INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

Laboratorio de Procesamiento de Lenguaje Natural

Extracción automática de información semántica basada en estructuras sintácticas

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA

HONORATO AGUILAR GALICIA

DIRECTORES

Dr. Grigori Sidorov

Dra. Yulia Nikolaevna Ledeneva

México, D.F.

Diciembre de 2012

SIP-14-Bis

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 12:00 horas del día 28 del mes de noviembre de 2012 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Extracción automática de información semántica basada en estructuras sintácticas"

Presentada por el alumno:

AGUILAR	**GALICIA**	**HONORATO**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	1	0	1	6	2	9

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de Tesis

Dr. Grigori Sidorov

Dra. Yulia Nikolaevna Ledeneva

Dr. Sergio Suárez Guerra

Dr. Alexander Gelbukh

Dr. Miguel Jesús Torres Ruiz

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 28 del mes de noviembre del año 2012, el que suscribe *Honorato Aguilar Galicia* alumno del *Programa de Maestría en Ciencias de la Computación* con número de registro *B101629*, adscrito al *Centro de Investigación en Computación*, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del *Dr. Grigori Sidorov* y la *Dra. Yulia Nikolaevna Ledeneva* y cede los derechos del trabajo intitulado *EXTRACCIÓN AUTOMÁTICA DE INFORMACIÓN SEMÁNTICA BASADA EN ESTRUCTURAS SINTÁCTICAS*, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección *aguilargh@hotmail.com*. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Honorato Aguilar Galicia

RESUMEN

La información en un texto se conforma de párrafos, cada párrafo por un conjunto de oraciones y cada oración por unidades de texto más pequeñas. Estas pequeñas unidades, con información semántica propia se llaman hechos y se pueden obtener a través de la descomposición de la oración en una colección de hechos. Cada hecho tiene información independiente y puede ser utilizado como una unidad independiente.

Los hechos pueden ser utilizados por otras tareas de Procesamiento Automático de Lenguaje Natural, como llenar bases de conocimiento, crear resúmenes automáticos, desarrollar sistemas de pregunta-respuesta y evaluar la calidad de contenido de un documento mediante el número de hechos encontrados en él en relación a su longitud. Además, lograr que la computadora guarde conocimiento y no sólo textos.

En la presente investigación se desarrolla un sistema para extraer hechos de forma automática empleando árboles de dependencias de las oraciones. Al sistema se le proporciona un conjunto de oraciones y devuelve los hechos de cada una de ellas. Los hechos se muestran en la interfaz del sistema y se guardan en una base de datos.

Para el desarrollo del sistema, se estudian patrones sintácticos en los árboles de dependencias, que identifiquen a los hechos. Con base en estos patrones sintácticos se desarrollan algoritmos heurísticos para extraer hechos. La investigación está orientada a textos en el idioma español, pero el método se puede aplicar a cualquier idioma en el que se puedan construir árboles sintácticos.

El sistema obtiene un alto desempeño: una precisión de 87% y un recall de 91%, lo que es mejor comparado con otros sistemas. Para la evaluación se diseñó un corpus de 68 oraciones, en donde 166 hechos son extraídos manualmente.

ABSTRACT

The information in a text is organized in paragraphs, each paragraph is represented by a set of sentences and each sentence consists of smaller text units. These small units, that have semantic nature, are called facts and can be obtained by decomposition of sentence. Each fact contains independent information and can be used as a standalone unit.

Facts can be used for other tasks of Automatic Natural Language Processing, such as filling knowledge bases, creating automatic summaries, developing question-answering systems and evaluating the quality of content of a document by considering the number of facts found in it as related to its length. In addition, computer stores the knowledge and not just texts.

In this research we develop a system for automatic extraction of facts using dependency trees that correspond to sentences. The system is given a set of sentences and returns the facts of each one of them. The events are displayed in the interface of the system and stored in a database.

For the development of the system, we studied syntactic patterns in dependency trees, identifying the facts. Based on these syntactic patterns, we developed heuristic algorithms for fact extractions. The research is aimed at Spanish language texts, but the method can be applied to any language, where we can construct syntactic trees.

The system obtained high performance: precision 87%, recall 91% that is better than comparable systems. For evaluation we designed a corpus of 68 sentences, where 166 facts are marked manually.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

Como capítulo inicial de la tesis, aquí se ubica el área al que pertenece la presente investigación, después se describe el planteamiento del problema, la justificación, los objetivos y como está organizada la tesis.

1.1 Ubicación

Uno de los recursos más importantes que posee la humanidad es la información, la cual se guarda primordialmente en forma de lenguaje natural (como español, inglés, ruso o algún otro idioma) en libros, revistas, periódicos u otros textos; esencialmente en formato digital.

Al ser un recurso importante de donde se pueden obtener múltiples beneficios, se buscan maneras de aprovechar esta información almacenada. Y es a través de "*Procesamiento Automático de Lenguaje Natural* (PLN), un área formada por la intersección e interacción de la lingüística y la computación" (Gelbukh & Sidorov, 2010), como se logran procesar grandes volúmenes de texto, por su sentido, consiguiendo agruparlos según la información contenida en ellos o de extraerles información útil.

Actualmente hay dos áreas principales de procesamiento inteligente de texto en PLN: Recuperación de Información y Extracción de Información. La "*Recuperación de Información* (Information Retrieval (IR), en inglés) consiste en seleccionar automáticamente, en una determinada colección de documentos, normalmente muy grande, aquellos que se ajustan a una pregunta del usuario" (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003). A modo de ejemplos de sistemas IR se tienen a los motores de búsqueda Google, Yahoo, y Bing.

La "*Extracción de Información* (Information Extraction (IE), en inglés) consiste en obtener información de forma selectiva de un documento (Quién hizo qué, cuándo, cómo, etc). Para ello se definen unas plantillas o esquemas correspondientes al dominio que se desea tratar, que deben ser rellenadas para cada documento" (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003). La IE busca entidades en el texto sobre categorías predefinidas como nombres de personas, organizaciones, lugares, cantidades, valores

monetarios, porcentajes, expresiones de hora; y relaciones entre ellas u otros tópicos específicos dentro de los textos.

Un paradigma reciente dentro de la IE es la *Extracción de Información Abierta* (Open Information Extraction (OIE), en inglés) que consiste en obtener información de forma selectiva, pero a diferencia de la IE tradicional, esta no maneja categorías de entidades predefinidas, además el dominio es independiente y la extracción de entidades y relaciones es escalable (Jain & Pennacchiotti, 2010). Se puede observar un avance respecto a la IE tradicional. Un ejemplo puede consultarse en el sitio web de la Universidad (Turing Center University of Washington).

Otro paradigma, también reciente, dentro de la IE es la Extracción de Hechos (Fact Extraction (FE), en inglés) cuya meta es extraer información semántica, es decir, datos bien conformados sintácticamente y significativos. El dominio es independiente, pero a diferencia de la OIE aquí no se necesitan escalar entidades y relaciones ya que en la FE se extrae toda la información semántica que existe en el texto.

Es en la FE donde se ubica la presente investigación y a la información semántica se le llama hechos. En la siguiente sección se define "hecho".

1.2 Planteamiento del problema

La información en un texto se conforma de párrafos, cada párrafo por un conjunto de oraciones y estas, por "unidades de texto más pequeñas que la oración, que se pueden obtener a través de la descomposición de la oración en una colección de frases. Cada frase tiene información independiente que puede ser usada como una unidad independiente" (Hovy, Zhou, & Kwon, 2007).

Estas frases se encuentran fusionadas en la oración para enunciar algo de manera más amplia, pero al separarse de la oración tienen sentido completo, es decir, tienen información semántica por ellas mismas. Una oración tiene sentido completo si contiene sujeto y predicado (Fuentes de la Corte, 2010).

De acuerdo a (Hovy, Zhou, & Kwon, 2007) y (Fuentes de la Corte, 2010), en la presente investigación estas frases que por ellas mismas contienen "información semántica", se les llama "hechos", y su definición formal es:

*Un **hecho** es la unidad mínima de texto que se puede extraer de una oración, tiene independencia semántica, únicamente un verbo y su forma es una tripleta conformada así*:

Hecho = [*Sujeto*] + [*Verbo*] + [*Objeto/Complemento*]

Por ejemplo en la oración: "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*", se pueden identificar los hechos que se muestran en la Tabla 1.1.

Tabla 1.1 Hechos identificados en la oración "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*".

No.	Sujeto	Verbo	Objeto/Complemento
1	La civilización China	heredó	el papel
2	La civilización China	heredó	la pólvora
3	La civilización China	heredó	una forma de imprenta rudimentaria
4	La civilización China	heredó	la brújula

Se puede observar que cada hecho tiene independencia semántica, es decir, ninguno necesita a otro para tener sentido completo o informar algo. Todos tienen un solo verbo, todos cumplen la tripleta que define hecho. Y que una oración puede tener varios hechos.

Así que la presente investigación consiste en desarrollar un método para identificar y extraer la información semántica o hechos que se encuentran fusionados en las oraciones, con base en análisis de estructuras sintácticas.

La extracción de información semántica o hechos, se hace desde un corpus formado por un conjunto de oraciones extraídas de libros de texto de educación primaria y secundaria, en el idioma español.

En adelante, a la "información semántica" que se extrae de las oraciones se le llamará también "hechos", utilizando los dos conceptos para referirse a lo mismo.

1.3 Justificación

La extracción de hechos de corpus textuales es una tarea que ayuda, proporcionándoles un producto, a varias tareas de LPN relacionadas con la comprensión del texto, tales como:

- Llenar bases de conocimiento.
- Sistemas de pregunta-respuesta (Question Answering, en inglés).
- Generación automática de resúmenes extractivos y abstractivos.
- Evaluar la calidad de contenido de un documento mediante el número de hechos encontrados en él en relación a su longitud –densidad de hechos– (Lex & Horn, 2012).

Además:

- Se pretende que la computadora adquiera conocimientos, leyendo libros y separando datos relevantes (hechos) que almacene en una base de conocimientos y posteriormente utilizarlos. Haciendo una analogía con las personas: las personas adquieren conocimientos leyendo libros y memorizando datos importantes para ellas y cuando son requeridos los utiliza.
- Lograr que la computadora almacene conocimiento y no sólo textos.

La razón de que la investigación sea con textos en español es porque no existen por el momento sistemas para extraer hechos desde textos en español. Y se han elegido libros de educación primaria y secundaria porque tienen áreas de estudio bien definidas, además por tener un objetivo didáctico contienen una gran cantidad de hechos.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un método y los recursos para la extracción automática de información semántica (hechos), desde un corpus de oraciones, con base en análisis de estructuras sintácticas que identifiquen la información semántica en las oraciones; y crear la aplicación de este método.

1.4.2 Objetivos específicos

- Elaborar un manual para identificar la información semántica o hechos en una oración.
- Crear un corpus de prueba, conformado de un conjunto de oraciones a las cuales dos humanos expertos identifica los hechos que contiene cada una oración.
- Identificar los patrones sintácticos que identifican la información semántica o hechos en una oración.
- Desarrollar el algoritmo basado en heurística para extraer la información semántica o hechos.
- Desarrollar una base de datos para guardar la información semántica o hechos.
- Desarrollar la aplicación.
- Evaluar los resultados obtenidos por el sistema.

1.5 Organización de la Tesis

El presente documento de tesis se compone de siete capítulos y cuatro anexos. El resto del documento se organiza de la siguiente manera:

Capítulo 2: Marco teórico. Aquí se describen: algunas tareas de lingüística computacional, conceptos sobre gramática, enfoques de análisis sintáctico, la herramienta FreeLing para análisis sintáctico automático, teoría acerca de heurísticas y corpus, la definición de hecho utilizada en la investigación.

Capítulo 3: Estado del arte. Revisa algunos trabajos relacionados con la extracción de información semántica, lo que en la presente investigación se llaman hechos.

Capítulo 4: Método propuesto. Presenta y explica: la arquitectura general del método propuesto; un diagrama y su descripción de los patrones sintácticos que identifican los hechos, además de un ejemplo; los algoritmos de las heurísticas desarrolladas con base a los patrones sintácticos y ejemplos de hechos extraídos con estos algoritmos.

Capítulo 5: Desarrollo del sistema. Describe la construcción del corpus que se utiliza como estándar de oro para evaluar al sistema; se explica el comando y sus parámetros para ejecutar FreeLing; luego se habla sobre el diseño de la base de datos para guardar los hechos; inmediatamente se comentan las herramientas utilizadas para el desarrollo del sistema, se muestra y describe un diagrama de bloques para explicar su funcionamiento, a continuación se muestra y describe su interfaz gráfica de forma detallada.

Capítulo 6: Evaluación y resultados. Explica cómo se evalúa el método propuesto, las métricas de evaluación, se presentan y describen los resultados obtenidos por el sistema, y una comparación con otros sistemas.

Capítulo 7: Conclusiones y trabajo futuro. Presenta las conclusiones, aportaciones de la investigación, trabajo futuro, presentaciones y publicaciones durante el desarrollo de la tesis.

Anexo A. Presenta el corpus de prueba, compuesto por un conjunto de oraciones. Se muestra el corpus como estándar de oro utilizado para la evaluación, donde cada oración tiene sus hechos extraídos por personas. Se muestran los hechos de cada oración obtenidos por el sistema. Se muestran ejemplos de árboles de dependencias.

Anexo B. Presenta la guía para extraer hechos de forma manual, utilizada para construir el estándar de oro.

Anexo C. Presenta las etiquetas sintácticas empleadas por FreeLing.

Anexo D. Presenta las etiquetas morfológicas empleadas por FreeLing.

2 MARCO TEÓRICO

Se describen algunas tareas de lingüística computacional, conceptos sobre gramática, enfoques de análisis sintáctico, la herramienta FreeLing para análisis sintáctico automático, teoría acerca de heurísticas y corpus, la definición de hecho utilizada en la investigación.

2.1 Tareas de lingüística computacional

2.1.1 Recuperación de información

La Recuperación de Información (Information Retrieval (IR), en inglés) puede ser definida como la aplicación de tecnología informática para la adquisición, organización, almacenamiento, recuperación y distribución de información (Jackson & Moulinier, 2007).

(Jackson & Moulinier, 2007) Explica que la IR se ocupa de las bases teóricas y la mejora práctica de la tecnología de motores de búsqueda, incluyendo la construcción y el mantenimiento de grandes repositorios de información. En años recientes, los investigadores han expandido sus preocupaciones desde la búsqueda bibliográfica y texto completo de repositorios de documentos a búsqueda en la Web, con su hipertexto asociado y bases de datos multimedia.

La IR es una actividad, como muchas actividades esta tiene un propósito. Un usuario de un motor de búsqueda comienza con una necesidad de información, que él o ella formula como una consulta con el objetivo de encontrar documentos relevantes. Esta consulta podría no ser la mejor formulada para indicar esa necesidad, quizás no está bien escrito, se hayan seleccionado mal las palabras, podría contener demasiadas palabras o insuficientes. Sin embargo es la única pista que el motor de búsqueda recibe para lograr su objetivo.

Un observación muy novedosa de (Jackson & Moulinier, 2007), dice que a menudo se habla de los documentos en el conjunto de resultados como más o menos relevante para la consulta, pero, en sentido estricto esto es incorrecto. Pues como bien dice, es el usuario quien juzga la relevancia con respecto a la necesidad de información y no la consulta.

A continuación se revisa otra definición. "*Recuperación de Información*, consiste en seleccionar automáticamente, en una determinada colección de documentos, normalmente muy grande, aquellos que se ajustan a una pregunta del usuario" (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003). La IR no se involucra mucho en la comprensión del lenguaje natural.

Para llevar a cabo su tarea la IR utiliza ciertos métodos, los más usados son el modelo booleano, el vectorial y el probabilístico; pero existen otros como las redes Bayesianas, redes neuronales, redes de inferencia y demás.

2.1.2 Extracción de información

Extracción de Información (Information Extraction (IE), en inglés), (Jackson & Moulinier, 2007) indica que difiere de IR, en que el objetivo no está en la búsqueda de documentos, sino en la búsqueda de información útil dentro de los documentos. Por lo general, los textos en una base de datos o documento electrónico son examinados por programas para ver si contienen cierta información objetivo, que podrían ser términos lingüísticos simples, tal como nombres propios o podrían ser estructuras lingüísticas más complejas, tal como la relación a un cierto tipo de eventos.

Sobre la IE (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003) comenta que la "*Extracción de Información* consiste en obtener información de forma selectiva de un documento (Quién hizo qué, cuándo, cómo, etc). Para ello se definen unas plantillas o esquemas correspondientes al dominio que se desea tratar, que deben ser rellenadas para cada documento".

La IE busca entidades en el texto sobre categorías predefinidas como nombres de personas, organizaciones, lugares, cantidades, valores monetarios, porcentajes, expresiones de hora; y relaciones entre ellas u otros tópicos específicos dentro de los textos.

2.2 Estructuras sintácticas

La extracción automática de información semántica o extracción de hechos que se realiza en el presente trabajo de investigación, es con base en análisis de estructuras sintácticas, específicamente en el análisis de lo que se conoce en gramática como *oración*, el objeto de análisis. Este análisis ayuda a identificar la función sintáctica de las palabras y con ello definir los patrones sintácticos que identifican a los hechos.

Así que en esta sección se explica, qué es una oración, cuáles son sus componentes y como se relacionan estos, cómo se construyen las oraciones y quién se encarga de su estudio.

Una "*estructura*", en el diccionario de la lengua española de la Real Academia Española (RAE) se define como "*1. f. Distribución y orden de las partes importantes de un edificio*", y "*2. f. Distribución de las partes del cuerpo o de otra cosa*". De acuerdo a estas definiciones se puede concluir que "*estructura*" para el objeto de análisis de la investigación se refiere a "*la distribución y orden de las palabras en una oración*" y considerando la sintaxis, podemos agregar que "*las palabras son funciones unas de otras*", como dice (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000) "*las palabras adquieren un significado preciso y cumplen una función sintáctica determinada*", por ejemplo en las siguientes oraciones:

Se lastimó la muñeca izquierda mientras jugaba a la pelota.
La muñeca que le regalé a mi hija cierra los ojos.

Aisladamente, la palabra *muñeca* tiene varias acepciones, pero en cada oración toma una de ellas; además, esta misma palabra cumple una función distinta, en la primera oración es objeto directo y en la segunda, es sujeto.

Dentro de las distintas partes de la gramática, *la sintaxis es la que se dedica al estudio de la oración*. Su estudio se basa en las diferentes funciones que desempeñan los componentes de la oración (Fuentes de la Corte, 2010).

(Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000) Explica que "la sintaxis es la parte de la gramática que estudia la manera como se combinan y ordenan las palabras para formar oraciones; analiza las funciones que aquéllas desempeñan, así como los fenómenos de concordancia que pueden presentarse entre sí".

2.2.1 La oración

Para ubicar más el objeto desde dónde se extraen los hechos: la oración, a continuación se revisa un conjunto de definiciones de este.

- La *oración* es una unidad lingüística dotada de significación que no pertenece a otra unidad lingüística superior y que se caracteriza porque expresa un sentido completo. Por ejemplo, la *palabra* tiene un significado completo, pero no expresa nada si no se combina con otras palabras (Fuentes de la Corte, 2010).
- "La *oración* es la serie o cadena de palabras que transmite un sentido completo" (Fuentes de la Corte, 2010).
- En el diccionario de la Real Academia Española (RAE) se define *oración* como: "5. f. Gram. Palabra o conjunto de palabras con que se expresa un sentido gramatical completo".
- (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000) Expone que la "*oración* es la unidad, dentro del discurso, que expresa un sentido completo y está constituida por sujeto y predicado. El sujeto es de quien se habla en la oración y muchas veces es el agente de la acción del verbo. El predicado es lo que se dice sobre el sujeto".
- (Gartz, 2011) Clasifica a la oración como aquella estructura lingüística que, en la lengua oral, se pronuncia entre dos pausas [pausa = fase de silencio]. Y en el texto escrito toma los puntos como límite de la oración, así como los signos de exclamación e interrogación.

¿Pero cómo es qué una oración tiene sentido completo?, bueno, para que lo tenga una oración se compone de dos partes principalmente: el sujeto y el predicado. "El predicado contiene lo que quiere comunicar el hablante; y el sujeto es una consecuencia del

predicado, una exigencia para la comprensión del sentido. Si digo que algo o alguien *chillaba, sube, aspiró o volvieron* es lógico que la frase haga referencia al ser u objeto en que se realiza esa acción de chillar, subir, etc. (Fuentes de la Corte, 2010)".

"Cuando una persona habla o escribe lo hace empleando frases, esto es, trozos coherentes de su lengua. De otro modo no le entenderíamos. Normalmente decimos que una cosa 'tiene sentido' cuando se entiende, cuando encontramos en ella esa coherencia por la que podemos saber qué se nos está diciendo o qué estamos leyendo. 'Tiene sentido' un libro, un artículo de un periódico, una carta, una conversación. Con sujeto y predicado una oración tiene sentido completo" (Fuentes de la Corte, 2010).

2.2.1.1 Elementos de la oración

La estructura general de la oración es bimembre, es decir se compone de dos elementos: el sujeto y el predicado. Pero estos elementos están formados a su vez por subestructuras más pequeñas y que son en cada miembro el *núcleo* y tienen uno o varios *modificadores*.

Desde el punto de vista semántico, podemos decir que la oración se conforma de dos elementos principalmente: el sujeto y el predicado. El sujeto responde a las preguntas: ¿De qué o de quién se habla?, y ¿Quién o qué realiza la acción? El predicado responde a la pregunta: ¿Qué se dice, de quién se habla o de lo que se habla? (Mora, 2004) En la Figura 2.1 se muestra la oración y sus componentes desde el punto de vista semántico.

