, G.D. Ritchie y P. Thanisch, "Natural Language Interfaces to Databases – An Introduction", *Journal of Natural Language Engineering*, vol. 1, no. 1, pp. 29–481, 1995.

[Atserias, 2006] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró y M. Padró, "FreeLing 1.3: Syntactic and Semantic Services in an Open-source NLP Library", *Proc. 5th International Conference on Language Resources and Evaluation*, pp. 48–55, ene. 2006.

[Auxerre, 1986] P. Auxerre y R. Inder, *MASQUE Modular Answering system for queries in english – user's manual*, Technical Report AIAI/SR/10, Artificial Intelligence Applications Institute, University of Edinburgh, jun. 1986.

[Azcoaga, 1989] J.E. Azcoaga, "Información Semántica, Distancias Semánticas y Conceptos", *Revista Interuniversitaria de Formación del Profesorado*, no. 4, pp.71–78, 1989.

[Bais, 2016] H. Bais, M. Machjour y L. Koutti, "Querying Database Using a Universal Natural Language Interface Based on Machine Learning", *Proc. International Conference on Information Technology for Organizations Development*, pp. 1–6, mar. 2016.

[Ballard, 1986] B.W. Ballard y D.E. Stumberger, "Semantic Acquisition in TELI: A Transportable, User-customized Natural Language Processor", *Proc. 24th Annual Meeting of ACL*, pp. 20–29, jul. 1986.

[Borland International, 1988] Borland International y S. Val, *Turbo Prolog 2.0 Reference Guide*, ene. 1988.

[Brants, 2000] T. Brants, "TnT – A statistical Part-of-speech Tagger", *Proc. 6th Applied Natural Language Processing Conference*, pp. 224–231, apr. 2000.

[Brinton, 2010] L.J. Brinton y D.M. Brinton, *The Linguistic Structure of Modern English*, 2da edición, John Benjamins Publishing Company, 2010.

[Brock, 2016] A. Brock, "Who Makes Grammar Rules?", VOA Learning English, https://learningenglish.voanews.com/a/who-makes-grammar-rules/ 3325780.html, may. 2016.

[Cardiff School of Computer Science & Informatics, 1997] Cardiff School of Computer Science & Informatics, "Regular Expressions", http://users.cs.cf.ac.uk/Dave.Marshall/Internet/NEWS/regexp.html, 1997.

[Celce-Murcia, 1999a] M. Celce-Murcia, D. Larsen-Freeman y H. Williams, *The Grammar Book*, 2da edición, Heinle & Heinle, E.U.A., https://flaviamcunha.files.wordpress.com/2013/03/the-grammar-book-an-eslefl-teachers-course-second-editiona4.pdf, 1999.

[Celce-Murcia, 1999b] M. Celce-Murcia, D. Larsen-Freeman y H. Williams, *The Grammar Book*, 2da edición, Heinle & Heinle, E.U.A., página 104, https://flaviamcunha.files.wordpress.com/2013/03/the-grammar-book-an-eslefl-teachers-course-second-editiona4.pdf, 1999.

[Cellan-Jones, 2009] R. Cellan-Jones, "Wolfram 'search engine' goes live", BBC News, http://news.bbc.co.uk/2/hi/technology/8052798.stm, may. 2009.

[Cervantes, 2005] J.A. Cervantes, *Analizador Sintáctico de Oraciones en Español Usando el Método de Dependencias*, tesis de maestría, Depto de Ciencias Computacionales, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, México, jul. 2005.

[Cohen, 2004] K.B. Cohen y L. Hunter, "Natural Language Processing and Systems Biology", *Artificial Intelligence Methods and Tools for Systems Biology*, Springer, 2004, pp. 147–173.

[Date, 1974] C.J. Date y E.F. Codd, "The Relational and Network Approaches: Comparison of the Application Programming Interfaces", *Proc. 1974 ACM-SIGMOD Workshop on Data Description, Access, and Control*, ACM, pp 83–113, may. 1974.

[Date, 2003] C.J. Date, *An introduction to Database Systems*, 8va edición, Pearson Education, Inc., 2003.

[Davies, 2008] M. Davies, "The COCA Corpus", https://www.english-corpora.org/coca/help/coca2020_overview.pdf, mar. 2008.

[Davies, 2009] M. Davies, "The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights", *International Journal of Corpus Linguistics*, vol.14, no.2, DOI: 10.1075/ijcl.14.2.02dav, pp. 159–190, ene. 2009.

[Davies, 2010] M. Davies, "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English", *Literary and Linguistic Computing*, vol. 25, no. 4, pp 447–464, dic. 2010.