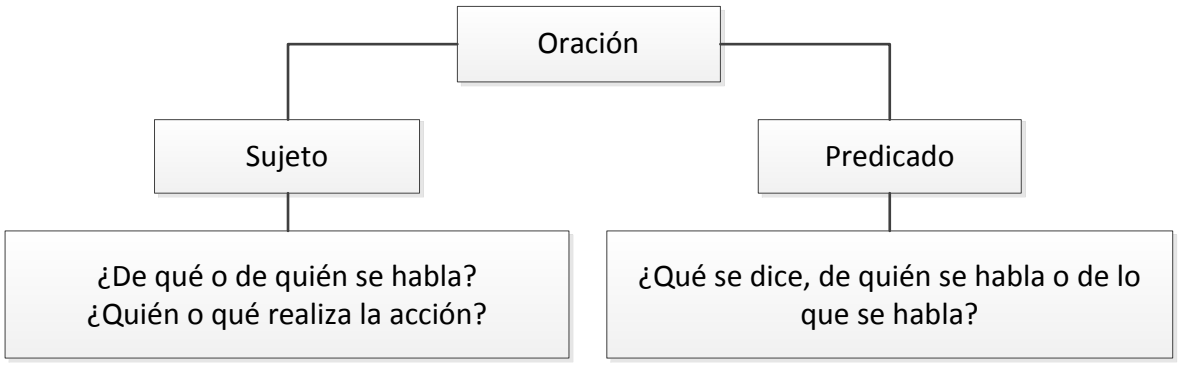

Figura 2.1 La oración y sus componentes desde el punto de vista semántico.

Y desde el punto de vista sintáctico (Fuentes de la Corte, 2010) dice que "sintácticamente una oración tiene un sintagma nominal y otro predicativo", como se muestra en la Figura 2.2.

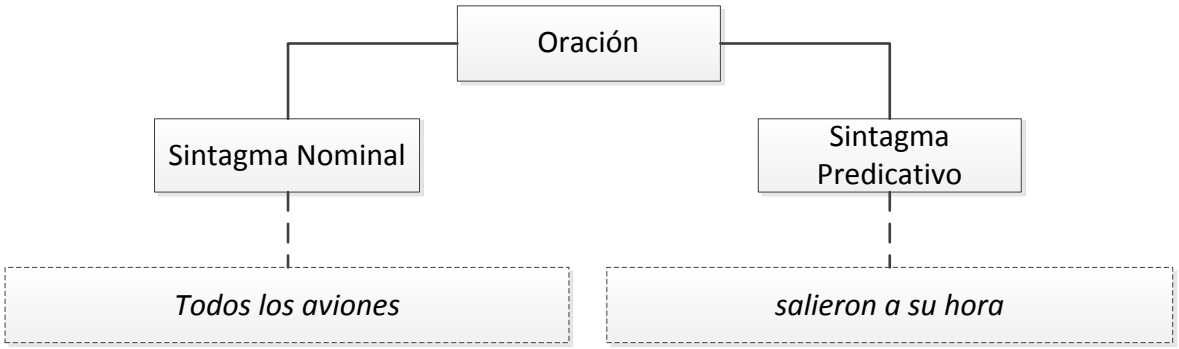

Figura 2.2 La oración y sus componentes desde el punto de vista sintáctico.

"Sintagma es una unidad conformada por una palabra que es la más importante y que funciona como núcleo; éste puede ir acompañado de complementos o modificadores y juntos forman un bloque. Es posible distinguir el núcleo en los sintagmas porque éste es imprescindible y las palabras que lo acompañan se pueden omitir" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). A continuación se definen los sintagmas que pueden existir en una oración.

- **Sintagma nominal**. Tiene como núcleo un nombre o sustantivo; también puede ser un pronombre o una palabra sustantivada.
- **Sintagma adjetivo**. Tiene como núcleo un adjetivo, el cual puede ir acompañado de un adverbio o sintagma adverbial que funciona como su complemento o modificador. El núcleo adjetivo también puede tener como complemento o modificador indirecto, un sintagma prepositivo.
- **Sintagma adverbial**. Su núcleo es un adverbio que puede ser modificado por otro adverbio.
- **Sintagma prepositivo o preposicional**. Está constituido por una preposición, que es el núcleo, y un sintagma nominal que recibe el nombre de término, el cual funciona como complemento de la preposición. Dado que el término es un sintagma

nominal, dentro de él es posible encontrar un núcleo sustantivo con modificadores directos e indirectos.

- **Sintagma verbal**. Tiene como núcleo un verbo y por ello, siempre constituye el predicado de una oración; sus complementos son el objeto directo, el indirecto, los circunstanciales, el predicativo y el agente.

2.2.1.1.1 El sujeto

Es la palabra que se refiere a una idea, un concepto, una persona, un animal o una cosa, de los cuales se dice algo; es de quien se habla en la oración; el sujeto, generalmente realiza la acción del verbo (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

El sujeto también se puede reconocer, además de las preguntas de la Figura 2.1, porque siempre concuerda en número (singular o plural) con el verbo. El sujeto puede encontrarse al principio, en medio o al final de la oración.

El sujeto puede estar constituido por un *pronombre* o un *sustantivo* con o sin modificadores; o sea un sintagma nominal. O puede estar constituido por una *oración*.

A veces el sujeto puedo omitirse; al hacerlo se dice que es *morfológico* y se reconoce por la desinencia del verbo; también suele llamársele sujeto *tácito*.

2.2.1.1.2 Núcleo y modificadores del sujeto

Todo sujeto explícito que sea sintagma nominal tiene un núcleo que es la palabra más importante; puede estar acompañada de modificadores.

- **Modificadores directos**. Acompañan al nombre para agregar algo a su significado o para precisarlo; deben concordar con él en género y número. Esta función la desempañan el artículo y el adjetivo.
- **Modificadores indirectos**. Son sintagmas prepositivos preposicionales que modifican el núcleo del sujeto. Se introducen mediante una preposición; también se llaman complementos adnominales.

- **Aposición**. Es otro tipo de complemento de los nombres; es un sintagma nominal que se caracteriza por escribirse entre comas y por ser intercambiable con el núcleo del sujeto.

2.2.1.1.3 El predicado

Es la parte de la oración que expresa la acción que realiza el sujeto o los diferentes estados en los que éste puede encontrarse; es decir, es todo lo que se dice del sujeto. Está formado por un verbo y sus complementos (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

El verbo puede aparecer sin complementos y constituir, por sí solo, un predicado. El predicado puede estar al principio o al final de la oración; también puede encontrarse dividido, porque el sujeto se ha colocado en medio.

2.2.1.1.4 Núcleo del predicado

El núcleo del predicado siempre es un verbo, simple o perifrástico, es la palabra más importante y concuerda en número y persona con el núcleo del sujeto.

2.2.1.1.5 Predicado verbal y predicado nominal

El predicado verbal es aquel que tiene como núcleo un verbo con significado pleno; es decir, por sí mismo puede predicar o dar información: “hervir, votar, explicar”.

El predicado nominal se construye con verbos copulativos, los cuales se caracterizan por no tener un significado pleno; se acompañan de un adjetivo, un sustantivo o una oración, estos elementos son los que aportan la información del predicado. En estas oraciones el verbo sólo cumple la función de enlazar el sujeto con el predicado, de ahí que reciba el nombre de copulativo. Los verbos copulativos más comunes son *ser* y *estar*.

2.2.1.1.6 Complementos del núcleo del predicado

La estructura del predicado está conformada por el verbo que funciona como núcleo y por los complementos de éste. Los complementos son: objeto o complemento directo, objeto o complemento indirecto, complemento circunstancial, predicativo o atributivo y complemento agente.

Objeto o complemento directo. Se refiere a la persona, animal o cosa que recibe directamente la acción del verbo; se conoce también como paciente, dado que es el que resulta afectado o modificado por la acción del verbo. Se presenta con verbos transitivos.

El complemento directo puede estar formado por:

- Un pronombre: me, te, se, lo, la, los, las, nos, os, todo, algo, etc.
- Un sintagma nominal, constituido por un sustantivo con o sin modificadores.
- Un sintagma preposicional introducido por la preposición *a*. Esta forma sólo se usa cuando le objeto se refiere a personas o seres personificados o singularizados.
- Una oración.

Objeto o complemento indirecto. Es la persona, animal o cosa que recibe indirectamente la acción del verbo; es el beneficiado o perjudicado por la acción. Siempre se une al verbo mediante la preposición *a* y, en algunas ocasiones, acepta la preposición *para*. Es muy frecuente que un pronombre repita el complemento indirecto en una oración.

El complemento indirecto está constituido por:

- Un sintagma prepositivo.
- Un pronombre: me, nos, te, os, se, le, les. El pronombre de complemento indirecto se antepone al verbo, aunque en algunos casos se presenta como enclítico.
- Una oración.

Complemento circunstancial. Expresa la manera, el tiempo, el lugar y demás circunstancias en las que se realiza la acción del verbo. Puede estar formado por:

- Un adverbio, un sintagma adverbial o una locución adverbial.
- Un sintagma prepositivo o preposicional.
- Un sintagma nominal.
- Una oración.

Las múltiples circunstancias en las que se realiza la acción del verbo pueden ser de:

- Modo. Se refieren a la manera como se realiza la acción; responden a la pregunta ¿cómo?
- Tiempo. Expresan el momento en el cual se lleva a cabo la acción; responden a la pregunta ¿cuándo?
- Lugar. Indican el sitio, espacio o lugar donde se realiza la acción, responden a la pregunta ¿dónde?
- Cantidad. En general, sólo se emplean adverbios que indican medida, puesto que denotan cantidad. Responden a la pregunta ¿cuánto?
- Instrumento. Aluden al objeto con el cual se realiza la acción; responden a la pregunta ¿con qué?
- Compañía. Señalan con quién o con quiénes se realiza la acción.
- Tema. Se presentan con verbos, que aluden a las acciones de *leer, hablar, escribir, conservar, pensar*; expresan el asunto, argumento o tema sobre el que tratan dichos verbos, responden a la pregunta ¿sobre qué?
- Causa. Manifiestan las razones o los motivos por los que se realiza la acción; responden a la pregunta ¿por qué?
- Finalidad. Expresan el objetivo o propósito que se persigue con el cumplimiento de la acción verbal. Responden a la pregunta ¿para qué?
- Duda. Expresan incertidumbre.

Complemento predicativo o atributivo. Es el complemento que predica o informa sobre cualidades, atributo o peculiaridades del sujeto. Aparece en las oraciones con predicado nominal, es decir, con los verbos copulativos *ser* y *estar*, también puede presentarse con verbos de significado pleno.

El predicativo se caracteriza porque siempre se refiere al sujeto y, en muchas ocasiones, concuerda con él en género y número. Puede estar formado por:

- Un sintagma nominal.
- Un sintagma adjetivo.
- Un pronombre.
- Una oración.

Complemento agente. Este complemento aparece solamente en las oraciones en voz pasiva y designa al agente de la acción verbal; a pesar de referirse a quién realiza la acción, no es el sujeto. Se introduce por la preposición por.

2.2.2 ¿Cómo se construyen las oraciones?

Para explicar cómo se construyen las oraciones se hará desde el aspecto sintáctico y semántico, pero recordando que la presente investigación analiza la oración desde el aspecto sintáctico. Para dicha explicación se toman las descripciones de (Giammatteo & Albano, 2009).

2.2.2.1 Aspectos sintácticos

Aunque las palabras parecen seguir un orden lineal o secuencial dentro de la oración, en realidad su esquema organizativo es bastante más complejo. Así, mientras en algunos casos es posible alterar el orden entre los elementos que la conforman:

a. Ayer vino Pedro (orden lineal: 1-2-3),
b. Pedro vino ayer (orden lineal: 3-2-1),
c. Vino ayer Pedro (orden lineal: 2-1-3).

En otros, tales alteraciones resultan imposibles:

a. El gato saltó desde el techo (orden lineal: 1-2-3-4-5-6),
b. (*) Gato el saltó desde el techo (orden lineal: 2-1-3-4-5-6),
c. (*) El gato saltó el techo desde (orden lineal: 1-2-3-5-6-4),

d. (*) Desde el gato saltó el techo (orden lineal: 4-1-2-3-5-6).

Sin embargo, algunos cambios de orden son aceptables:

a. Desde el techo el gato saltó (orden lineal: 4-5-6-1-2-3),
b. Saltó desde el techo el gato (orden lineal: 3-4-5-6-1-2).

De acuerdo a lo anterior, algunos componentes de la oración, a los que se llaman *constituyentes*, necesitan desplazarse juntos: esto viene a significar que, por detrás de su aparente linealidad, las palabras que forman una oración se agrupan según principios de jerarquía. En otros términos, las palabras mantienen diferentes relaciones entre sí dentro de la oración (algunas de estas relaciones son más estrechas; otras, más débiles) y forman una *estructura sintáctica* compleja, la que se puede analizar en niveles o grados. Retomando los ejemplos anteriores, se tiene:

a. El gato saltó desde el techo (orden jerárquico: {[1-2]-[3-[4-5-6]]}),
b. Desde el techo el gato saltó (orden jerárquico: {[4-5-6]-[[1-2]-[3]]}),
c. Saltó desde el techo el gato (orden jerárquico: {[3-[4-5-6]]-[1-2]}).

Los constituyentes, ya no las palabras aisladas, son las unidades que dan forma a una oración. De manera típica, la oración consta de dos constituyentes mayores, a saber, el sujeto y el predicado, y de un número no determinado de constituyentes menores, que son los que están incluidos dentro de ambos componentes mayores. Entre el sujeto y el predicado verbal de una oración se establece una relación de concordancia en persona y número expresada a través de la flexión del verbo; esta relación permite no sólo vincular ambos componentes, sino también reconocerlos como los constituyentes mayores de la oración.

2.2.2.2 Aspectos semánticos

La relación de concordancia no agota las relaciones que se establecen entre el sujeto y el predicado. Se vinculan, además, por relaciones de significado. Así podemos decir:

Clara compró una computadora,

Pero no resultaría aceptable:

(*) La computadora compró una casa.

La causa de esta restricción es semántica: el verbo *comprar* exige que su sujeto sea 'humano', una característica de significado (o sea, un rasgo semántico) aplicable a "Clara", pero no a "la computadora".

Las relaciones de significado, sin embargo, no son exclusivas de la relación que se establece entre sujeto y predicado, también se dan entre los distintos constituyentes del predicado. Así, por ejemplo, podemos aceptar:

Caín mató a Abel,

Pero vamos a rechazar:

(*) Caín mató la piedra.

Nuevamente, la causa de esta incompatibilidad semántica deriva de que el verbo *matar* solicita que su complemento sintáctico (o sea, su objeto directo) sea 'animado', un rasgo semántico aplicable a "Abel", pero no a "la piedra". En definitiva, la buena formación de una oración está condicionada tanto por las buenas relaciones sintácticas entre sus constituyentes como de las buenas relaciones semánticas que los vinculan.

Asimismo, en el establecimiento de este condicionamiento sintáctico-semántico para formación oracional, el verbo adquiere un papel fundamental.

El *verbo* es la clase de palabra que, de manera típica, organiza la estructura de la oración: por un lado, realiza la función de núcleo del predicado y, por el otro, selecciona, según su significado (aspecto semántico) y según sus relaciones jerárquicas (aspectos sintácticos) los constituyentes requeridos para que la oración esté bien formada.

Aunque puede funcionar solo – *Trabaja*. / *Corre*. –, el verbo, por lo general, se rodea de un conjunto muy específico de modificadores, que, sin embargo, no se encuentran todos en el mismo nivel de relación con el núcleo. Así, en:

a. Dijo la verdad (inmediatamente).
b. (*) Dijo inmediatamente.
c. Dijo la verdad.

Mientras que la omisión del objeto directo (OD) –la verdad– deja a la oración sin un elemento necesario para su interpretación, el circunstancial de tiempo, al no estar exigido por la semántica del verbo, puede omitirse sin poner en peligro la gramaticalidad de la oración.

Dentro de los modificadores del predicado se reconocen dos grupos:

1) Los *complementos*, que son seleccionados o exigidos por el verbo y desempeñan un papel en la estructura argumental, y
2) Los adjuntos, que son más externos o "periféricos" y no son obligatorios, por lo que pueden agregarse o quitarse libremente, como se puede ver en la siguiente oración:

(Esta madrugada) Pedro se marchó (con su hermana) (en tren) (sin avisar) (hacia la ciudad).

En cuanto a los constituyentes requeridos por la semántica verbal, no todos los verbos tienen las mismas exigencias. Justamente, a los que conocemos como *impersonales* son los que no requieren ningún argumento y pueden funcionar bien solos:

Nieva. / Llueve. / Truena.

Con estos verbos, cualquier agregado que se efectúe será sintácticamente opcional, aunque, desde, el punto de vista pragmático – en relación con el contexto de situación – la información agregada sea, sin duda, relevante:

(El 9 de julio) nevó (copiosamente) (en Buenos Aires).

Un segundo grupo, los tradicionalmente conocidos como *intransitivos*, sólo exigen un acompañante: el sujeto, como en:

Pedro *trabaja*. / El perro *corre*. / El árbol *floreció*.

Por el contrario, hay verbos que requieren más argumentos para completar o precisar su significación. Los *transitivos* exigen dos: sujeto y objeto (a), mientras que los *ditransitivos*, tres: sujeto, objeto directo e indirecto (b), o bien sujeto, objeto y lugar (c):

a. Pedro *pintó* su casa. Ana *escribió* un libro de poemas.
b. Lucas le *compró* una nueva computadora a su hijo.
c. Laura *puso* el jarrón sobre la mesa.

Una consecuencia importante de que se ha venido exponiendo es que dado que los argumentos son obligatorios para la interpretación semántica del verbo, las posiciones argumentales deben estar saturadas, puesto que su ausencia provoca mala formación -(a, b de abajo)- o cambia la interpretación del verbo (c):

a. (*) Pedro *puso* el libro. (Falta dónde; María puso la mesa = 'preparó' la mesa.)
b. (*) Mario *realizó* con meticulosidad. (Falta qué.)
c. Esteban *bebe* un vaso de vino en las comidas. / Esteban *bebe* (= 'es bebedor').

No obstante lo dicho, algunos verbos admiten la omisión de alguno de sus argumentos, pero en estos casos el elemento faltante se interpreta en sentido general –(a) y (b) de abajo– o puede deducirse del contexto discursivo o situacional (c):

a. María *pinta*. (Se interpreta 'obras de arte': cuadros, acuarelas, óleos → 'es pintora'.)
b. Ya *comimos* ('cualquier cosa comestible').
c. Pedro se *puso* (con lo que se esperaba que se pusiera, por lo general dinero o alguna otra cosa valiosa).

El *verbo*, en resumen, elige sus argumentos (es decir, los constituyentes sintácticamente requeridos) y les asigna un determinado valor semántico asociado con su signidicado al que se denomina *papel temático*.

2.3 Enfoques sintácticos de la oración

Para describir formalmente la estructura sintáctica de una oración existen dos enfoques principalmente: constituyentes y dependencias (Galicia Haro & Gelbukh, 2007). Los dos enfoques emplean árboles para representar la estructura sintáctica de una oración, pero se diferencian por el significado de los nodos y sus relaciones en el árbol.

2.3.1 Enfoque de constituyentes

Los constituyentes y la suposición de la estructura de frase, sugerida por Leonard Bloomfield en 1933, es el enfoque en que las oraciones se analizan mediante un proceso de segmentación y clasificación.

Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso para cada parte dividiéndola en subconstituyentes, y así sucesivamente hasta que las partes sean las partes de la palabra indivisibles dentro de la gramática (morfemas).

La suposición de frase y la noción de constituyentes, se aplican de la siguiente manera forma. La frase "*los niños pequeños estudian pocas horas*" se divide en el grupo nominal (GN) "*los niños pequeños*" más el grupo verbal (GV) "*estudian pocas horas*", este último a su vez, se divide en el verbo "*estudian*" más el grupo nominal "*pocas horas*" y así sucesivamente. En la Figura 2.3 se puede ver el árbol de constituyentes de esta frase.

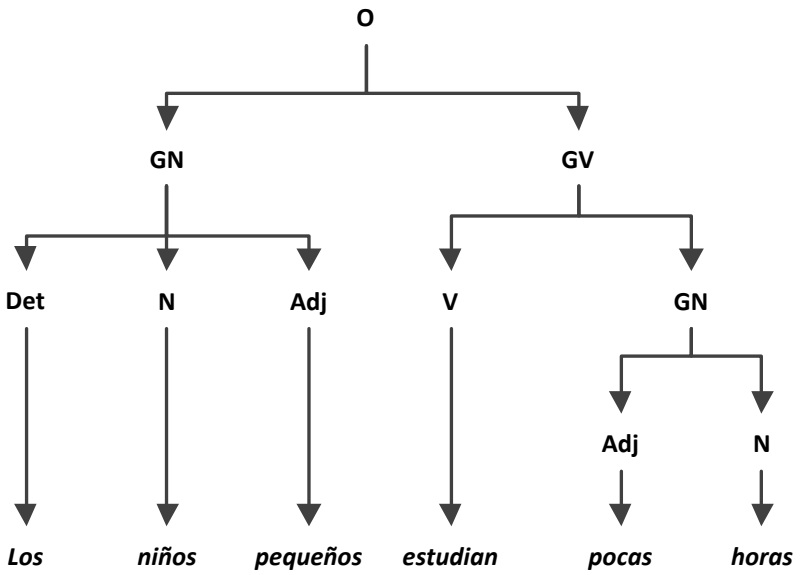

Figura 2.3 Árbol de constituyentes de la oración "*Los niños pequeños estudian pocas horas*".

En el árbol de constituyentes aparecen los siguientes símbolos: O (oración), GN (grupo nominal), GV (grupo verbal), N (sustantivo), GP (grupo preposicional), V (verbo), etc., como etiquetas en los nodos y se tiene el supuesto que estas únicas etiquetas determinan las funciones sintácticas de los nodos correspondientes.

2.3.2 Enfoque de dependencias

Fue Lucien Tesnière en 1959 el primero en construir una teoría que describiera las gramáticas de dependencias.

En este enfoque las dependencias se establecen entre pares de palabras, donde una es principal o rectora y la otra está subordinada a (o dependiente de) la primera (Galicia Haro & Gelbukh, 2007). Por lo tanto si cada palabra de la oración tiene una palabra propia rectora, la oración entera se ve como una estructura jerárquica de diferentes niveles, o sea como un árbol de dependencias. Y la única palabra que no está subordinada a otra es la raíz del árbol.

La motivación de muchas dependencias sintácticas es el sentido de las palabras. Por ejemplo en la frase "*Los niños pequeños estudian pocas horas*", las palabras "*pequeños*" y

"*pocas*" son modificadoras de atributo de las palabras "*niños*" y "*horas*" respectivamente, y "*niño*" es el sujeto de "*estudiar*". Algo muy importante de las dependencias es que no son iguales: una sirve para modificar el significado de la otra, así la secuencia "*los niños pequeños*" denota ciertos niños, y "*estudian pocas horas*" denota una forma de estudiar. En la Figura 2.4 se puede observar el árbol de dependencias de la oración "*Los niños pequeños estudian pocas horas*".

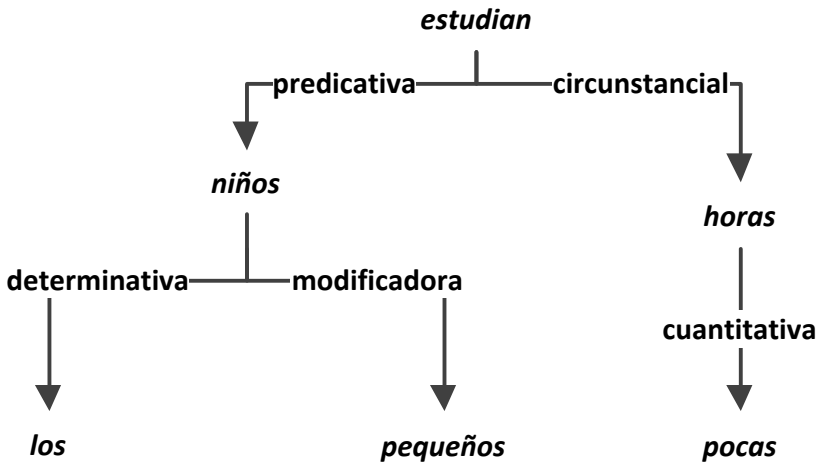

Figura 2.4 Árbol de dependencias de la oración "*Los niños pequeños estudian pocas horas*".

Se pueden observar las siguientes características: los roles sintácticos se indican de forma explícita mediante etiquetas especiales; se muestran cuáles elementos se relacionan con cuáles otros y en qué forma; contiene solamente nodos terminales, no se requiere una representación abstracta de agrupamientos.

2.4 Análisis sintáctico automático

"El proceso de análisis sintáctico consiste en asignar a cada oración de un texto su estructura sintáctica" (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003). El análisis sintáctico automático se logra por medio de un programa informático que toma una oración como entrada y le asigna su estructura sintáctica.

2.4.1 FreeLing

FreeLing es el analizador sintáctico automático (parser, en inglés) que se utiliza en la presente investigación, en su versión 2.2; actualmente se encuentra en su versión 3.0 ya disponible en su sitio web (FreeLing).

Se utiliza porque fue diseñado para trabajar con el idioma español, y porque proporciona los árboles de dependencia de las oraciones analizadas. Los árboles de dependencia representan las estructuras sintácticas de la oración y es con base al análisis de estos árboles como se buscan hechos en las oraciones.

2.4.1.1 Descripción y servicios de FreeLing

FreeLing es un conjunto de herramientas de análisis del lenguaje (diccionarios, lexicones, gramáticas, etc), de código abierto, publicado bajo la licencia GNU General Public de la Free Software Foundation. El proyecto FreeLing fue creado y actualmente es liderado por Lluís Padró. Está desarrollado por el TALP Research Center de la Universitat Politècnica de Catalunya (FreeLing).

Algunos servicios principales que ofrece FreeLing, son los siguientes (Padró, 2011):

- Tokenización de texto. Recibe un texto plano y devuelve una lista de palabras, cada palabra se compone de una tupla <lema, etiqueta, probabilidad lista de sentidos>.
- Separación de sentencias.
- Análisis morfológico. En la oración analiza morfológicamente cada una de sus palabras, contemplando posibles opciones para cada palabra.
- Reconocimiento de multipalabras.
- Reconocimiento de fechas/horas.
- Reconocimiento de expresiones de moneda.
- Reconocimiento de expresiones numéricas (números, cantidades, porcentajes, etc).
- Reconocimiento de expresiones de medidas físicas: velocidad (120 Km/h), longitud (23 cm.) presión (12.3 in/ft2), frecuencia, temperatura, densidad, etc.
- Reconocimiento de nombres propios.

- Etiquetado Part-of-Speech. Para cada oración desambigua la categoría morfosintáctica de cada palabra en la oración.
- Análisis de dependencias basado en reglas: Árboles de dependencias de las oraciones.

A continuación se muestran ejemplos de análisis morfológico, etiquetado Part-of-Speech y el árbol de dependencias de la oración "*El gato come pescado y bebe agua.*", que se pueden consultar en el sitio web de (FreeLing).

Figura 2.5 Análisis morfológico, de FreeLing, de la oración "*El gato come pescado y bebe agua.*".

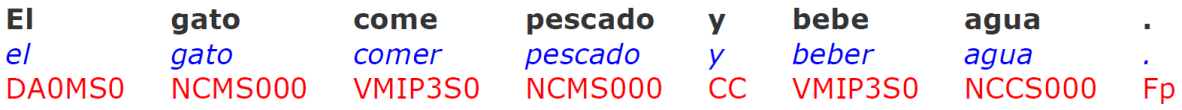

Figura 2.6 Etiquetado Part-of-Speech, de FreeLing, de la oración "*El gato come pescado y bebe agua.*".

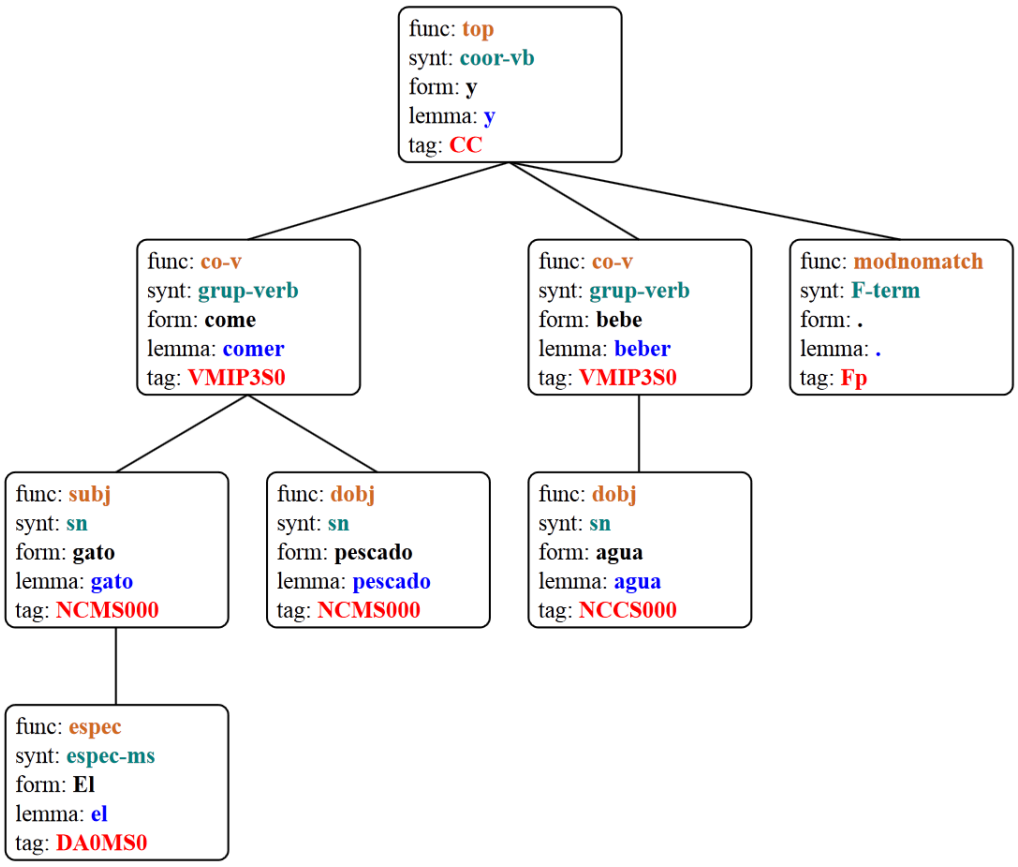

Figura 2.7 Árbol de dependencias, de FreeLing, de la oración "*El gato come pescado y bebe agua.*".

Los idiomas soportados actualmente son el español, catalán, gallego, italiano, inglés, ruso, portugués, el galés y el asturiano.

2.4.1.2 Archivos de dependencias de FreeLing

FreeLing 2.2 se instala en la computadora de desarrollo y ese ahí donde se ejecuta. Al ejecutarse genera los archivos de dependencias utilizados por el sistema del método propuesto. Los archivos son en formato de texto plano y lo que representan es el árbol de dependencias de la oración.

Cada nodo del árbol de dependencias representa una palabra de la oración, contiene información sintáctica y morfológica de cada una de ellas, organizados de forma jerárquica.

A continuación en la Figura 2.8 se muestra un archivo de dependencias generado por FreeLing y a su lado se muestra de forma gráfica para una mejor visualización el árbol de dependencias.

```
grup-verb/top/(procede proceder VMIP3S0 -) [
  sn/subj/(numeración numeración NCFS000 -) [
    espec-fs/espec/(La el DA0FS0 -)
    s-a-fs/adj-mod/(arábiga arábigo AQ0FS0 -)
  ]
  sp-de/sp-obj/(de de SPS00 -) [
    sn/obj-prep/(India india NP00000 -)
  ]
  F-term/term/(. . Fp -)
]
```

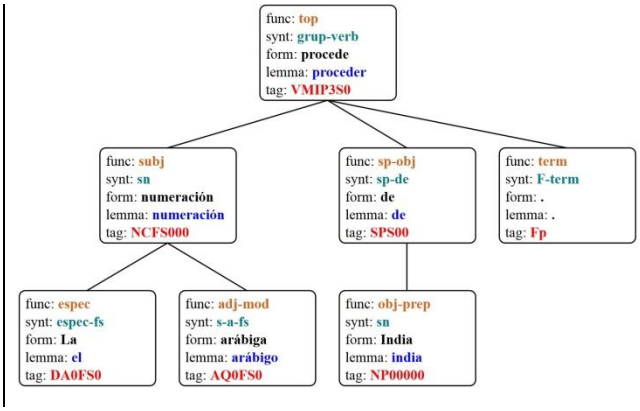

Figura 2.8 Archivo y gráfica, del árbol de dependencias de la oración "*La numeración arábiga procede de India.*".

El archivo y la gráfica representan el árbol dependencias de la oración y como ambos son lo mismo, se pueden observar las mismas etiquetas y la misma jerarquía. En el archivo la jerarquía se representa por medio de los símbolos "[" y "]".

El archivo de dependencias, y no la gráfica, es lo que procesa el sistema del método propuesto, cargándolo en memoria por medio de una estructura de datos del tipo árbol n-ario y es aquí donde se buscan los patrones sintácticos para identificar hechos en la oración.

2.4.1.3 Etiquetas func

La etiqueta *func* en los nodos representa las relaciones o funciones sintácticas del nodo y sus descendientes en el árbol, por ejemplo *sujeto* (subj) u *objeto directo* (dobj).

En la Figura 2.9 se puede ver que el nodo y sus descendientes cumplen la función de sujeto, indicado a través de la etiqueta {func: subj}. La etiqueta *func* obtiene su valor de un conjunto de etiquetas definidas en el proyecto FreeLing (véase Anexo C: Etiquetas sintácticas de dependencias).

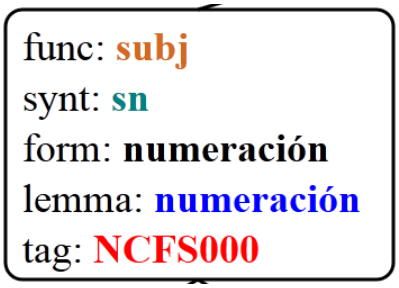

Figura 2.9 Un nodo y sus etiquetas.

2.4.1.4 Etiquetas synt

La etiqueta *synt* en los nodos representa la sintaxis superficial de la palabra en el árbol, por ejemplo *sintagma nominal* (sn) o *grupo pre-posicional* (grup-sp). En la Figura 2.9 se puede ver que la palabra "numeración" es un sustantivo o sintagma nominal, indicado a través de la etiqueta {synt: sn}. La etiqueta *synt* obtiene su valor de un conjunto de etiquetas definidas en el proyecto FreeLing (véase Anexo C: Etiquetas sintácticas superficiales).

2.4.1.5 Etiquetas form y lemma

La etiqueta *form* representa la palabra tal y como está escrita en la oración y la etiqueta *lemma* representa la palabra simple. En la Figura 2.9 se puede ver que la etiqueta *form* y *lemma* son iguales (numeración) pero no siempre es así.

2.4.1.6 Etiquetas tag

El analizador morfológico para el español, de FreeLing, utiliza el conjunto de etiquetas (EagV2.0) para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo "Expert Advisory Group on Language Engineering Standards" (EAGLES) para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas, y por lo tanto del español.

La etiqueta *tag* representa la marca morfológica que ha asignado el analizador morfológico a la palabra. En la Figura 2.9 se puede ver la marca "NCFS000" de la palabra "numeración".

La etiqueta *tag* toma sus valores de un conjunto de categorías definidas en (EagV2.0), dependiendo de la categoría de la palabra en el nodo. Para el ejemplo los valores de la etiqueta *tag* corresponden a los de la Tabla 2.1. Se asigna cero para casos indeterminados.

Tabla 2.1 Valores para la categoría "Nombre" de la etiqueta "tag".

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Las categorías actuales son: Adjetivos, Adverbios, Determinantes, Nombres, Verbos, Pronombres, Conjunciones, Interjecciones, Preposiciones, Signos de puntuación, Numerales y, Fechas y Horas.

2.5 Heurísticas para la extracción de hechos

La Real Academia Española define heurística como: "2. f. Técnica de la indagación y del descubrimiento. 4. f. En algunas ciencias, manera de buscar la solución de un problema mediante métodos no rigurosos, como por tanteo, reglas empíricas, etc.". (Simon, 1999) Define heurística como "conjunto de reglas de sentido común, que pueden aplicarse para resolver problemas complejos y poco estructurados".

Tomando como referencia estas definiciones de heurística, se puede decir que la extracción de hechos por el método propuesto es con base en heurísticas, ya que se lleva a cabo

mediante un conjunto de reglas formuladas por el estudio de patrones sintácticos observados en los árboles de dependencias generados por FreeLing. Los patrones sintácticos son la conjunción de las etiquetas *func*, *synt* y *tag*, y la jerarquía del nodo en el árbol.

De acuerdo a estos patrones se formulan un conjunto de reglas, las heurísticas, para identificar los elementos que componen un hecho en una oración, y extraerlo.

Cabe señalar que la extracción de hechos es un problema complejo y poco estructurado, pues una característica del lenguaje humano ya sea escrito u oral es su complejidad, pues presenta una gran diversidad de estructuras gramaticales. Es por ello que se hace uso de heurísticas.

2.6 Corpus

Un corpus es una colección de elementos lingüísticos seleccionados y ordenados de acuerdo con criterios lingüísticos explícitos, con la finalidad de ser usada como muestra de la lengua (Sinclair, 1996).

Los corpus también se definen como fuente de información de todos los fenómenos del lenguaje y se usan para varios tipos de investigación lingüística – gramatical, sintáctica, pragmática, etc. – (Gelbukh & Sidorov, 2010).

En general se puede decir que los corpus se emplean como un recurso en la investigación y en PLN son muy comunes. El concepto de corpus se ha actualizado con los años, pero actualmente un corpus tiene características como: son textos en formato electrónico, pueden ser escritos u orales, monolingües o multilingües, corpus para identificar errores, corpus para entrenar algún sistema, autenticidad de los datos, la elección de los textos se hace con base a un criterio específico del objetivo del corpus.

Entonces se puede decir que un corpus tiene una estructura específica, se ha conformado con un criterio lingüístico específico, y principalmente se usa para la investigación.

A continuación se dan algunos ejemplos de corpus escritos, que pueden encontrarse en formato electrónico, y son utilizados por las áreas de PLN.

- Un conjunto de palabras en general o específicas del lenguaje. Por ejemplo un conjunto de verbos para identificar cuales expresan sentimientos.
- Un conjunto de mensajes tomados de la red social llamada Twitter. Para estudiar el tipo de abreviaturas que la gente utiliza en estos mensajes. O para clasificar los mensajes como de opinión positiva o negativa.
- Un conjunto de imágenes recolectadas de la Web. Que sirva como conjunto de pruebas para buscadores de imágenes, buscadores que trabajan con base en la imagen misma y no de palabras.
- Un conjunto de textos para entrenar sistemas que identifican la autoría de dichos textos.
- Un conjunto de resúmenes de textos creado por humanos para evaluar a los sistemas informáticos que crean resúmenes automáticos.
- Un conjunto de artículos de noticias tomados de periódicos de la Internet para realizar minería de texto.
- Internet como un corpus enorme. Internet es un corpus muy especial, porque no cuenta con el marcaje y la estructura que usualmente ofrecen otro tipo de corpus, lo que resulta en el desarrollo de métodos especiales para su análisis (Gelbukh & Sidorov, 2010).

El corpus que se utiliza en la presente investigación está formado por un conjunto de oraciones y cada oración por un conjunto de hechos. Las oraciones se han tomado de lecciones de libros de educación primaria y secundaria.

Se han seleccionado este tipo de libros porque contienen muchas definiciones e información enunciativa; ya que han sido redactados para cumplir un propósito educativo, y por lo tanto contienen gran cantidad de hechos.

Este corpus se utiliza para ayudar a formular un método de extracción de hechos, mediante la búsqueda de patrones sintácticos que nos identifican hechos en una oración. Y también

se utiliza para evaluar el método propuesto, comparando la cantidad y hechos correctos extraídos por el experto contra los extraídos por el sistema.

2.7 Definición de hecho

Para llegar a la definición formal de hecho que se utiliza en la presente investigación antes se revisan algunas definiciones en la literatura.

2.7.1 Algunas definiciones de hecho

- La Real Academia Española define hecho como: "m. Acción u obra, m. Cosa que sucede, m. Asunto o materia de que se trata" (RAE).
- Cierto evento o estado de situación, que puede ser una acción, un proceso, un estado físico o mental (Giammatteo & Albano, 2009). Por ejemplo: "Mamá *cocina* una torta", "La papa *aumentó* terriblemente", "La ventana *quedó* abierta" y "El estudiante *sabe* matemática".
- Unidades de texto más pequeñas que la oración, que se pueden obtener a través de la descomposición de la oración en una colección de frases. Cada frase tiene información independiente que puede ser usada como una unidad independiente (Hovy, Zhou, & Kwon, 2007).
- (Joosse, 2007) clasifica a los hechos en dos tipos: hechos simples o básicos y hechos complejos. Ejemplos de hechos simples son: "Albert Einstein" y "14 de marzo de 1879", dos entidades que por separadas no transmiten información semántica, pero si se establece una relación entre ellas llamada "fecha de nacimiento", inmediatamente toman sentido las dos entidades e indican la fecha de nacimiento de Albert Einstein. La relación entre las dos entidades es un hecho complejo, y lo que (Joosse, 2007) llama "un hecho".

2.7.2 Definición formal de hecho en esta investigación

Considerando las definiciones expuestas y en particular la de (Hovy, Zhou, & Kwon, 2007) y tomando en cuenta que con sujeto y predicado una oración tiene sentido completo (Fuentes de la Corte, 2010); en este trabajo de tesis a estas frases que por ellas mismas

contienen "información semántica", se les llama "hechos", y la definición formal que se utiliza en la presente investigación es:

*Un **hecho** es la unidad mínima de texto que se puede extraer de una oración, tiene independencia semántica, únicamente un verbo y su forma es una tripleta conformada así*:

Hecho = [*Sujeto*] + [*Verbo*] + [*Objeto/Complemento*]

Por ejemplo en la oración: "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*", se pueden identificar los hechos que se muestran en la Tabla 2.2.

Tabla 2.2 Hechos identificados en la oración "La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula".

No.	Sujeto	Verbo	Objeto/Complemento
1	La civilización China	heredó	el papel
2	La civilización China	heredó	la pólvora
3	La civilización China	heredó	una forma de imprenta rudimentaria
4	La civilización China	heredó	la brújula

Observaciones del ejemplo:

- Cada hecho tiene independencia semántica, es decir, ninguno necesita a otro para tener sentido completo o informar algo.
- Todos tienen un solo verbo.
- Todos cumplen la tripleta que define un hecho.
- Una oración puede contener varios hechos.

Con base en esta definición se identifican los patrones sintácticos, el desarrollo de las heurísticas y algoritmos para la extracción de hechos.

En el resto del documento de la tesis, al tercer componente de la tripleta del hecho ([*objeto/Complemento*]) se le llama solamente [*Complemento*] para simplificar la escritura de la tripleta o para referirse a este componente del hecho.

2.7.3 Características de un hecho

Después de extraer los hechos se podría revisar que su información semántica concuerde con el mundo real, y para ello se podrían revisar ciertas características que determinen si es un hecho. Este filtro no se considera en el método propuesto debido a que el corpus se conforma de oraciones de libros de educación, los cuales contienen información objetiva, concreta y verídica.

De todas formas se consideran en la redacción para futuras investigaciones. No son hechos: las creencias, opiniones, suposiciones, ideas subjetivas, expresiones del futuro, expresiones de duda, expresiones de probabilidad.

Si son hechos: Definiciones, acciones del pasado.

3 ESTADO DEL ARTE

A continuación se revisan algunos trabajos relacionados con la extracción de información semántica, lo que en la presente investigación se llaman hechos.

3.1 Extracción de hechos con intervención de usuario y entrenamiento

Extracción de hechos con intervención de usuario y entrenamiento (User Trainable Fact Extraction (UTFE), en inglés) es un sistema desarrollado por (Joosse, 2007) y explica que es un proceso en el cual un usuario le dice al sistema que tipo de información, en cuales documentos, él está interesado. El sistema debe usar estos datos para encontrar el mismo tipo de información en documentos similares. El usuario anota la información etiquetando fragmentos de texto en un documento basado en una ontología. Información adicional puede ser especificada mediante la anotación de hechos, los cuales representan las relaciones entre las anotaciones. Las anotaciones proporcionan al sistema la información necesaria para extraer información concreta desde los nuevos documentos. El sistema UTFE presenta la información recientemente descubierta junto con los documentos al usuario.