[Faili, 2009] H. Faili, "From Partial toward Full Parsing", *Proc. International Conference RANLP-2009*, Association for Computational Linguistics, pp. 71–75, 2009.

[Farlex, 2020] Farlex Inc, "The Free Dictionary", consulta "grammatical rule", Princeton University, https://www.thefreedictionary.com/grammatical+rule, consultado en 2020.

[Finegan, 2008] E. Finegan, *Language: Its Structure and Use*, 5ta edición, Boston, MA, 2008.

[Floridi, 2015] L. Floridi, "Concepciones Semánticas de la Información", *Diccionario Interdisciplinar Austral*, http://dia.austral.edu.ar/Concepciones_semánticas_de_la_información, 2015.

[Ford, 2004] B. Ford, "Parsing Expression Grammars: A Recognition Based Syntactic Foundation", *Proc. 31st ACM SIGPLAN Notices*, vol. 39, no 1, pp. 111–122, ene. 2004.

[Garside, 1987] R. Garside, "The CLAWS Word-tagging System", *The Computational Analysis of English: A Corpus-based Approach*, Longman, London, http://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf, pp. 30–41, 1987.

[González, 2005] J.J. González, *Traductor de Lenguaje Natural Español a SQL para un Sistema de Consultas a Bases de Datos*, tesis doctoral, Depto. de Ciencias Computacionales, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, México, dic. 2005.

[Green, 1961] B.F. Green, A.K. Wolf, C. Chomsky y K. Laughery, “Baseball: An Automatic Question-Answerer”, *Proc. Western Joint IRE-AIEE-ACM Computer Conference*, pp. 219–224, may. 1961.

[Grosz, 1983] B. Grosz, “TEAM: A Transportable Natural-Language Interface System”, *Proc. Conference on Applied Natural Language Processing*, pp. 39–45, feb. 1983.

[Hafner, 1984] C.D. Hafner, “Interaction of Knowledge Sources in a Portable Natural Language Interface”, *Proc. 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics*, pp. 57–60, jul. 1984.

[Halliday, 2006] M.A.K. Halliday, *On Language and Linguistics*, *Vol. 3*, Continuum International Publishing Group, mar. 2006.

[Hardeniya, 2015] N. Hardeniya, *NLTK Essentials: Build Cool NLP and Machine Learning Applications Using NLTK and Other Python Libraries*, Packt Publishing, jul. 2015.

[Hendrix, 1977] G.G. Hendrix, E.D. Scaerdoti, D. Sagalowicz y J. Slocum, “Developping a Natural Language Interface to Complex Data”, SRI International, Menlo Park, CA, EUA, Nota técnica 152, ago. 1977.

[Hemphill, 1990] C.T. Hemphill, J.J. Godfrey y G.R. Doddington, “The ATIS Spoken Language Systems Pilot Corpus”, *Proc. Speech and Natural Language*, pp. 96–101, https://aclanthology.org/H90-1021.pdf, jun. 1990

[Herring, 2016] P. Herring. *The Farlex Grammar Book: Complete English Grammar Rules*, Farlex Intenational, 2016.

[Jurafsky, 1996] D. Jurafsky, “A Probabilistic Model of Lexical and Syntactic Access and Disambiguation”, *Cognitive Science*, vol. 20, no. 2, pp. 137–194, abr. 1996.

[Kokare, 2014] R.B. Kokare y K.H. Wanjale, “A Survey of Natural Language Query Builder Interface for Structured Databases Using Dependency Parsing”, *International Journal of Computer Applications*, vol. 107, no. 5, pp. 9–14, dic. 2014.

[Krishnamurthy, 2014] J. Krishnamurthy y T.M. Mitchell, “Joint Syntactic and Semantic Parsing with Combinatory Categorial Grammar”, *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1188–1198, jun. 2014.

[Kroeger, 2019] P.R. Kroeger, *Analyzing Meaning: An Introduction to Semantics and Pragmatics*, 2da edición, Language Science Press, 2019.

[Lewis, 1997] M. Lewis, *Implementing the Lexical Approach: Putting Theory into Practice*, 1ra edición, Heinle ELT, ene. 1997.

[Li, 2005] Y. Li, H. Yang y H.V. Jagadish, “Nalix: An Interactive Natural Language Interface for Querying XML”, *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 900–902, jun. 2005.

[Li, 2014] F. Li y H. Jagadish, “NaLIR: An Interactive Natural Language Interface for Querying Relational Databases”, *Proc. 2014 ACM SIGMOD International Conference on Management of Data*, pp. 709–712, jun. 2014.