Para la implementación de UTFE se desarrollaron herramientas de anotación y extracción de información. El usuario interactúa con la herramienta de anotación mientras la herramienta de extracción de información intenta encontrar los hechos en los que el usuario está interesado, en los nuevos documentos.

Para evaluar el rendimiento del sistema el usuario probó con un dominio formado por los ganadores del premio Nobel. Aquí el usuario trata de entrenar al sistema para encontrar el nombre del ganador del premio Nobel y el premio Nobel que él o ella ganó. La parte de extracción de información del sistema UTFE es evaluada por el rendimiento en un número de corpus.

Diferente a la Recuperación de Información tradicional, el enfoque de extracción de información de UTFE es mostrar al sistema algunos documentos ejemplo donde la información que el usuario está buscando está indicada en el documento. Entonces el

sistema emplea estos documentos para regresar información similar extraída desde otros documentos.

(Joosse, 2007) Explica que la Extracción de Hechos (Fact Extraction (FE), en inglés) es una capa adicional de extracción sobre la Extracción de Información. La meta de la FE es encontrar piezas de información y relacionarlas entre ellas. Por ejemplo, si alguien está interesado en fechas de cumpleaños de personas, entonces la tarea del motor FE es primero encontrar nombres de personas y fechas de cumpleaños en un documento, después las personas encontrados y fechas de cumpleaños tienen que ser relacionadas para determinar a cual persona que fecha de cumpleaños le pertenece. Se puede ver que la primera tarea pertenece al área de Extracción de Información, pero relacionar la información es lo que distingue la FE de la Extracción de Información.

La información obtenida por EI es llamada un hecho simple, una secuencia de texto señalada con etiqueta proporcionando información adicional. El programa FE obtiene *hechos complejos*; *una relación entre anotaciones la cual proporciona información adicional acerca de las anotaciones*.

En la investigación de (Joosse, 2007), *un hecho simple es conocido como una anotación*, y un *hecho complejo es conocido como hecho*. La Figura 3.1 muestra un fragmento de texto donde se indica la fecha de nacimiento de Albert Einstein con hechos simples y complejo.

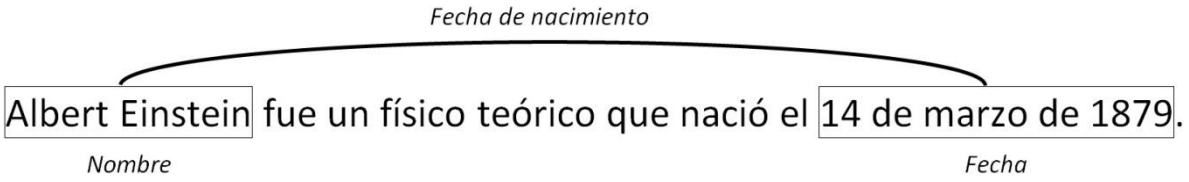

Figura 3.1 Fragmento de texto anotado con hechos simples y complejo.

El sistema UTFE combina técnicas de Extracción de Información y Recuperación de información con una interfaz de usuario amigable para automáticamente descubrir hechos en nuevos documentos. En resumen, un usuario anota, en la "*herramienta de anotación del sistema*", un número de documentos con información (anotaciones y hechos) en los que él

está interesado. Una vez que un número adecuado de documentos han sido anotados, el usuario ordena al sistema encontrar la información en otros documentos.

Se puede ver que la tarea del usuario que interactúa con el sistema es etiquetar ciertas piezas de información dentro del documento con información adicional. Esta información adicional viene de una ontología y especificación de hechos. La ontología es una lista de etiquetas de información que el usuario quiere agregar un fragmento de texto. Por ejemplo, la palabra o elemento "*William*" no significa nada, sólo si un usuario etiqueta a este como el nombre de una persona, es como se convierte en información para el sistema. Estas piezas de texto etiquetadas representan hechos simples o anotaciones y es la tarea de los algoritmos de EI descubrirlos. Los hechos son las relaciones entre las anotaciones, relaciones entre anotaciones y otros hechos o relaciones entre hechos (Joosse, 2007).

Después de esta descripción general del sistema UTFE, se tienen las siguientes conclusiones:

- Este sistema se enfoca sólo en documentos en el idioma inglés.
- La definición de hecho que propone (Joosse, 2007), diferenciando hecho simple y complejo es realmente muy interesante y novedosa.
- El sistema trabaja a partir de un conjunto de documentos, elegidos inicialmente, o sea un dominio específico.
- En los documentos elegidos inicialmente se hacen anotaciones por el usuario de la información que le interesa, luego lanza una consulta buscando nuevos documentos con la información etiquetada anteriormente.
- Utiliza una herramienta para recuperar documentos similares al conjunto de documentos elegidos durante el entrenamiento del usuario, de donde se extraerá la información requerida.
- Se puede decir que es un sistema semiautomático ya que las anotaciones y la relación entre ellas las tiene que hacer el usuario. Las hace mediante una herramienta de anotaciones, y ya en la ejecución del sistema un algoritmo de EI realiza nuevas anotaciones en los nuevos documentos, tomando como referencia los documentos anotados por el usuario.

- Para trabajar, el sistema necesita la intervención del usuario y de entender ciertos conceptos como el de hechos simples y las relaciones entre ellos (hecho complejo).
- El usuario necesita conocer y entender la ontología para representar la información que le interesa.
- Se realiza entrenamiento del sistema.

En la presente investigación a diferencia de (Joosse, 2007) se construye un sistema que extrae hechos de forma completamente automática mediante heurísticas que no necesitan entrenamiento. La definición de hecho que se maneja, es la de hecho complejo comparado con Joosse. Una diferencia muy importante del sistema desarrollado en la presente investigación es que este extrae todos los hechos existentes en una oración de un documento, pues aquí no se definen hechos o información específica de forma previa. El corpus son documentos en el idioma español.

Otras diferencias, no se realiza anotaciones en las oraciones ni se utilizan ontologías ya que la extracción es con base al análisis de estructuras sintácticas. Los hechos se guardan en una base de datos relacional.

3.2 Un esquema de evaluación semiautomática

(Hovy, Zhou, & Kwon, 2007) Explican que en muchas tareas de procesamiento de lenguaje natural, se encuentra el problema de determinar el nivel de granularidad adecuado para las unidades de información. Comúnmente los investigadores emplean las oraciones como la unidad individual de información. Sin embargo, un gran número de aplicaciones de PLN requieren que se determinen "*unidades de texto más pequeñas que las oraciones, básicamente descomponiendo sentencias dentro de una colección de frases. Cada frase contiene una pieza independiente de información que puede utilizarse como una unidad independiente. Estas unidades de información más específicas les llaman nuggets*".

Por ejemplo, de la siguiente oración:

> The Danube at Cernavoda village, where the reactor is located, fell to a depth of less than three meters on Saturday, down from its usual level of almost seven meters.

Se obtiene, ligeramente reescritos para mejorar la legibilidad, los siguientes nuggets:

- Danube
- Danube is at Cernavoda village
- the Cernavoda village is where the reactor is located
- Danube fell
- Danube fell to a depth of less than three meters
- Danube fell on Saturday
- Danube fell down from its usual level of almost seven meters

Para la extracción de nuggets emplean árboles sintácticos producidos por el analizador sintáctico Collins para obtener la representación estructural de las oraciones. Los nuggets se extraen mediante la identificación de subárboles que son descripciones de entidades y eventos. Para nuggets entidad, examinan subárboles encabezados por "NP", para nuggets evento, se examinan subárboles encabezados por "VP" y sus correspondientes sujetos (hermanos encabezados por "NP") se tratan como entidad adjunta para la frase verbal.

De la oración ejemplo y sus nuggets, se puede observar que la definición de nugget o hecho para la presente investigación es diferente, pues Hovy considera un nivel granularidad de hechos diferente, como:

- Sustantivo (Sujeto)
- Sustantivo (Sujeto) + Verbo
- Sustantivo (Sujeto) + Verbo + Objeto/Complemento

Lo que marca una diferencia con respecto a la forma de un hecho como una tripleta únicamente: Sujeto + Verbo + Objeto/Complemento. Y otra diferencia muy importante es que trabajan con textos en el idioma inglés.

Los resultados obtenidos comparados contra los de un humano tienen una media geométrica de 0.7465.

3.3 Sistema de extracción automática de información semántica de los libros de texto estructurados

Es un método propuesto por (Herrera de la Cruz, 2010), extrae hechos mediante heurísticas creadas con base al análisis de estructuras sintácticas, de libros de texto en español. Las heurísticas examinan árboles de dependencias creados por el analizador sintáctico llamado "*Connexor Versión 3.8.5"* para español de la compañía *Connexor Oy*. Los hechos se almacenan en una base de datos relacional.

Connexor es un software con derechos de autor, es decir, para usarlo se debe comprar una licencia, no tiene un buen desempeño para el análisis de oraciones en español ya que en ocasiones los árboles de dependencias que proporciona no están bien formados (Herrera de la Cruz, 2010), dejando algunas veces nodos huérfanos o palabras donde no se reconoce la función sintáctica.

Las heurísticas desarrolladas son muy genéricas, no tocan algunos tópicos comunes, y la descripción de cómo se relacionan entre ellas es poco clara. Por ejemplo, no maneja la coordinación de adjetivos, la coordinación de preposiciones, el atributo nominal para los verbos copulativos. Sobre la aparición de varios patrones sintácticos en una misma oración y como trabajan las heurísticas sólo se trata de manera concreta la coordinación de sustantivos con la coordinación de verbos.

Los resultados obtenidos comparados contra los de un humano son: precisión del 80%, recall del 73% y F1 del 76%.

4 MÉTODO PROPUESTO

Se presenta y explica la arquitectura general del método propuesto; un diagrama y su descripción de los patrones sintácticos que identifican los hechos, además de un ejemplo; los algoritmos de las heurísticas desarrolladas con base a los patrones sintácticos y ejemplos de hechos extraídos con estos algoritmos.

4.1 Arquitectura general

La arquitectura general del método propuesto para la extracción de hechos se muestra en la Figura 4.1.

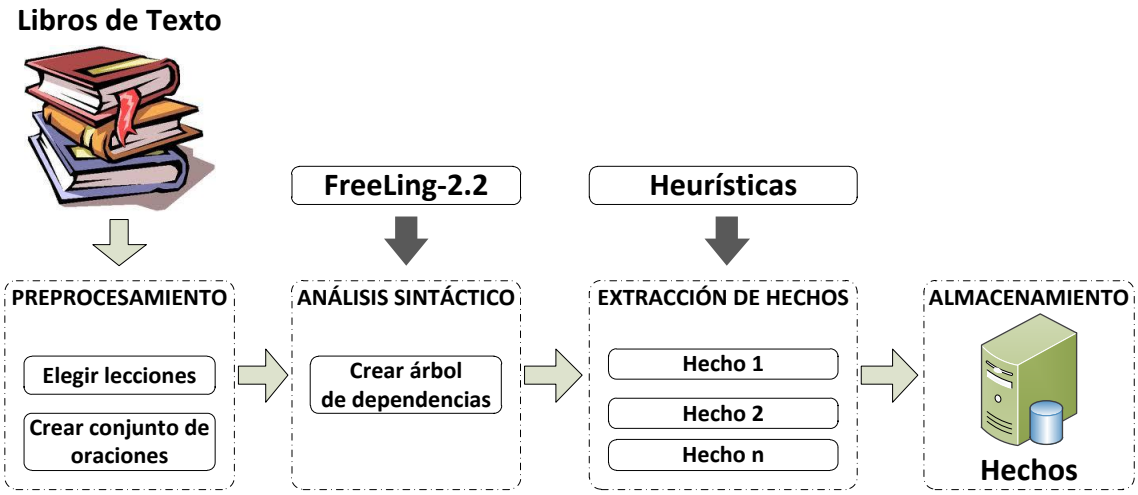

Figura 4.1 Arquitectura general del método propuesto.

De libros de texto se eligen las lecciones de interés, y de estas se crea un conjunto de oraciones. Después FreeLing realiza un análisis sintáctico de cada oración y crea un árbol de dependencias de cada una de ellas. Luego mediante un conjunto de heurísticas se analizan los árboles de dependencias y se extraen los hechos de cada oración. En seguida se almacenan los hechos en una base de datos relacional.

4.2 Libros de texto

Libros de texto se refiere a libros didácticos utilizados para la enseñanza de alguna materia de estudio. Estos libros, por su objetivo, están estructurados por temas y subtemas, también llamados capítulos y subcapítulos. Cada subtema está compuesto por un título, un conjunto de párrafos y cada párrafo por un conjunto de oraciones. En la investigación a un subtema se le llama lección.

La estructura que presentan los libros de texto es una de las razones de seleccionarlos para extraer hechos. Otra razón y la más importante, porque contienen muchas definiciones e información enunciativa; ya que han sido redactados para cumplir un propósito educativo, y por lo tanto contienen gran cantidad de hechos.

4.3 Preprocesamiento

En el preprocesamiento se eligen las lecciones de interés para extraer hechos, de cada lección se extraen solamente los párrafos de información que la conforman, es decir elementos como, tablas, imágenes, gráficas, ecuaciones, indicaciones o preguntas para los estudiantes; no son tomados en cuenta.

Después cada párrafo se separa en un conjunto de oraciones, considerando que el delimitador es el símbolo de punto. Todas las oraciones se guardan en un archivo de texto plano.

4.4 Análisis sintáctico

Se realiza análisis sintáctico de cada oración en el archivo, mediante (FreeLing) quien hace un análisis morfológico y sintáctico y da como resultado un árbol de dependencias por cada oración.

Para el análisis sintáctico FreeLing utiliza sus propias etiquetas (véase anexo C), y para el análisis morfológico emplea un conjunto de etiquetas (véase anexo D) propuesto por el grupo Expert Advisory Group on Language Engineering Standards (EagV2.0).

4.4.1 Árbol de dependencias

Cada nodo del árbol de dependencias representa una palabra de la oración, contiene información sintáctica y morfológica de cada una de ellas, organizados de forma jerárquica.

El árbol de dependencias generado por FreeLing se representa en un archivo de texto plano, como el que se muestra en la Tabla 4.1 de la oración "*La numeración arábiga procede de India.*".

Tabla 4.1 Árbol de dependencias en formato de texto de la oración "*La numeración arábiga procede de India.*".

```
grup-verb/top/(procede proceder VMIP3S0 -) [
  sn/subj/(numeración numeración NCFS000 -) [
    espec-fs/espec/(La el DA0FS0 -)
    s-a-fs/adj-mod/(arábiga arábigo AQ0FS0 -)
  ]
  sp-de/sp-obj/(de de SPS00 -) [
    sn/obj-prep/(India india NP00000 -)
  ]
  F-term/term/(. . Fp -)
]
```

En la tabla, las etiquetas sintácticas y morfológicas se han puesto en negritas, el orden en que aparecen es: { **synt**/**func**/(**form lemma tag** -) }, cada una de estas líneas es un nodo.

El símbolo "/" sirve para separar las etiquetas, los símbolos "**(**, -, **)**" encierran la etiqueta morfológica "**tag**", los símbolos "**[,]**" representan la jerarquía del árbol.

En la Figura 4.2 se muestra el mismo árbol de dependencias, pero ahora, en forma de gráfica.

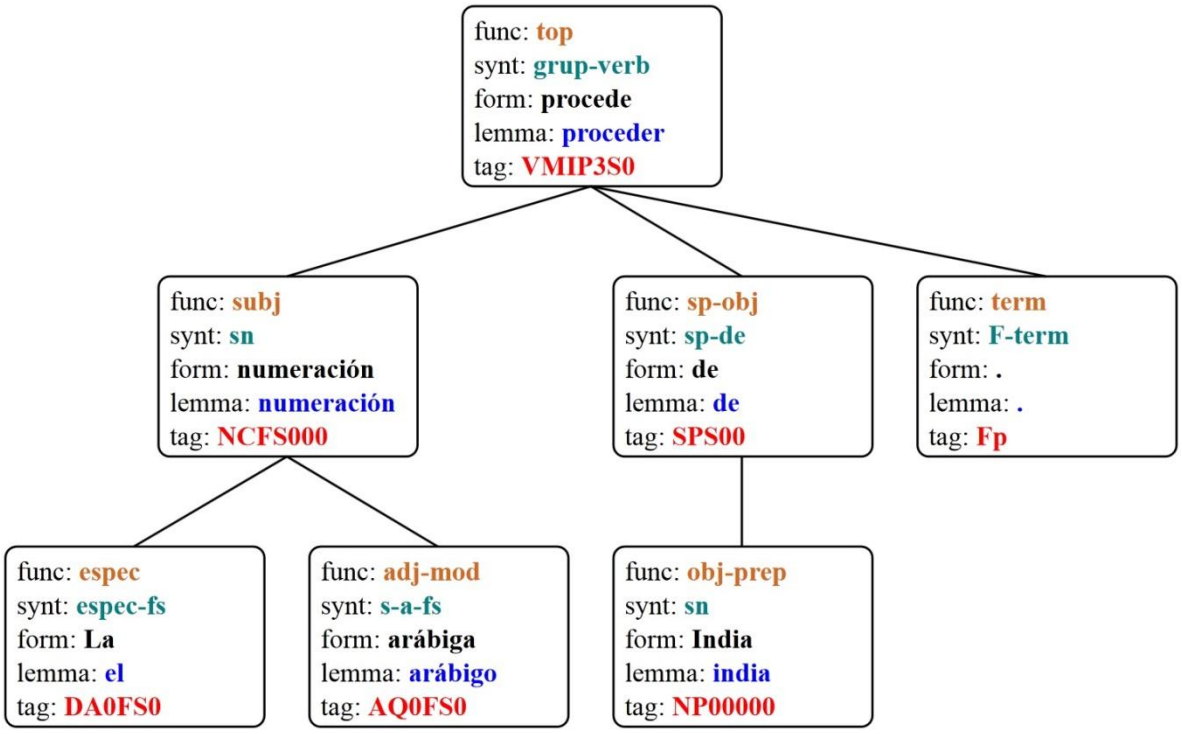

Figura 4.2 Árbol de dependencias en forma de gráfica de la oración "*La numeración arábiga procede de India.*".

En el anexo A se pueden consultar otros ejemplos de árboles de dependencias más complejos.

4.5 Extracción de hechos

4.5.1 Heurísticas

Para la extracción de hechos el método propuesto utiliza los archivos de los árboles de dependencias generados por FreeLing. Y mediante un conjunto de heurísticas se extraen los hechos que contienen las oraciones.

4.5.1.1 Cómo trabajan las heurísticas

Las heurísticas hacen uso de los datos morfológicos y sintácticos de los árboles de dependencias y buscan patrones sintácticos en ellos. Estos patrones son la conjunción de las etiquetas *func*, *synt* y *tag*, y la jerarquía del nodo en el árbol. Mediante estos patrones se extraen los componentes de la tripleta que define un hecho.

La búsqueda de los patrones inicia en el nodo raíz del árbol, luego se revisan sus descendientes y si es necesario se avanza otro u otros niveles hacia abajo, pues los patrones pueden estar anidados. Es a través de esta búsqueda como se van obteniendo los componentes de los hechos y al mismo tiempo construyéndolos.

En las siguientes secciones se describen los patrones sintácticos y las heurísticas.

4.5.2 Convenciones para describir las heurísticas

De acuerdo a su definición, la forma de un hecho es la tripleta [*Sujeto*] + [*Verbo*] + [*Objeto/Complemento*], pero en la descripción de las heurísticas se utilizará la palabra "Complemento", refiriéndose a "Objeto/Complemento", esto para simplificar la escritura de la tripleta.

En la descripción de las heurísticas se utilizan algunos símbolos para agrupar y separar algunos términos.

- { } = Representan a un nodo y puede contener una o más etiquetas. Por ejemplo {func: cc / synt: subord / tag: CS}.
- / = Se utiliza para separar las etiquetas del nodo. Por ejemplo: {func: cc / synt: subord / tag: CS}, el símbolo '/' separa tres etiquetas.
- [ ] = Encierra a un componente del hecho: sujeto, verbo u objeto/complemento. Por ejemplo: Hecho = [Sujeto] + [Verbo] + [Complemento].
- ; = Significa disyunción.
- , = Significa conjunción.
- < > = Representa algún patrón en los diagramas de los patrones sintácticos. Por ejemplo: < Coordinación de Sustantivos >, significa el patrón de este nombre.
- * = Representa cualquier carácter. Por ejemplo: {V******} significa una 'V' seguida de seis carácter cualesquier.
- En la construcción de los hechos el símbolo de punto y los determinantes no son tomados en cuenta.

- En los diagramas de los patrones sintácticos, las líneas punteadas indican que el patrón podría presentarse o no.

En la descripción de las heurísticas se sigue la siguiente estructura:

- Nombre de la heurística.
- Descripción teórica.
- Diagrama del patrón sintáctico.
- Ejemplo del patrón.
- Se describe de forma detallada el algoritmo de la heurística.
- Ejemplo de hechos extraídos: Los ejemplos de la heurística "Básica" y "Coordinación de Verbos" son los que se muestran para todas las otras heurísticas.

4.5.3 Algoritmo clasificador

La ejecución de las heurísticas se inicia por medio de un algoritmo que analiza la raíz del árbol para determinar la heurística inicial que se aplicará. Se tienen dos heurísticas principales que llaman a las demás.

- Si la raíz es un verbo se llama a la heurística "Básica".
- Si la raíz es una conjunción se llama a la heurística "Coordinación de Verbos".

En la Tabla 4.2 se describe de forma detallada el algoritmo clasificador.

Tabla 4.2 Algoritmo "*Clasificador*".

```
1   Clasificador:
2   Construir árbol con el archivo de dependencias.
3   Si raíz del árbol tiene la etiqueta {tag: V******}:
4         Llamar algoritmo de la heurística
          Básica (Apuntador a nodo raíz).
5   Si raíz del árbol tiene las etiquetas: {synt: coor-vb / tag: CC}:
6         Llamar algoritmo de la heurística
          Coordinación de Verbos (Apuntador a nodo raíz).
7   Fin.
```

4.5.4 Complemento simple

Al buscar el complemento para el hecho, puede presentarse de forma simple, es decir, un nodo con etiquetas que no representa ningún otro patrón sintáctico, por eso mismo representa el complemento directo del hecho.

A continuación se describen los tipos de complemento simple y las etiquetas que FreeLing les coloca, pues en la descripción de las heurísticas se hace referencia a este complemento.

Complemento circunstancial. "Expresa la manera, el tiempo, el lugar y demás circunstancias en las que se realiza la acción del verbo" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Puede estar formado por:

- Un sintagma prepositivo o preposicional. Se etiqueta con {func: cc / synt: grup-sp} o {func: cc / synt: sp-de}.
    - Se comporta de una forma extraña.
- Un sintagma nominal. Se etiqueta con {func: cc / synt: sn}.
    - Esta tarde comenzará un diluvio.
- Un verbo en gerundio subordinado a otro verbo. Se etiqueta con: {func: cc / synt: subord-ger}.
    - Pedro y Luis llegaron siguiendo a los caballos.

Objeto directo. "El complemento directo se refiere a la persona, animal o cosa que recibe directamente la acción del verbo; se conoce también como paciente, dado que es el que resulta afectado o modificado por la acción del verbo" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Puede estar formado por:

- Un sintagma nominal, constituido por un sustantivo con o sin modificadores. Se etiqueta con {func: dobj / synt: sn}.
    - Esa canción transmitía alegría.

- Sintagma preposicional. Se etiqueta con {func: dobj / synt: grup-sp}.
    - Buscaban a los estudiantes.

Sintagma prepositivo o preposicional. "Ésta compuesto por una preposición, que es el núcleo, y un sintagma nominal que recibe el nombre de término, el cual funciona como complemento de la preposición" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Por ejemplo la preposición "en". Se etiqueta con {func: sp-obj / synt: grup-sp}.

- Pedro nació en París.

O la preposición "de". Se etiqueta con {func: sp-obj / synt: sp-de}.

- Luis viene de Inglaterra.

Cuando el verbo principal de la oración es copulativo (*ser*, *estar*). Se etiqueta con {func: att / synt: grup-sp}.

- Las hojas son para la impresora.

Cuando se adjunta alguna frase con preposición, por ejemplo "hasta, entre, mediante". Se etiqueta con {func: ador / synt: grup-sp}.

4.5.5 Heurística: Básica

Esta heurística se aplica a los árboles que tienen el patrón sintáctico donde la raíz es un verbo, y se le llama básica porque es la forma común de los árboles. La estructura del patrón puede ser simple o compleja.

En la estructura simple el nodo raíz es el verbo del hecho, y en los descendientes inmediatos se tiene un nodo que representa al sujeto y a otro que representa al complemento. La Figura 4.3 muestra la estructura simple del patrón.

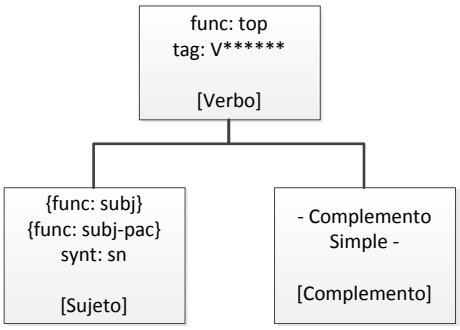

Figura 4.3 Diagrama del patrón sintáctico "*Básico*": Estructura Simple.

Se puede ver que el *verbo* esta etiquetado con {func: top / tag: V******}, el *sujeto* con {func: subj / synt: sn} ó {func: subj-pac / synt: sn} y {synt: sn}; y el *complemento* tiene alguna de las etiquetas enumeradas como complemento simple.

Las Figura 4.4 y Figura 4.5 muestran ejemplos del patrón sintáctico básico de estructura simple en las oraciones.

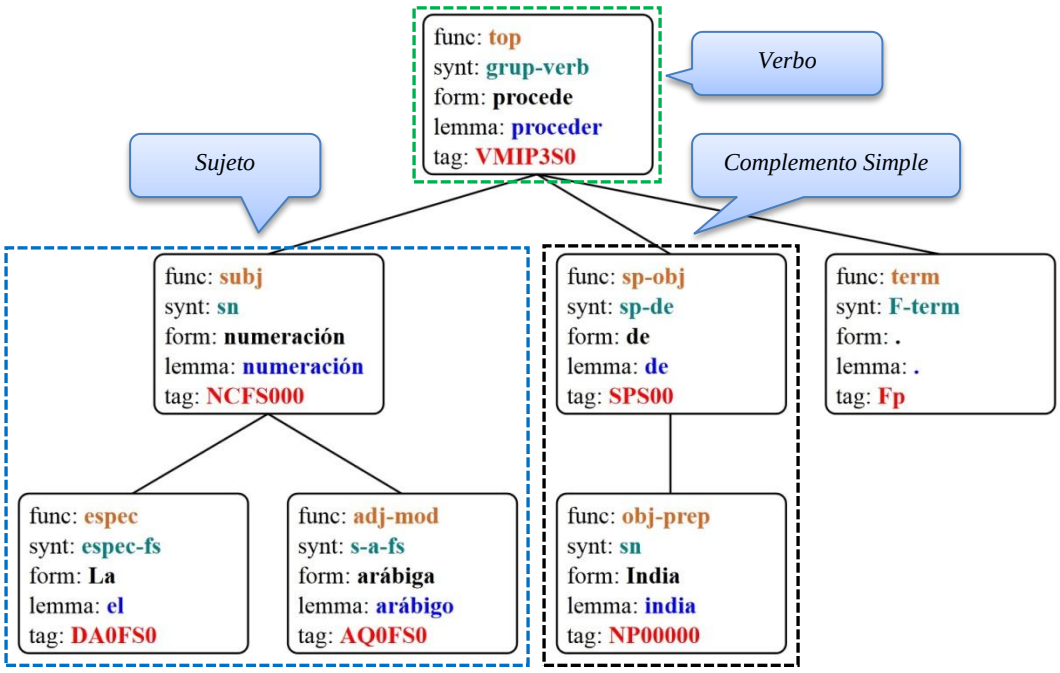

Figura 4.4 Patrón sintáctico "*Básico*" en el árbol de dependencias de la oración "*La numeración arábiga procede de India*".

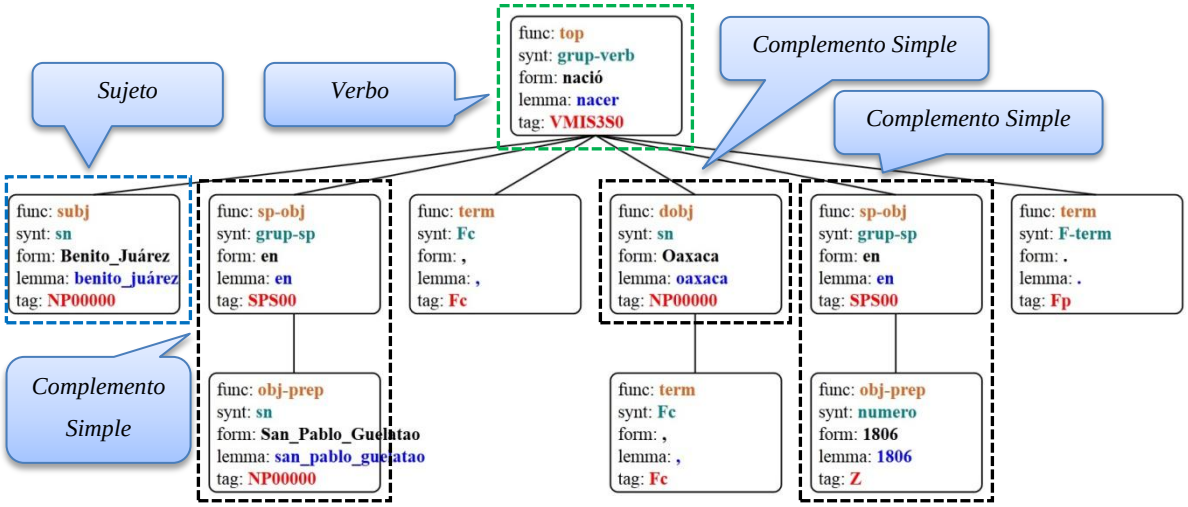

Figura 4.5 Patrón sintáctico "*Básico*" en el árbol de dependencias de la oración "*Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806*".

El otro tipo de estructura en el patrón sintáctico básico es la compleja, en donde el nodo raíz es el verbo del hecho, y en los descendientes inmediatos se tiene un nodo que representa al sujeto y otro que podría ser algún otro patrón sintáctico. La Figura 4.6 muestra la estructura compleja del patrón.

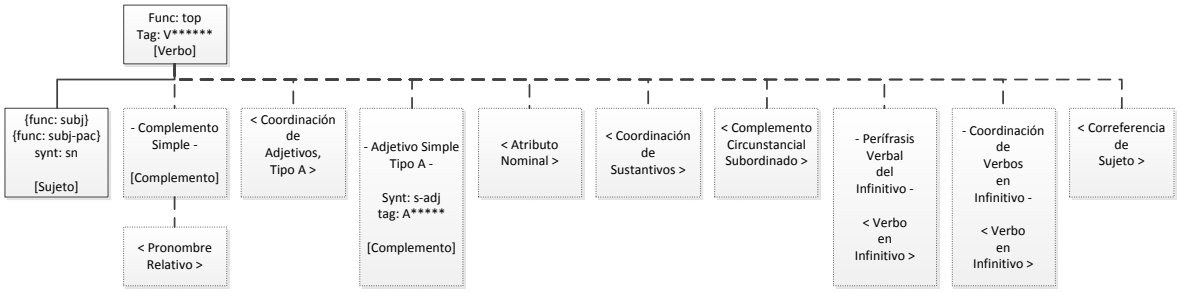

Figura 4.6 Diagrama del patrón sintáctico "*Básico*": Estructura Compleja.

Se puede ver que el *verbo* esta etiquetado con {func: top / tag: V******}, el *sujeto* con {func: subj} ó {func: subj-pac} y {synt: sn}; y el *complemento* podría encontrarse en algún otro patrón sintáctico donde se aplicaría otra heurística para extraer los hechos.

La forma de trabajar de la heurística es la siguiente: toma el *verbo* del nodo raíz para el hecho, luego en sus nodos descendientes busca el *sujeto*, posteriormente busca el *complemento* en algún otro patrón sintáctico.

En la Tabla 4.3 se describe el algoritmo de manera detallada.

Tabla 4.3 Algoritmo de la heurística "*Básica*".

```
1   Básica:
2   Parámetro de entrada: Apuntador a nodo raíz.
3   Asignar: Verbo = Valor de {form:} de nodo raíz.
4   Buscar si existe nodo Sujeto en los hijos del nodo raíz, debe tener
    las etiquetas {func: subj / synt: sn} ó { func: subj-pac / synt: sn }
5   Si existe, extraer ese nodo y sus descendientes como Sujeto.
6         Recorrer todos los hijos del nodo raíz y revisar:
7               Si tiene etiquetas de Complemento Simple.
8                     Buscar si existe un nodo Pronombre Relativo en la
                      rama del hijo, debe tener las etiquetas {func:
                      subord-mod / synt: subord-rel}.
9                     Si existe, extraer el nodo hijo hasta antes del nodo
                      Pronombre Relativo como Complemento.
10                          Construir: Hecho = [Sujeto] + [Verbo] +
                            [Complemento].
11                          Llamar algoritmo de la heurística Pronombre
                            Relativo (Apuntador a Pronombre Relativo,
                            Complemento).
12                    Sino existe, extraer el nodo hijo y sus
                      descendientes como Complemento.
13                          Construir: Hecho = [Sujeto] + [Verbo] +
                            [Complemento].
14              Si tiene la etiqueta {synt: s-adj}, revisar:
15                    Si tiene la etiqueta {synt: CC}.
16                          Llamar algoritmo de la heurística Coordinación
                            de Adjetivos, tipo A (Apuntador a hijo,
                            Sujeto, Verbo).
17                    Sino la tiene, extraer el nodo hijo y sus
                      descendientes como Complemento.
18                          Construir: Hecho = [Sujeto] + [Verbo] +
                            [Complemento].
19              Si tiene las etiquetas {func: att / synt: sn}:
20                    Llamar algoritmo de la heurísitica Atributo Nominal
                      (Apuntador a hijo, Sujeto, Verbo).
21              Si tiene las etiquetas {synt: coor-n / tag: CC}:
22                    Llamar algoritmo de la heurísitica Coordinación de
                      Sustantivos (Apuntador a hijo, Sujeto, Verbo).
23              Si tiene las etiquetas {func: cc / synt: subord / tag:
                CS}:
24                    Llamar algoritmo de la heurísitica Complemento
                      Circunstancial Subordinado (Apuntador a hijo,
                      Sujeto).
25              Si tiene las etiquetas {synt: grup-verb-inf}:
26                    Llamar algoritmo de la heurísitica Verbo en
                      Infinitivo (Apuntador a hijo, Sujeto).
27              Si tiene las etiquetas {synt: grup-sp-inf}:
28                    Llamar algoritmo de la heurísitica Correferencia de
                      Sujeto (Apuntador a hijo, Sujeto).
29  Fin
```

4.5.6 Heurística: Coordinación de Verbos

"Las oraciones coordinadas se encuentran unidas mediante una conjunción coordinante. Éstas pueden ser copulativas o disyuntivas" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Las conjunciones pueden representarse con "y, e, ni" y las disyuntivas con "o, u".

Aquí las palabras que se coordinan son palabras, en este, verbos. Esta heurística se aplica a los árboles que tienen el patrón sintáctico donde la raíz representa una conjunción y sus hijos son verbos, es decir, existe coordinación de verbos. La estructura del patrón puede ser simple o compleja.

En la estructura simple el nodo raíz representa la conjunción y sus hijos son verbos, en el primer hijo verbo en sus descendientes se encuentra el sujeto para los hechos que se extraen, cada hijo verbo contiene un complemento simple. La Figura 4.7 muestra la estructura simple del patrón.

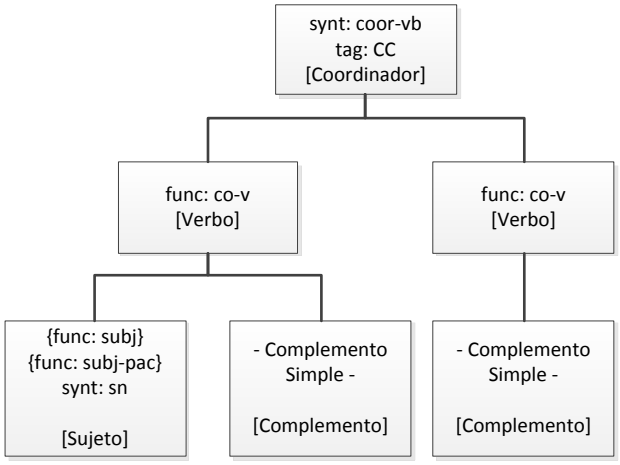

Figura 4.7 Diagrama del patrón sintáctico "*Coordinación de Verbos*": Estructura Simple.

La Figura 4.8 muestra un ejemplo del patrón sintáctico "Coordinación de Verbos" de estructura simple.

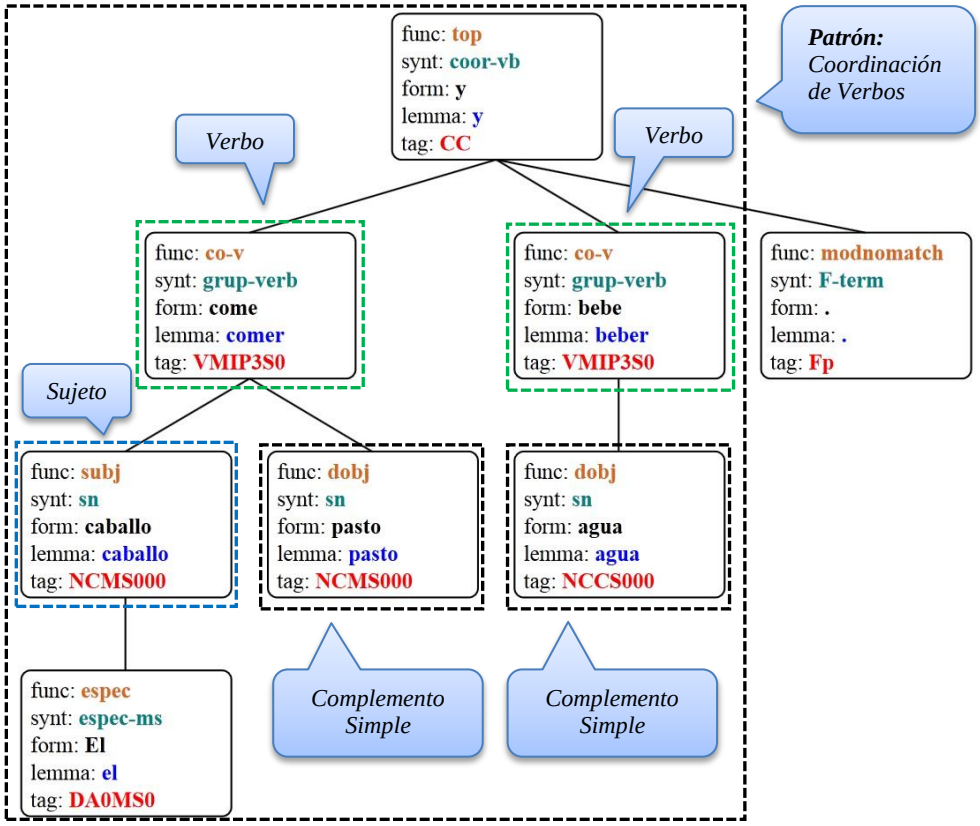

Figura 4.8 Patrón sintáctico "*Coordinación de Verbos*" en el árbol de dependencias de la oración "*El caballo come pasto y bebe agua*".

El otro tipo de estructura en el patrón sintáctico "Coordinación de Verbos" es la compleja, donde el nodo raíz representa la conjunción y su primer hijo es un verbo que tiene como descendiente el sujeto para los hechos que se extraen y como complemento tiene algún otro patrón sintáctico, los demás hijos de la raíz es algún otro patrón sintáctico. La Figura 4.9 muestra la estructura compleja del patrón.

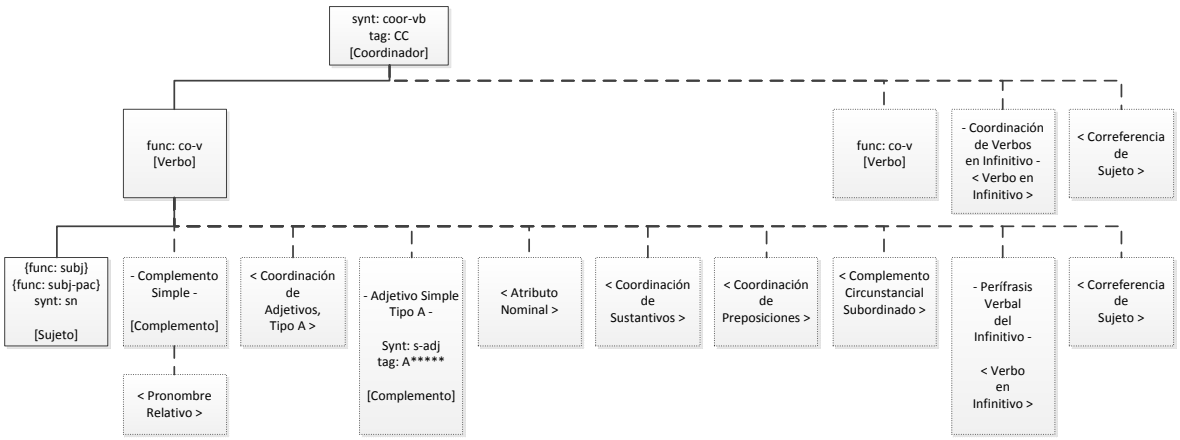

Figura 4.9 Diagrama del patrón sintáctico "*Coordinación de Verbos*": Estructura Compleja.

La forma de trabajar de la heurística es la siguiente: del primer hijo verbo se busca el sujeto para todos los hechos, el verbo coordinado se toma como verbo del hecho, si el complemento es simple se toma como complemento para el hecho, pero si existe algún patrón sintáctico se llama la heurística indicada.

En la Tabla 4.4 se describe el algoritmo de manera detallada.