[Linguistic Data Consortium, 1990]
Linguistic Data Consortium, *The 2884 ATIS0 Speaker-dependent Training Prompts*, No. de catálogo LDC93S4B-3, https://catalog.ldc.upenn.edu/docs/LDC93S4B-3/TRN_PRMP.TXT, 1990.

[Lyons, 1991] J. Lyons, *Natural Language and Universal Grammar: Volume 1: Essays in Linguistic Theory, (African Studies Series)*, Cambridge University Press, 1991.

[Machine Learning Research Group, 2004]
Machine Learning Research Group, "Semantic Mapping Corpora", descargar Geoquery250 en la liga "GeoQueries250 zip file", http://disi.unitn.it/~agiordani/corpora.htm, 2004.

[Macmillan Dictionary, 2021]
Macmillan Dictionary, https://www.macmillandictionary.com, consultado en 2021.

[Majhadi, 2021] K. Majhadi y M. Machkour, "The History and Recent Advances of Natural Language Interfaces for Databases Querying", *Proc. 3rd International Conference of Computer Science and Renewable Energies*, vol. 229, pp. 1–7, ene. 2021.

[Marimon, 2018] M. Marimon, L. Padró y J. Turmo. "Coreference Resolution in FreeLing 4.0", *Proc. 11th International Conference on Language Resources and Evaluation*, European Language Resources Association, pp. 376–381, https://aclanthology.org/L18-1057.pdf, may. 2018.

[Martínez de Sousa, 1995]
J.M. Souza, *Diccionario de Lexicografia Práctica*, Bibliograf, Barcelona, España, 1995.

[Mellado, 2014] O.M. Mellado, *Implementación de un Analizador Sintáctico del Idioma Español para una Interfaz de Lenguaje Natural a Bases de Datos*, tesis de maestría, Div. de Estudios de Posgrado e Investigación, Inst. Tecnológico de Ciudad Madero, Cd. Madero, México, http://200.188.131.162:8080/jspui/handle/123456789/205, may. 2014.

[Moore, 1981] R.C. Moore, "Problems in Logical Form", *Proc. 19th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 117–124, jun. 1981.

[Newman, 2020] J. Newman y C. Cox, "Corpus Annotation", *A Practical Handbook of Corpus Linguistics*, Springer, pp. 25–48, 2020.

[OAG, 1990] Official Airline Guides, *Official Airline Guide: With Fares, North American Edition with Fares*, Oakbrook, IL, vol. 16, No. 7, ene. 1990.

[Oracle, 2011] Oracle. "ATG Search Query Reference Guide", https://docs.oracle.com/cd/E23507_01/Search.20073/ATGSearchQueryRef/html/s0102audience01.html, 2011.

[Oxford, 2020a] "Diccionario de Inglés y Español, Sinónimos y Traductor de Espñol a Inglés", consulta "formal language", https://www.lexico.com/en/definition/formal_language, consultado en 2020.

[Oxford, 2020b] "Diccionario de Inglés y Español, Sinónimos y Traductor de Espñol a Inglés", consulta "natural language", https://www.lexico.com/en/definition/natural_language, consultado en 2020.

[Oxford, 2020c] "Diccionario de Inglés y Español, Sinónimos y Traductor de Espñol a Inglés", Nota: obtiene la explicación de "Word classes (or parts of the speech)", https://www.lexico.com/grammar/word-classes-or-parts-of-speech, consultado en 2020.

[Oxford, 2021a] "Diccionario de Inglés y Español, Sinónimos y Traductor de Español a Inglés", consulta "syntax", https://www.lexico.com/en/definition/syntax, consultado en 2021.

[Padró, 1997] L. Padró, *A Hybrid Environment for Syntax–Semantic Tagging*, tesis doctoral, Depto. de Lenguas y Sistemas Infromáticos, Universidad Politécnica de Cataluña, http://www.tdx.cat/bitstream/handle/10803/6643/01Lpc01de01.pdf?sequence=1&isAllowed=y, dic. 1997.

[Padró, 2011] L. Padró, "Analizadores Multilingües en FreeLing", *Linguamática*, vol.3, no. 3, pp. 13–20, ene. 2011.

[Pazos, 2013] R.A. Pazos R., J.J. González B., M.A. Aguirre L., J. M. Martínez F. y H.J. Fraire H. "Natural Language Interfaces to Databases: An Analysis of the State of the Art", *Recent Advances on Hybrid Intelligent System*s, vol. 451, Springer, Berlin, Heidelberg, Alemania., 2013, pp. 463–480.