Tabla 4.4 Algoritmo de la heurística "Coordinación de Verbos".

```
1   Coordinación de Verbos:
2   Parámetro de entrada: Apuntador a nodo raíz, tiene las etiquetas
    {synt: coor-vb / tag: CC}.
3   Recorrer todos los hijos (nA) del nodo raíz y revisar:
4         Si tiene la etiqueta {func: co-v}.
5               Asignar: Verbo = Valor de {form:} de nA.
6               Si es primer nA con etiqueta {func: co-v}.
7                     Extraer Sujeto desde los hijos de nA.
8               Recorrer todos los hijos (nB) de nA.
9                     Si tiene etiquetas de Complemento Simple.
10                          Buscar si existe un nodo Pronombre Relativo en
                            la rama de nB, debe tener las etiquetas {func:
                            subord-mod / synt: subord-rel}.
11                          Si existe, extraer nB hasta antes del nodo
                            Pronombre Relativo como Complemento.
12                                Construir: Hecho = [Sujeto] + [Verbo] +
                                  [Complemento].
13                                Llamar algoritmo de la heurística
                                  Pronombre Relativo (Apuntador a nB,
                                  Complemento).
14                          Sino existe, extraer nB y sus descendientes
                            como Complemento.
15                                Construir: Hecho = [Sujeto] + [Verbo] +
                                  [Complemento].
16                    Si tiene la etiqueta {synt: s-adj}, revisar:
17                          Si tiene la etiqueta {synt: CC}.
18                                Llamar algoritmo de la heurística
                                  Coordinación de Adjetivos, tipo A
                                  (Apuntador a nB, Sujeto, Verbo).
19                          Sino la tiene, extraer a nB y sus
                            descendientes como Complemento.
20                                Construir: Hecho = [Sujeto] + [Verbo] +
                                  [Complemento].
21                    Si tiene las etiquetas {func: att / synt: sn}:
22                          Llamar algoritmo de la heurística Atributo
                            Nominal (Apuntador a nB, Sujeto, Verbo).
23                    Si tiene las etiquetas {synt: coor-n / tag: CC}:
24                          Llamar algoritmo de la heurística Coordinación
                            de Sustantivos (Apuntador a nB, Sujeto,
                            Verbo).
25                    Si tiene las etiquetas {synt: coor-sp / tag: CC}:
26                          Llamar algoritmo de la heurística
                            Coordinación de Preposiciones (Apuntador a nB,
                            Sujeto, Verbo).
27                    Si tiene las etiquetas {func: cc / synt: subord /
                      tag: CS}:
28                          Llamar algoritmo de la heurística Complemento
                            Circunstancial Subordinado (Apuntador a nB,
                            Sujeto).
29                    Si tiene las etiquetas {synt: grup-verb-inf}:
30                          Llamar algoritmo de la heurística
                            Verbo en Infinitivo (Apuntador a nB, Sujeto).
31                    Si tiene las etiquetas {synt: grup-sp-inf}:
32                          Llamar algoritmo de la heurística
                            Correferencia de Sujeto (Apuntador a nB,
                            Sujeto).
33        Si tiene las etiquetas {synt: grup-verb-inf}:
34              Llamar algoritmo de la heurística Verbo en Infinitivo
                (Apuntador a nB, Sujeto).
```

```
35        Si tiene las etiquetas {synt: grup-sp-inf}:
36              Llamar algoritmo de la heurística Correferencia de Sujeto
                (Apuntador a nB, Sujeto).
37  Fin.
```

4.5.7 Heurística: Pronombre Relativo

"Los pronombres relativos hacen referencia a alguien o a algo que se ha mencionado antes en el discurso o que ya es conocido por los interlocutores. Los pronombres relativos, funcionan, en la mayor parte de los casos, como elementos de subordinación de oraciones. Los pronombres relativos son: *que*, *quien*, *quienes*, *cual*, *cuales*, *cuanto*, *cuantos*, *cuanta*, *cuantas*" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Así que esta heurística se aplica a los árboles que tienen el patrón sintáctico donde aparece un pronombre relativo. La Figura 4.10 muestra el patrón sintáctico, donde se puede ubicar como esta etiquetado el pronombre, de quién desciende y quiénes podrían ser sus descendientes.

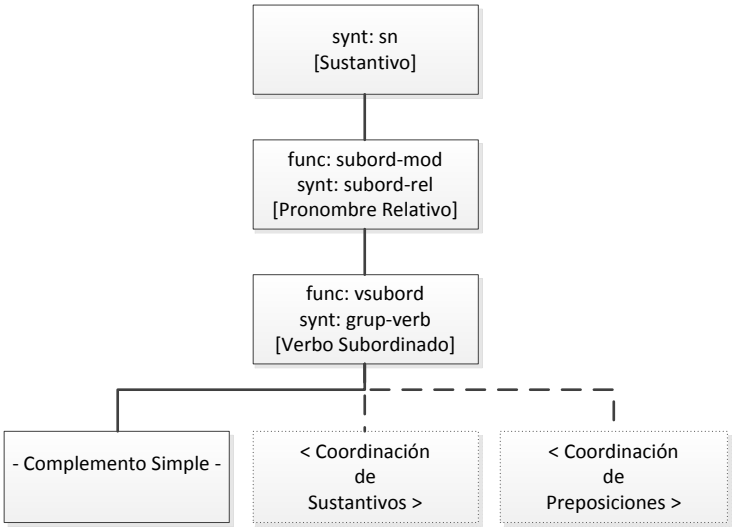

Figura 4.10 Diagrama del patrón sintáctico "¨*Pronombre Relativo*".

La Figura 4.11 muestra un ejemplo del patrón sintáctico "Pronombre Relativo".

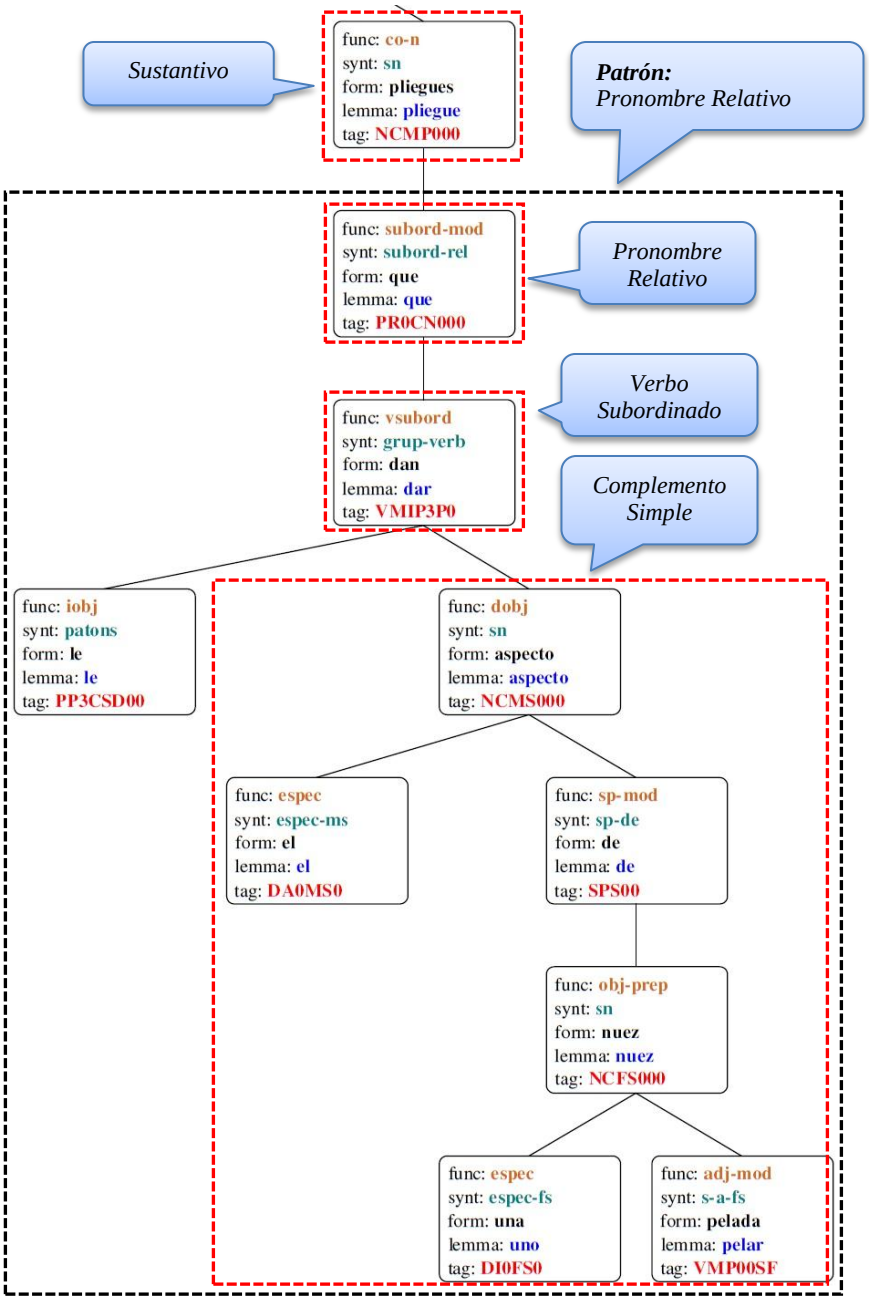

Figura 4.11 Patrón sintáctico "*Pronombre Relativo*" en el árbol de dependencias de la oración "*El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada*".

La forma de trabajar de la heurística es la siguiente: como el pronombre relativo hace referencia a alguien o a algo que se ha mencionado antes, entonces se busca al sujeto (sustantivo, pronombre personal) para el hecho en la parte inmediata que antecede al

pronombre relativo. El verbo de este hecho se encuentra localizado después del pronombre relativo y en sus descendientes del verbo se puede encontrar el complemento.

En la Tabla 4.5 se describe el algoritmo de manera detallada.

Tabla 4.5 Algoritmo de la heurística "*Pronombre Relativo*".

```
1   Pronombre Relativo:
2   Parámetros de entrada: Sujeto y un apuntador al nodo pronombre
    relativo (nPR), tiene las etiquetas
    {func: subord-mod / synt: subord-rel}.
3   Buscar nodo verbo subordinado (nVS) en los descendientes de nPR,
    tiene las etiquetas
    {func: vsubord / synt: grup-verb}.
4   Asignar: Verbo = Valor de {form:} de nVS.
5   Recorrer todos los hijos de nVS, y revisar:
6         Si tiene etiquetas de Complemento Simple, extraer ese nodo y
          sus descendientes como Complemento.
7               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
8         Si tiene las etiquetas {synt: coor-n / tag: CC}
9               Llamar algoritmo de la heurística
                Coordinación de Sustantivos (Apuntador a hijo, Sujeto,
                Verbo).
10        Si tiene las etiquetas {synt: coor-sp / tag: CC}.
11              Llamar algoritmo de la heurística
                Coordinación de Preposiciones (Apuntador a hijo, Sujeto,
                Verbo).
12  Fin.
```

En la Tabla 4.6 se puede ver el hecho que se ha extraído de una oración, con el algoritmo de la heurística "Pronombre Relativo".

Tabla 4.6 Hechos extraídos con la heurística "*Pronombre Relativo*" de la oración "*El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada*".

No.	Sujeto	Verbo	Complemento
1	pliegues	dan	aspecto de nuez pelada

4.5.8 Heurística: Coordinación de Adjetivos, tipo A

La coordinación de adjetivos de este tipo, se presenta en oraciones que tienen complemento predicativo o atributo, dicho "predicado informa sobre cualidades, atributos o peculiaridades del sujeto" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Aparece en oraciones con los verbos copulativos *ser* y *estar*, pero también puede presentarse con verbos de significado pleno.

Por lo tanto esta heurística se aplica a los árboles que presentan el patrón sintáctico que contiene coordinación de adjetivos que dependen de un verbo. La Figura 4.12 muestra el patrón sintáctico de forma general, en ella se puede ubicar como esta etiquetada la raíz de la coordinación, de quien desciende y quiénes son sus descendientes.

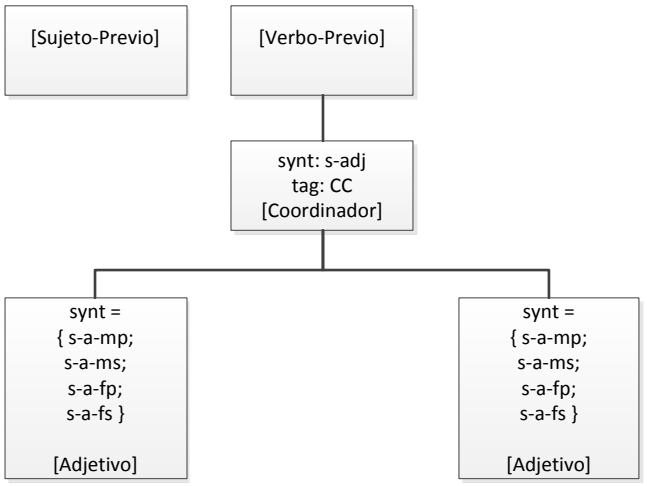

Figura 4.12 Diagrama del patrón sintáctico "*Coordinación de adjetivos, tipo A*".

La Figura 4.13 muestra un ejemplo del patrón sintáctico "Coordinación de Adjetivos, tipo A".

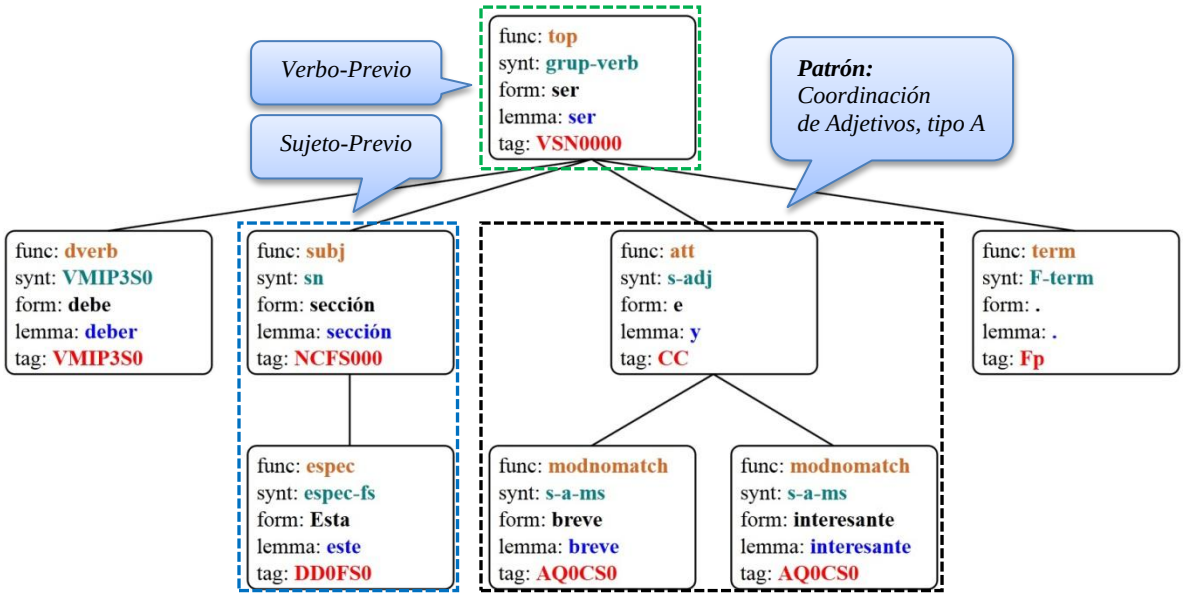

Figura 4.13 Patrón sintáctico "*Coordinación de Adjetivos, tipo A*" en el árbol de dependencias de la oración "*Esta sección debe ser breve e interesante*".

La forma de trabajar de la heurística es la siguiente: el sujeto y verbo ya se han extraído previamente, ahora se extraen los adjetivos coordinados como complementos para un hecho por cada uno de ellos.

En la Tabla 4.7 se describe el algoritmo de manera detallada.

Tabla 4.7 Algoritmo de la heurística "*Coordinación de Adjetivos, tipo A*".

```
1   Coordinación de Adjetivos, tipo A:
2   Parámetros de entrada: Sujeto, Verbo y un apuntador al nodo
    coordinador de adjetivos, tiene las etiquetas
    {synt: s-adj / tag: CC}.
3   Recorrer todos los hijos del nodo coordinador y revisar:
4         Si tiene la etiqueta
          {synt:} = {s-a-ms ó s-a-mp ó s-a-fs ó s-a-fp}, extraer ese nodo
          y sus descendientes como Complemento.
5               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
6   Fin.
```

En la Tabla 4.8 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Coordinacion de adjetivos, tipo A".

Tabla 4.8 Hechos extraídos con la heurística "*Coordinación de Adjetivos, tipo A*" de la oración "*Esta sección debe ser breve e interesante*".

No.	Sujeto	Verbo	Complemento
1	sección	ser	breve
2	sección	ser	interesante

4.5.9 Heurística: Coordinación de Adjetivos, tipo B

Aquí el complemento predicativo o atributo que se describe en la heurística "Coordinación de adjetivos, tipo A", se forma por un sustantivo más un adjetivo que actúa como modificador de ese sustantivo.

Así que esta heurística se aplica a los árboles que presentan el patrón que contiene un predicado con un sustantivo modificado por una coordinación de adjetivos. La Figura 4.14 muestra el patrón sintáctico, en ella se puede ubicar el sustantivo y el nodo coordinador de los adjetivos que lo modifican.

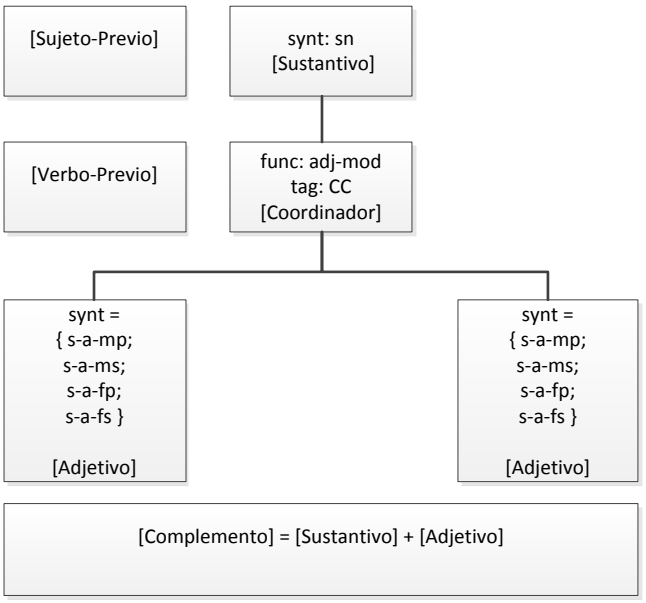

Figura 4.14 Diagrama del patrón sintáctico "*Coordinación de Adjetivos, tipo B*".

La Figura 4.15 muestra un ejemplo del patrón sintáctico "Coordinación de Adjetivos, tipo B".

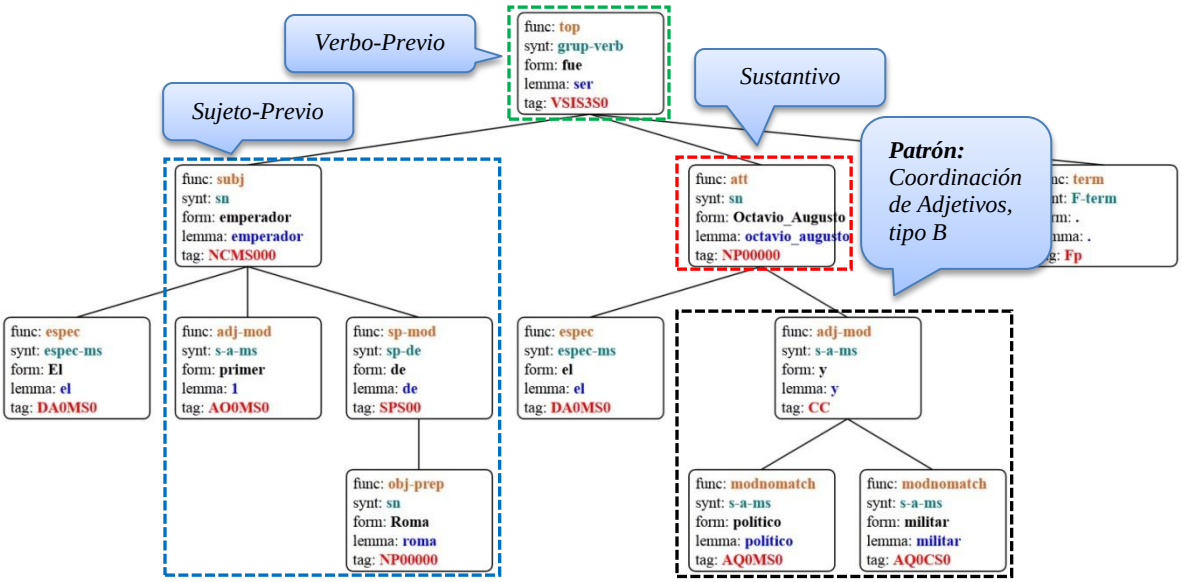

Figura 4.15 Patrón sintáctico "*Coordinación de adjetivos, tipo B*" en el árbol de dependencias de la oración "*El primer emperador de Roma fue el político y militar Octavio Augusto*".

La forma de trabajar de la heurística es la siguiente: el sujeto, verbo sustantivo se extraen previamente, ahora se extraen los adjetivos coordinados, cada adjetivo extraído se adjunta al sustantivo para formar un hecho por cada uno de ellos.

En la Tabla 4.9 se describe el algoritmo de manera detallada.

Tabla 4.9 Algoritmo de la heurística "*Coordinación de Adjetivos, tipo B*".

```
1   Coordinación de Adjetivos, tipo B:
2   Parámetros de entrada: Sujeto, Verbo, Sustantivo y un apuntador al
    nodo coordinador de adjetivos, tiene las etiquetas
    {func: adj-mod / tag: CC}.
3   Recorrer todos los hijos del nodo coordinador y revisar:
4         Si tiene la etiqueta
          {synt:} = {s-a-ms ó s-a-mp ó s-a-fs ó s-a-fp}, extraer ese nodo
          y sus descendientes como Adjetivo.
5               Formar: Complemento = [Sustantivo] + [Adjetivo].
6               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
7   Fin.
```

En la Tabla 4.10 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Coordinacion de adjetivos, tipo B".

Tabla 4.10 Hechos extraídos con la heurística "*Coordinación de Adjetivos, tipo B*" de la oración "*El primer emperador de Roma fue el político y militar Octavio Augusto*".

No.	Sujeto	Verbo	Complemento
1	primer emperador de Roma	fue	Octavio_Augusto político
2	primer emperador de Roma	fue	Octavio_Augusto militar

4.5.10 Heurística: Atributo Nominal

"El predicado nominal se construye con verbos copulativos, los cuales se caracterizan por no tener un significado pleno; se acompañan de un adjetivo, un sustantivo o una oración, estos elementos son los que aportan la información del predicado. En estas oraciones el verbo sólo cumple la función de enlazar el sujeto con el predicado, de ahí que reciba el nombre de copulativo. Los verbos copulativos más comunes son *ser* y *estar*" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Esta heurística se aplica a los árboles que presentan el patrón donde se tiene un verbo copulativo necesariamente, lo que la distingue de las heurísticas de coordinación de adjetivos; luego un sustantivo que puede ser modificado por un adjetivo o varios, o ese sustantivo puede ser el sujeto para otro hecho cuando se presenta el patrón "Pronombre

Relativo". La Figura 4.16 muestra el patrón sintáctico, en ella se puede ver al sustantivo que depende de un verbo copulativo y otros patrones sintácticos que pueden presentarse.

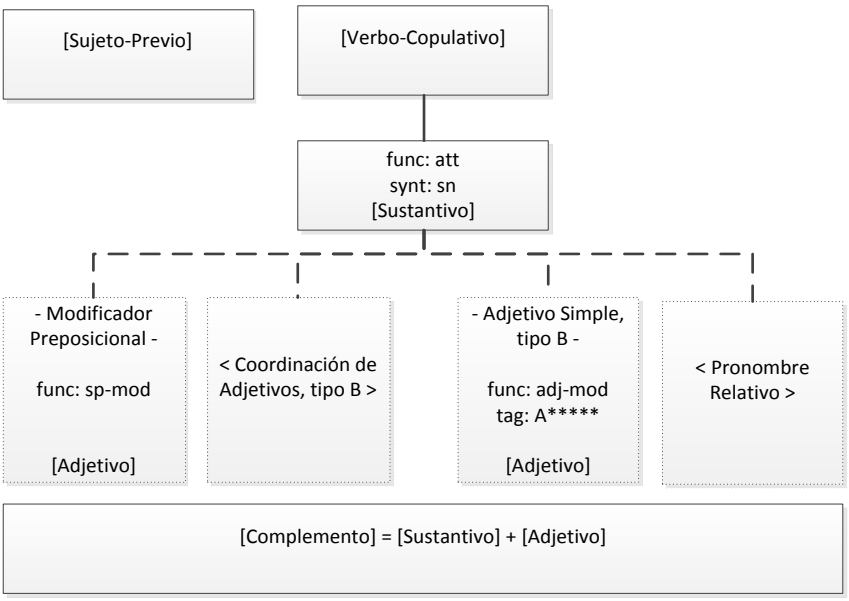

Figura 4.16 Diagrama del patrón sintáctico "*Atributo Nominal*".

La Figura 4.17 muestra un ejemplo del patrón sintáctico "Atributo Nominal".

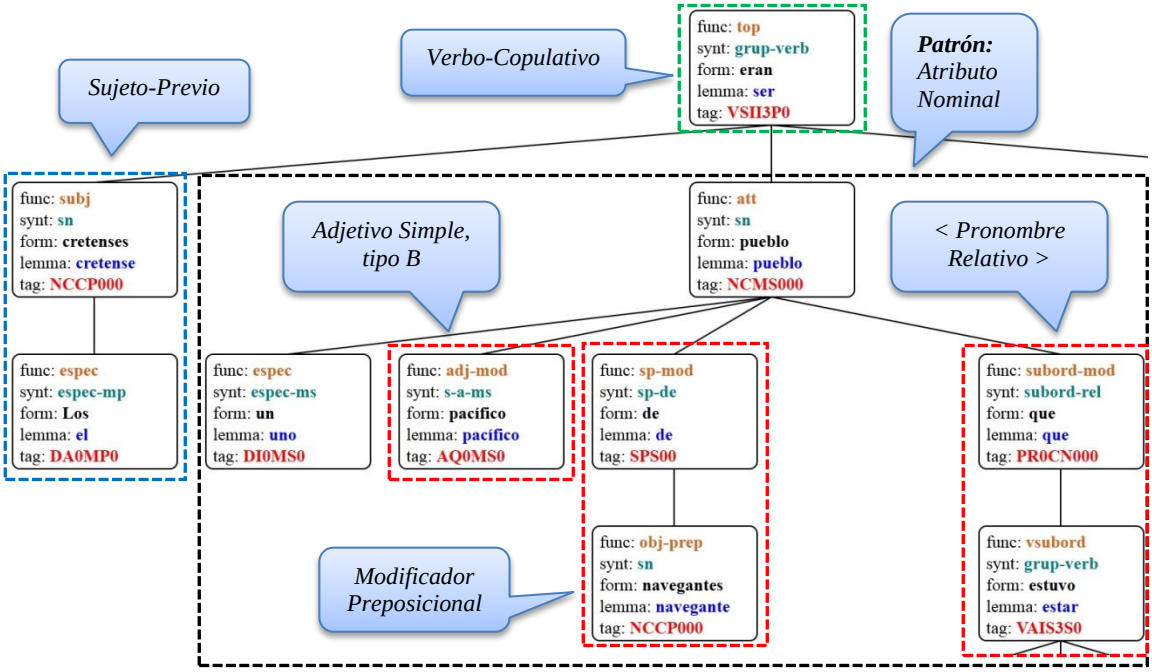

Figura 4.17 Patrón sintáctico "*Atributo Nominal*" en el árbol de dependencias de la oración "*Los cretenses eran un pueblo pacífico de navegantes que estuvo en contacto con Egipto y Medio Oriente*".

La forma de trabajar de la heurística es la siguiente: el sujeto ya se ha extraído previamente, el verbo copulativo se toma para el hecho, y los complementos se obtienen revisando los descendientes del nodo sustantivo que depende directamente del verbo copulativo.

En la Tabla 4.11 se describe el algoritmo de manera detallada.

Tabla 4.11 Algoritmo de la heurística "*Atributo Nominal*".

```
1   Atributo Nominal:
2   Parámetros de entrada: Sujeto, Verbo, y un apuntador al nodo
    sustantivo, tiene las etiquetas {func: att / tag: sn}.
3   Asignar: Sustantivo = Valor de {form:} del nodo sustantivo.
4   Asignar: SujetoPR = [Sustantivo] // Sujeto para Pronombre Relativo.
5   Recorrer todos los hijos del nodo sustantivo y revisar:
6         Si tiene la etiqueta {func: sp-mod}, extraer ese nodo y sus
          descendientes como Complemento.
7               Actualizar: SujetoPR = [SujetoPR] + [Complemento].
8               Actualizar: Complemento = [Sustantivo] + [Complemento].
9               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
10        Si tiene la etiquetas {func: adj-mod / tag: CC}.
11              Llamar al algoritmo de la heurística
                Coordinación de Adjetivos, tipo B (Apuntador a hijo,
                Sujeto, Verbo, Sustantivo)
12        Si tiene las etiquetas {func: adj-mod / tag: A*****}, extraer
          ese nodo y sus descendientes como Complemento.
13              Actualizar: SujetoPR = [SujetoPR] + [Complemento].
14              Actualizar: Complemento = [Sustantivo] + [Complemento].
15              Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
16        Si tiene las etiquetas {func: subord-mod /synt: subord-rel}.
17              Construir: Hecho = [Sujeto] + [Verbo] + [Sustantivo].
18              Llamar al algoritmo de la heurística
                Pronombre Relativo (Apuntador a hijo, SujetoPR).
19  Fin.
```

En la Tabla 4.12 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Atributo Nominal".

Tabla 4.12 Hechos extraídos con la heurística "*Atributo Nominal*" de la oración "*Los cretenses eran un pueblo pacífico de navegantes que estuvo en contacto con Egipto y Medio Oriente*".

No.	Sujeto	Verbo	Complemento
1	Cretenses	eran	pueblo pacífico
2	Cretenses	eran	pueblo de navegantes
3	Pueblo pacífico de navegantes	estuvo	en contacto
4	Pueblo pacífico de navegantes	estuvo	con Egipto
5	Pueblo pacífico de navegantes	Estuvo	Medio_Oriente

4.5.11 Heurística: Coordinación de Sustantivos

Se presenta cuando las palabras que se coordinan en la oración son sustantivos, los cuales se encuentran en el predicado. La coordinación puede especificarse con "y, o, u".

Así que esta heurística se aplica a los árboles que presentan el patrón donde aparece uno o más sustantivos coordinados. La Figura 4.18 muestra el patrón sintáctico, en ella se puede ver que el nodo coordinador depende de un verbo y sus descendiente son sustantivos (synt: sn), además se puede ver que estos sustantivos pueden presentar otros patrones sintácticos.

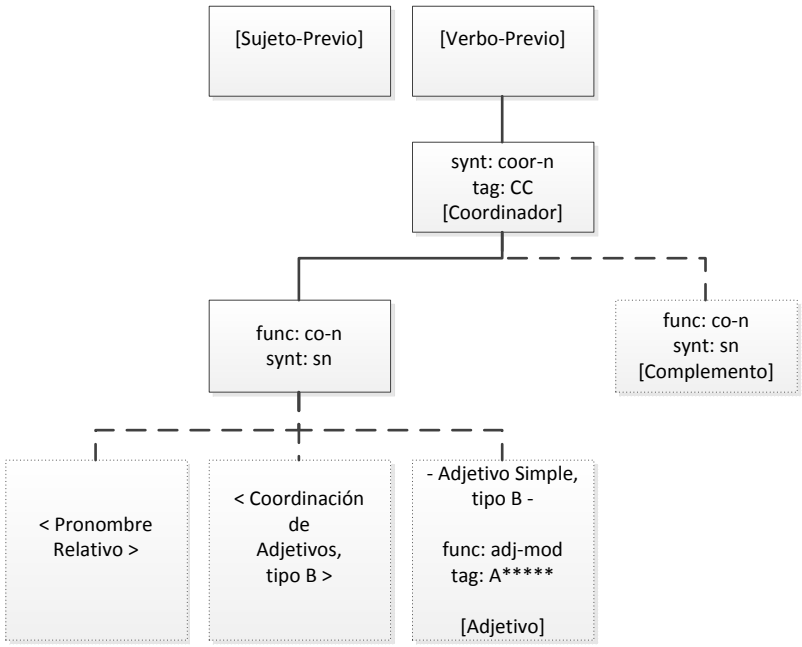

Figura 4.18 Diagrama del patrón sintáctico "*Coordinación de Sustantivos*".

La Figura 4.19 muestra un ejemplo del patrón sintáctico "Coordinación de Sustantivos".

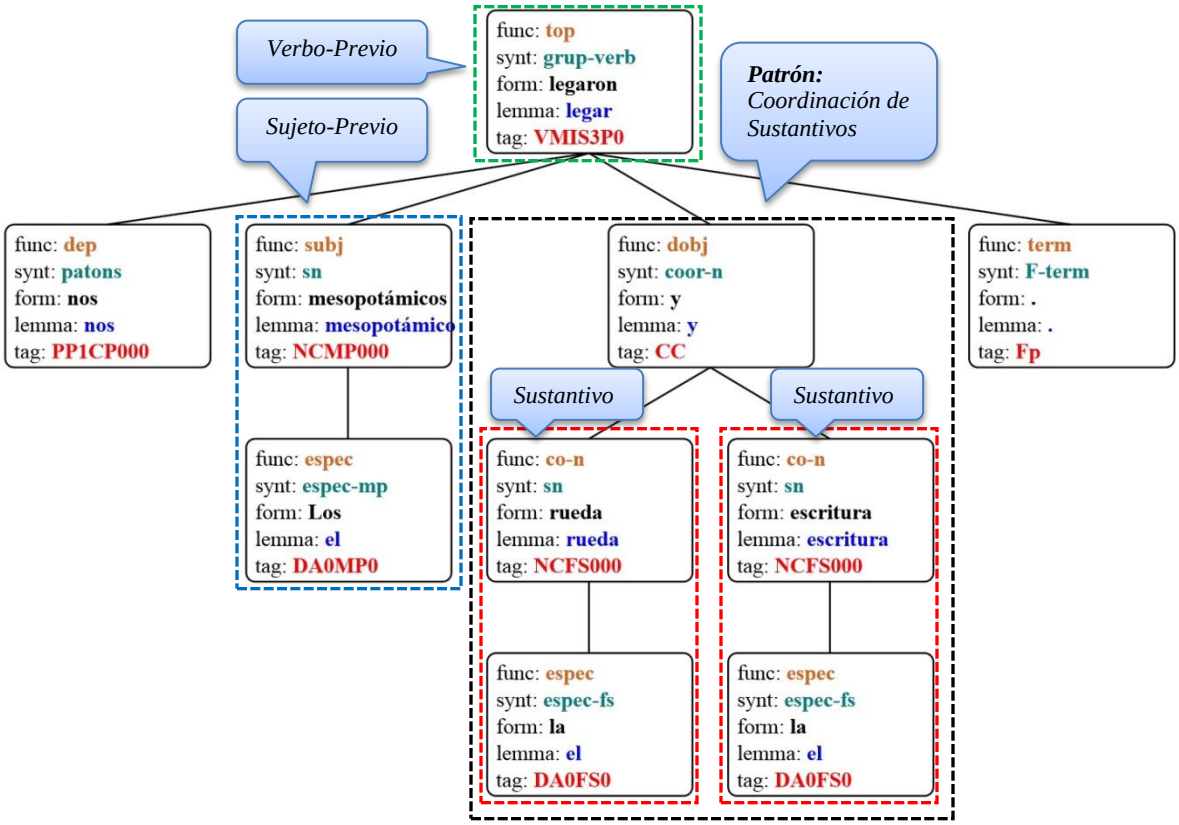

Figura 4.19 Patrón sintáctico "*Coordinación de Sustantivos*" en el árbol de dependencias de la oración "*Los mesopotámicos nos legaron la rueda y la escritura*".

La forma de trabajar de la heurística es la siguiente: el sujeto y verbo se obtienen previamente, ahora se obtienen los sustantivos, cada sustantivo es el complemento para un hecho.

En la Tabla 4.13 se describe el algoritmo de manera detallada.

Tabla 4.13 Algoritmo de la heurística "*Coordinación de Sustantivos*".

```
1   Coordinación de Sustantivos:
2   Parámetros de entrada: Sujeto, Verbo, y un apuntador al nodo
    coordinador de sustantivos, tiene las etiquetas
    {synt: coor-n / tag: CC}.
3   Recorrer todos los hijos (HijoA) del nodo coordinador y procesar los
    que tienen la etiqueta
    {func: co-n}. Revisar a los hijos (HijoB) de HijoA, lo siguiente:
4         Si HijoB tiene las etiquetas
          {func: subord-mod / synt: subord-rel}.
5               Asignar: Complemento = Valor de {form:} de HijoA.
6               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
7               Llamar al algoritmo de la heurística
                Pronombre Relativo (Apuntador HijoB, Complemento).
8         Si HijoB tiene las etiquetas {func: adj-mod / tag: CC}.
9               Asignar: Sustantivo = Valor de {form:} de HijoA.
10              Llamar algoritmo de la heurística
                Coordinación de Adjetivos, tipo B (Apuntador a HijoB,
                Sujeto, Verbo, Sustantivo).
11        Si HijoB tiene las etiquetas {func: adj-mod / tag: A*****},
          extraer ese nodo y sus descendientes como Complemento.
12              Asignar: Sustantivo = Valor de {form:} de HijoA.
13              Construir: Hecho = [Sujeto] + [Verbo] + [ [Sustantivo] +
                [Complemento] ].
14        Sino // Ningún Si anterior.
15              Extraer como Complemento al nodo {HijoA}.
16              Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
17  Fin.
```

En la Tabla 4.14 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Coordinación de Sustantivos".

Tabla 4.14 Hechos extraídos con la heurística "*Coordinación de Sustantivos*" de la oración "*Los mesopotámicos nos legaron la rueda y la escritura*".

No.	Sujeto	Verbo	Complemento
1	mesopotámicos	legaron	rueda
2	mesopotámicos	legaron	escritura

4.5.12 Heurística: Coordinación de Preposiciones

Se presenta cuando las palabras que se coordinan en la oración son preposiciones, los cuales se encuentran en el predicado. La coordinación puede especificarse con "y, o, u".

Así que esta heurística se aplica a los árboles que presentan el patrón donde aparece una o más preposiciones coordinadas. La Figura 4.21 muestra el patrón sintáctico, en ella se puede ver que el nodo coordinador depende de un verbo y sus descendiente son preposiciones (func: co-sp) y que se toman como complemento.

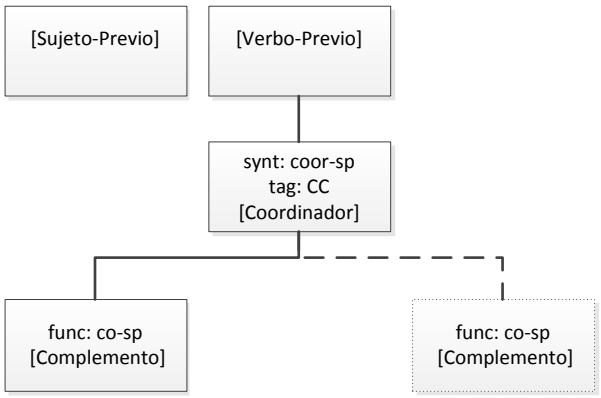

Figura 4.20 Diagrama del patrón sintáctico "*Coordinación de Preposiciones*".

La Figura 4.21 muestra un ejemplo del patrón sintáctico "Coordinación de Preposiciones".

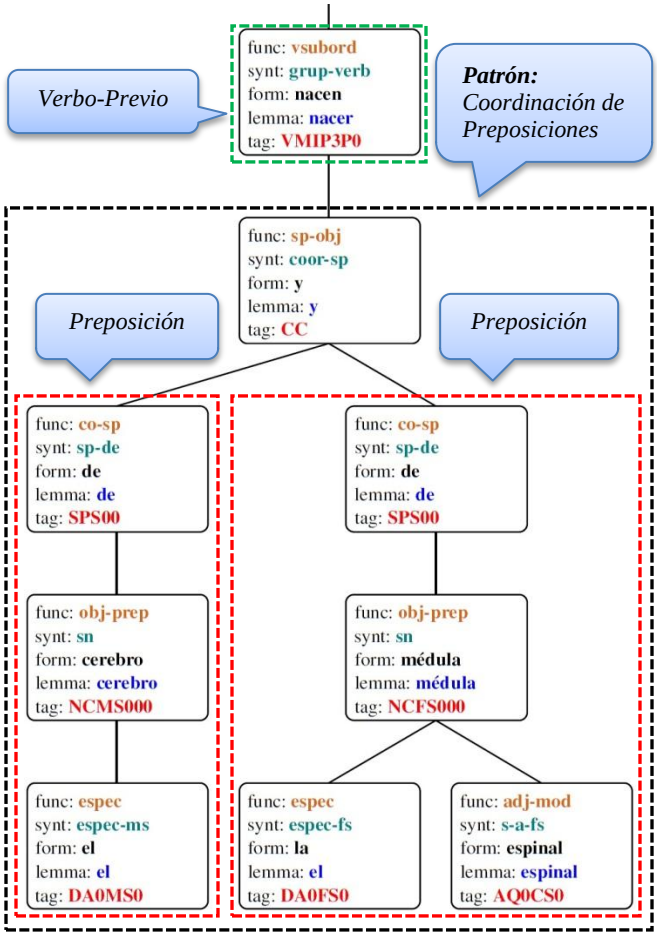

Figura 4.21 Patrón sintáctico "*Coordinación de Preposiciones*" en el árbol de dependencias de la oración "*El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas*".

La forma de trabajar de la heurística es la siguiente: el sujeto y verbo se obtienen previamente, ahora se obtienen las preposiciones, cada preposición es el complemento para un hecho.

En la Tabla 4.15 se describe el algoritmo de manera detallada.

Tabla 4.15 Algoritmo de la heurística "*Coordinador de Preposiciones*".

```
1   Coordinación de Preposiciones:
2   Parámetros de entrada: Sujeto, Verbo, y un apuntador al nodo
    coordinador de preposiciones, tiene las etiquetas
    {synt: coor-sp / tag: CC}.
3   Recorrer todos los hijos del nodo coordinador y revisar:
4         Si tiene la etiqueta {func: co-sp}, extraer ese nodo y sus
          descendientes como Complemento.
5               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
6   Fin.
```

En la Tabla 4.16 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Coordinación de Preposiciones".

Tabla 4.16 Hechos extraídos con la heurística "Coordinación de Preposiciones" de la oración "*El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas*".

No.	Sujeto	Verbo	Complemento
1	nervios	nacen	de cerebro
2	nervios	nacen	de médula espinal

4.5.13 Heurística: Complemento Circunstancial Subordinado

El complemento circunstancial "expresa la manera, el tiempo, el lugar y demás circunstancias en las que se realiza la acción del verbo" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Y se dice que es subordinado porque este tipo de complemento sirve para dar más detalles de la acción del verbo principal, así que para ello tienen su propio verbo. El complemento puede especificarse con "cuando, ya que".

Así que esta heurística se aplica a los árboles que presentan el patrón donde se especifica este complemento. La Figura 4.22 muestra el patrón sintáctico, en ella se puede ver el nodo que indica el complemento circunstancial subordinado, que depende de un verbo y que tiene como descendiente directo un verbo subordinado a él. En el verbo subordinado se puede encontrar un sujeto o sólo un complemento.

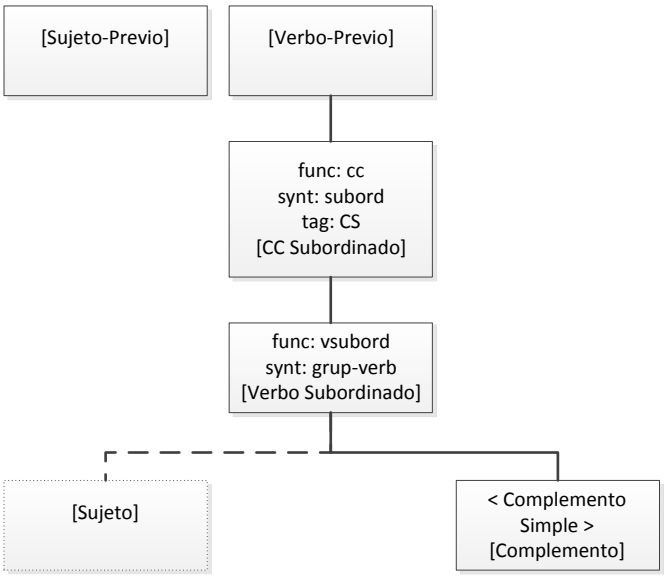

Figura 4.22 Diagrama del patrón sintáctico "*Complemento Circunstancial Subordinado*".

La Figura 4.23 muestra un ejemplo del patrón sintáctico "Complemento Circunstancial Subordinado".