[Pazos, 2021] R.A. Pazos R., J.M. Martínez F., J. Gaspar H., G. Rivera R. y Florencia-Juárez, "Natural Language Interfaces to Databases: Survey on Recent Advances", *Handbook of Research on Natural Language Processing and Smart Service Systems*, pp. 1–30, IGI Global, 2021.

[PCMag, 2020] PCMag, "PCMag Encyclopedia", consulta "natural language query", https://www.pcmag.com/encyclopedia/term/natural-language-query, consultada en 2020.

[Popescu, 2004] A. Popescu, A. Armanasu, O. Etzioni, D. Ko y A. Yates, "Modern Natural Language Interfaces to Databases: Composing Statistical Parsing with Semantic Tractability", *Proc. 20th International Conference on Computational Linguistics*, pp. 141–147, ago 2004.

[Quirk, 1985] R. Quirk, S. Greenbaum, G. Leech y J. Svartvik, *A Comprehensive Grammar of the English Language*, 2da edición, Longman, may. 1985.

[RAE, 2021] "Diccionario de la Lengua Española", consulta "oración", https://dle.rae.es/oraci%C3%B3n?m=form, consultado en 2021.

[Rayson, 1998] P. Rayson y R. Garside. "The CLAWS Web Tagger", *ICAME Journal*, vol. 22, pp. 121–123, https://www.researchgate.net/publication/2618590_The_CLAWS_Web_Tagger/link/00b495187f76917eef000000/download, jun. 1998.

[Saiz, 2003] M. Saiz, "Procesamiento del Lenguaje Natural: Presente y Perspectivas Futuras", *Memoria de Ingeniería de Software en la Década del 2000*, Cartagena de Indias, Colombia, pp. 196–198, ago. 2003.

[Sapir, 1954] E. Sapir, *El Lenguaje: Introducción al Estudio del Habla*, Fondo de Cultura Económica, México, 1954.

[Sujatha, 2016] B. Sujatha y S.V. Raju, "Ontology Based Natural Language Interface for Relational Databases", *Procedia Computer Science*, vol. 92, pp. 487–492, 2016.

[Templeton, 1983] M. Templeton y J. Burger, "Problems in Natural-Language Interface to DBMS with Examples from EUFID", *Proc. 1st Conference on Applied Natural Language Processing*, pp. 3–16, feb. 1983.

[The Boston Language Institute, 2013] The Boston Language Institute, "Language Regulation", https://bostonlanguage.wordpress.com/2013/02/19/language-regulation/, feb. 2013.

[Vanderwende, 2015] L. Vanderwende, *NLPwin – An Introduction*, Microsoft Research, Reporte técnico No. MSR-TR-2015-23, mar. 2015.

[Verástegui, 2020] A.A. Verástegui, *Tratamiento de los Problemas de Valores de Búsqueda de Difícil Detección en la Traducción de Consultas de Lenguaje Natural a SQL*, tesis doctoral, Div. de Estudios de Posgrado e Investigación, Inst. Tecnológico de Tijuana, Tijuana, México, 2020.

[Waltz, 1978] D.L. Waltz, "An English Language Question Answering System for a Large Relational Database", *Communications of the ACM*, vol. 21, no. 7, pp. 526–539, jul. 1978.

[Warren, 1982] D.H.D. Warren y F.C.N. Pereira, "An Efficient Easily Adaptable System for Interpreting Natural Language Queries", *American Journal of Computational Linguistics*, vol. 8, no. 3–4, pp. 110–122, https://aclanthology.org/J82-3002, dic. 1982.

[Wolfram, 2019] Wolfram Alpha LLC, "TextStructure", https://reference.wolfram.com/language/ref/TextStructure.html, 2019.

[Wolfram, 2021] Wolfram Alpha LLC, "What is Wolfram|Alpha?", https://www.wolframalpha.com/tour/, consultado en 2021.

[Woods, 1972] W.A. Woods, R.M. Kaplan y B. Nash-Webber, "The Lunar Sciences Natural Language Information System: Final Report", Bolt Beranek and Newman, Inc., Cambridge, MA, https://www.researchgate.net/publication/247926251_The_Lunar_Science_Natural_Language_Information_System_Final_Report, jun. 1972.

[World Wide Web Consortium, 1982] World Wide Web Consortium, "Notation: A. BNF Notation for Syntax", https://www.w3.org/Notation.html, ago. 1982.

[Zhong, 2017] V. Zhong, C. Xiong y R. Socher, "Seq2sql: Generating Structured Queries from Natural Language Using Reinforcement Learning", arXiv.org, Cornell University, New York, NY, http://arxiv.org/pdf/1709.00103.pdf, nov. 2017.