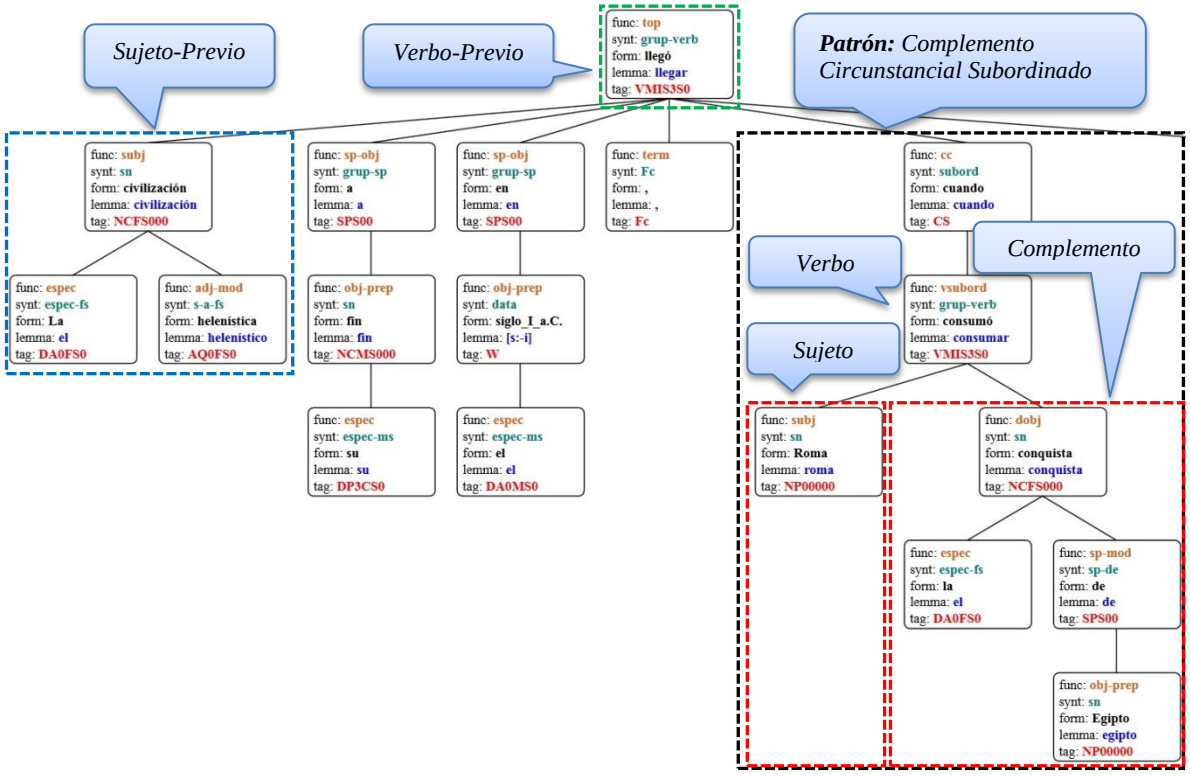

Figura 4.23 Patrón sintáctico "*Complemento Circunstancial Subordinado*" en el árbol de dependencias de la oración "*La civilización helenística llegó a su fin en el siglo I a.C., cuando Roma consumó la conquista de Egipto*".

La forma de trabajar de la heurística es la siguiente: el sujeto se obtiene previamente, el verbo subordinado del nodo indicador del complemento se toma para el hecho; luego en los descendientes del verbo subordinado se busca si existe un sujeto que se toma como sujeto para el hecho, sino existe el sujeto previo se mantiene como sujeto del hecho; en los descendientes del verbo subordinado se busca el complemento para el hecho.

En la Tabla 4.17 se describe el algoritmo de manera detallada.

Tabla 4.17 Algoritmo de la heurística "*Complemento Circunstancial Subordinado*".

```
1   Complemento Circunstancial Subordinado:
2   Parámetros de entrada: Sujeto y un apuntador al nodo CC Subordinado,
    tiene las etiquetas {func: cc / synt: subord / tag: CS}.
3   Buscar nodo verbo subordinado (nVS) en los descendientes del nodo CC
    Subordinado, debe tener las etiquetas
    {func: vsubord / synt: grup-verb}.
4         Asignar: Verbo = Valor de {form:} de nVS.
5         Buscar sujeto en los hijos de nVS, debe tener las etiquetas
          {func: subj / synt: sn} ó {func: subj-pac / synt: sn}.
6               Extraer ese nodo y sus descendientes como el nuevo Sujeto.
7         Recorrer todos los hijos de nVS y revisar:
8               Si tiene etiquetas de Complemento Simple, extraer ese nodo
                y sus descendientes como Complemento.
9                     Construir: Hecho = [Sujeto] + [Verbo] +
                      [Complemento].
10  Fin.
```

En la Tabla 4.18 se muestran un hecho extraído de una oración, con el algoritmo de la heurística "Complemento Circunstancial Subordinado".

Tabla 4.18 Hecho extraído con la heurística "*Complemento Circunstancial Subordinado*" de la oración "*La civilización helenística llegó a su fin en el siglo I a.C., cuando Roma consumó la conquista de Egipto*".

No.	Sujeto	Verbo	Complemento
1	Roma	consumó	conquista de Egipto

4.5.14 Heurística: Verbo en Infinitivo

El verbo en infinitivo se presenta en dos tipos de patrones sintácticos, como "Perífrasis Verbal" y "Coordinación de Verbos en Infinitivo".

4.5.14.1 Perífrasis verbal del Infinitivo

"Las perífrasis verbales son construcciones que se forman con dos o más verbos que, en ocasiones, pueden estar unidos por una palabra de enlace. El primer verbo se conjuga y el segundo se expresa por medio de una forma no personal, es decir, por un infinitivo, un gerundio o un participio, aunque también es posible encontrarlo conjugado" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000).

Las perífrasis normalmente tienen la siguiente forma: (verbo auxiliar) + (preposición o conjunción) + (infinitivo, gerundio o participio).

El patrón sintáctico "Perífrasis verbal del Infinitivo" se puede presentar en el patrón sintáctico "Básico" o en el patrón "Coordinación de Verbos".

Perífrasis Verbal del Infinitivo en el patrón "*Básico*"

En la Figura 4.24 se muestra este patrón, se puede ver un nodo descendiente de la raíz que contiene al verbo auxiliar, en la raíz del árbol se encuentra el verbo en participio, y un nodo descendiente de la raíz que contiene al verbo en infinitivo. El verbo en infinito es el que se toma como verbo para el hecho.

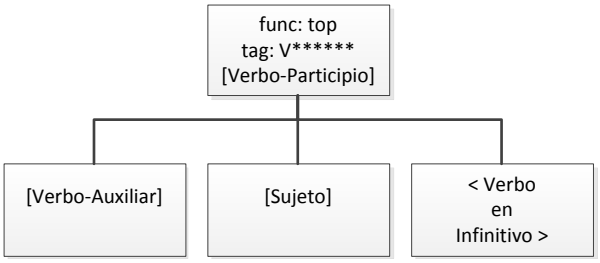

Figura 4.24 Diagrama del patrón sintáctico "*Perífrasis Verbal del Infinitivo*" en el patrón "*Básico*".

La Figura 4.25 muestra un ejemplo del patrón sintáctico "Perífrasis Verbal del Infinitivo", cuando se presenta en el patrón sintáctico "Básico".

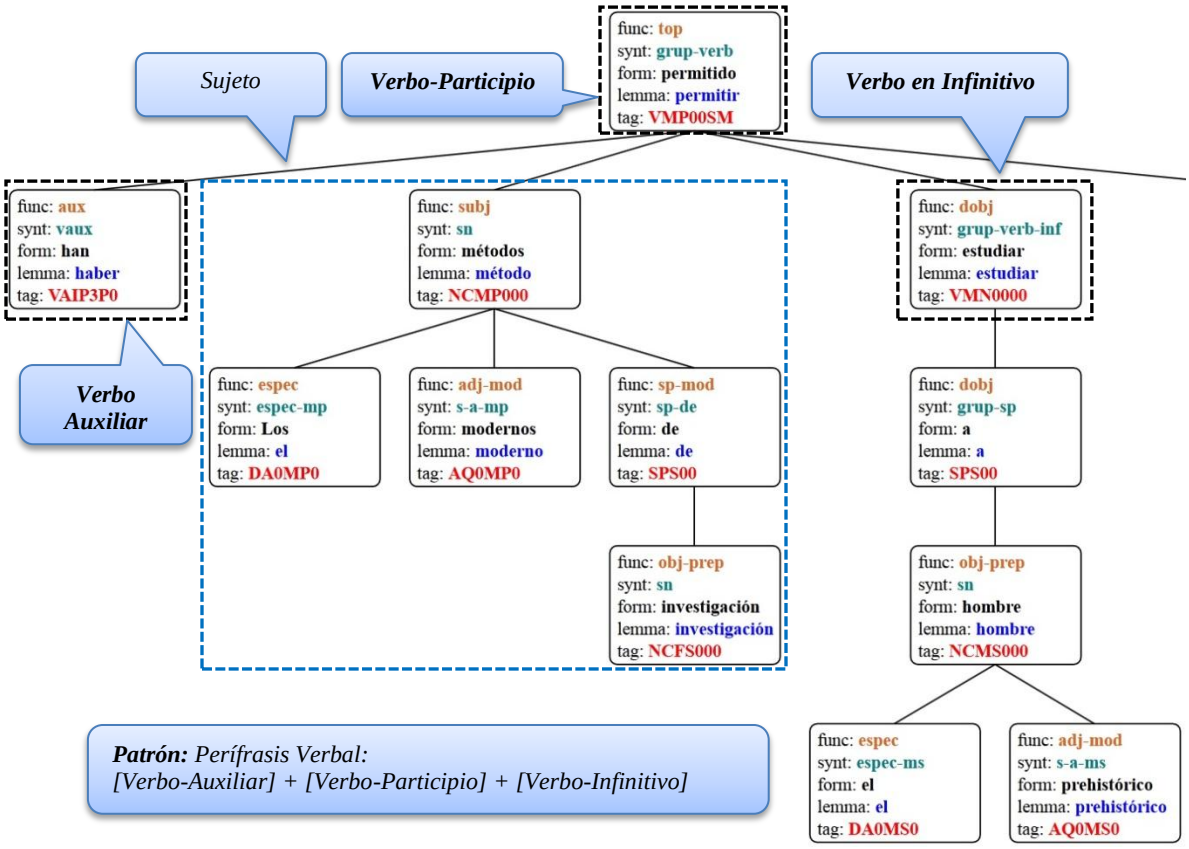

Figura 4.25 Patrón sintáctico "*Perífrasis Verbal del Infinitivo*" en el árbol de dependencias de la oración "*Los métodos modernos de investigación han permitido estudiar al hombre prehistórico*".

Perífrasis Verbal del Infinitivo en el patrón "*Coordinación de Verbos*"

En la Figura 4.26 se muestra este patrón, en este caso el verbo coordinado es el que representa al verbo en participio y tiene como descendientes al verbo auxiliar y al verbo en infinitivo. El verbo en infinito es el que se toma como verbo para el hecho.

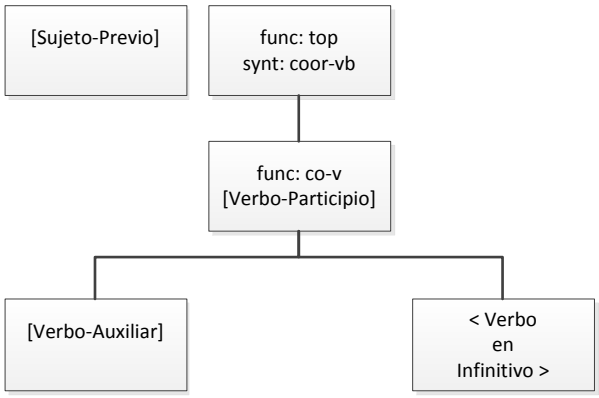

Figura 4.26 Diagrama del patrón sintáctico "*Perífrasis Verbal del Infinitivo*" en el patrón "*Coordinación de Verbos*".

La Figura 4.27 muestra un ejemplo del patrón sintáctico "Perífrasis Verbal del Infinitivo", cuando se presenta en el patrón sintáctico "Coordinación de Verbos".

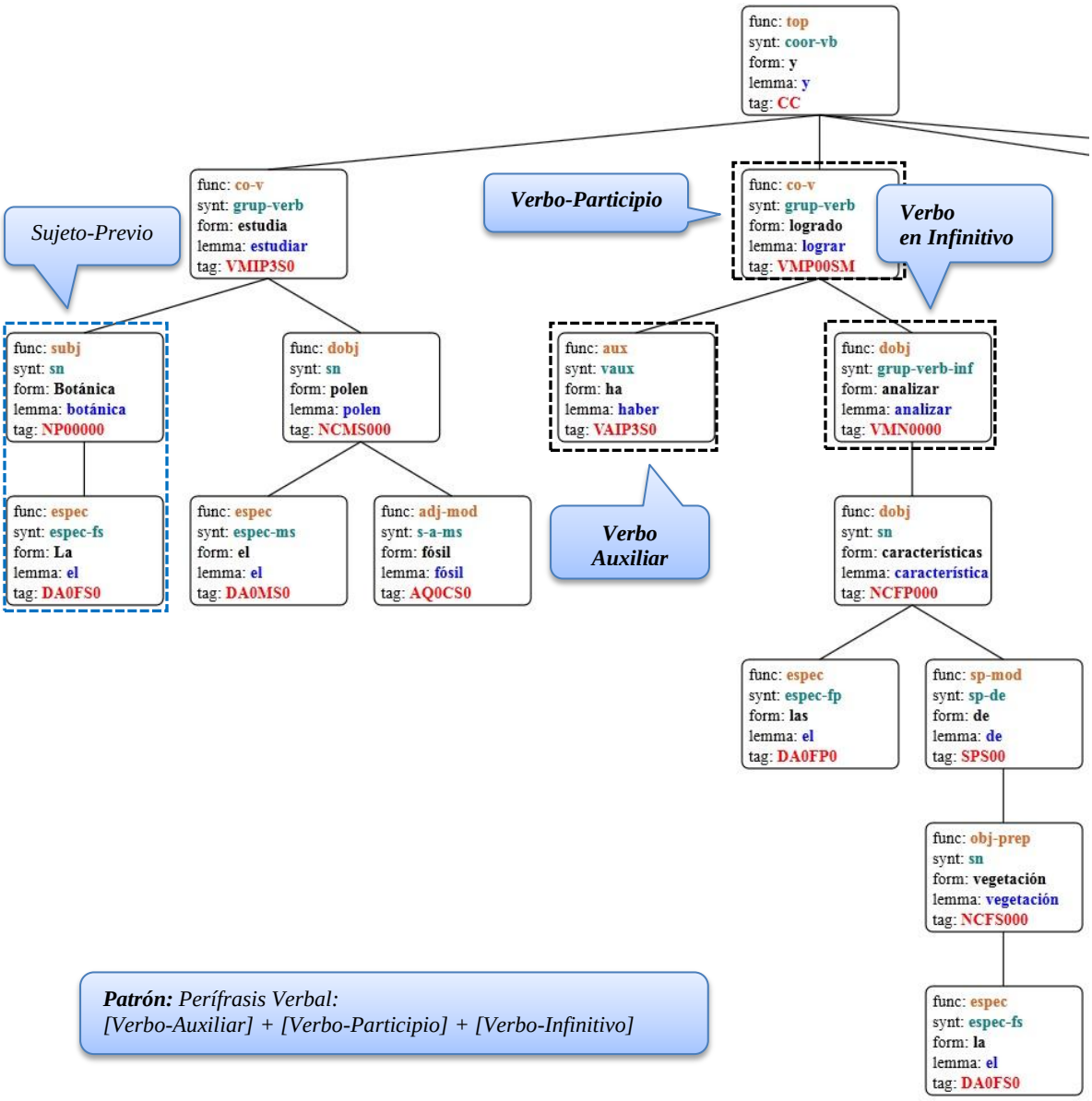

Figura 4.27 Patrón sintáctico "*Perífrasis Verbal del Infinitivo*" en el árbol de dependencias de la oración "*La Botánica estudia el polen fósil y ha logrado analizar las características de la vegetación e inferir los climas*".

4.5.14.2 Coordinación de Verbos en Infinitivo

La coordinación de verbos en una oración también puede tratarse de verbos en infinitivo, en lugar de verbos conjugados como en el patrón sintáctico "Coordinación de Verbos".

El patrón sintáctico "Coordinación de Verbos en Infinitivo" se puede presentar en el patrón sintáctico "Básico" o en el patrón "Coordinación de Verbos".

Coordinación de Verbos en Infinitivo en el patrón "*Básico*"

En la Figura 4.28 se muestra este patrón, se puede ver que la raíz es un verbo, a un nodo etiquetado con {synt: coord / tag: CC} el cual indica la existencia de la coordinación del verbo en infinitivo, y al nodo que contiene al verbo en infinitivo. El verbo en infinito es el que se ocupa para el hecho.

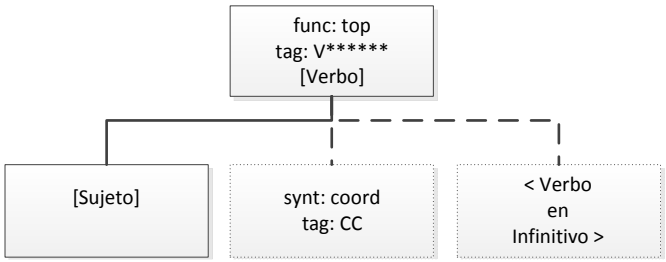

Figura 4.28 Diagrama del patrón sintáctico "*Coordinación de Verbos en Infinitivo*" en el patrón "*Básico*".

La Figura 4.29 muestra un ejemplo del patrón sintáctico "Coordinación de Verbos en Infinitivo", cuando se presenta en el patrón sintáctico "Básico".

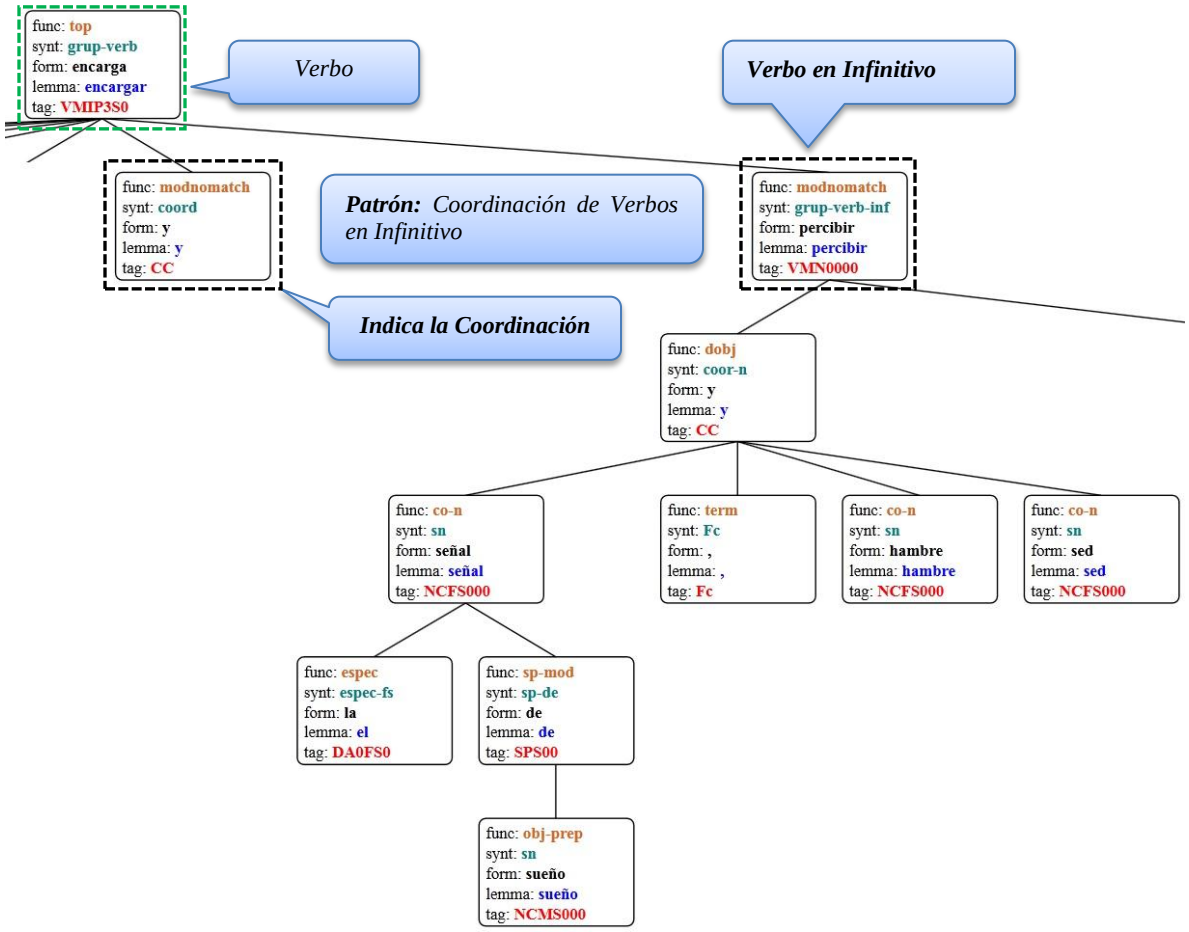

Figura 4.29 Patrón sintáctico "*Coordinación de Verbos en Infinitivo*" en el árbol de dependencias de la oración "*El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed*".

Coordinación de Verbos en Infinitivo en el patrón "*Coordinación de Verbos*"

En la Figura 4.30 se muestra este patrón, se puede ver que la raíz en este caso no es un verbo sino el nodo coordinador de verbos {func: top / synt: coor-vb}. La raíz tiene como descendientes al verbo conjungado {func: co-v} que es parte del patrón "Coordinación de Verbos", y al verbo en infinitivo. El verbo en infinito es el que se toma como verbo para el hecho.

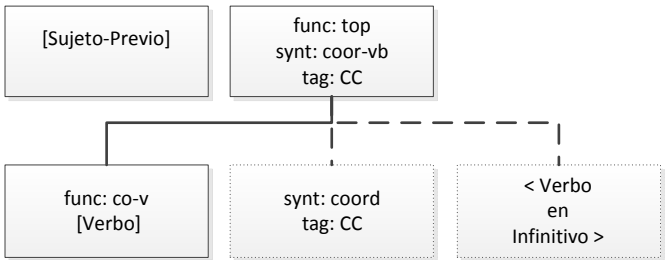

Figura 4.30 Diagrama del patrón sintáctico "*Coordinación de Verbos en Infinitivo*" en el patrón "*Coordinación de Verbos*".

La Figura 4.31 muestra un ejemplo del patrón sintáctico "Coordinación de Verbos en Infinitivo", cuando se presenta en el patrón sintáctico "Coordinación de Verbos".

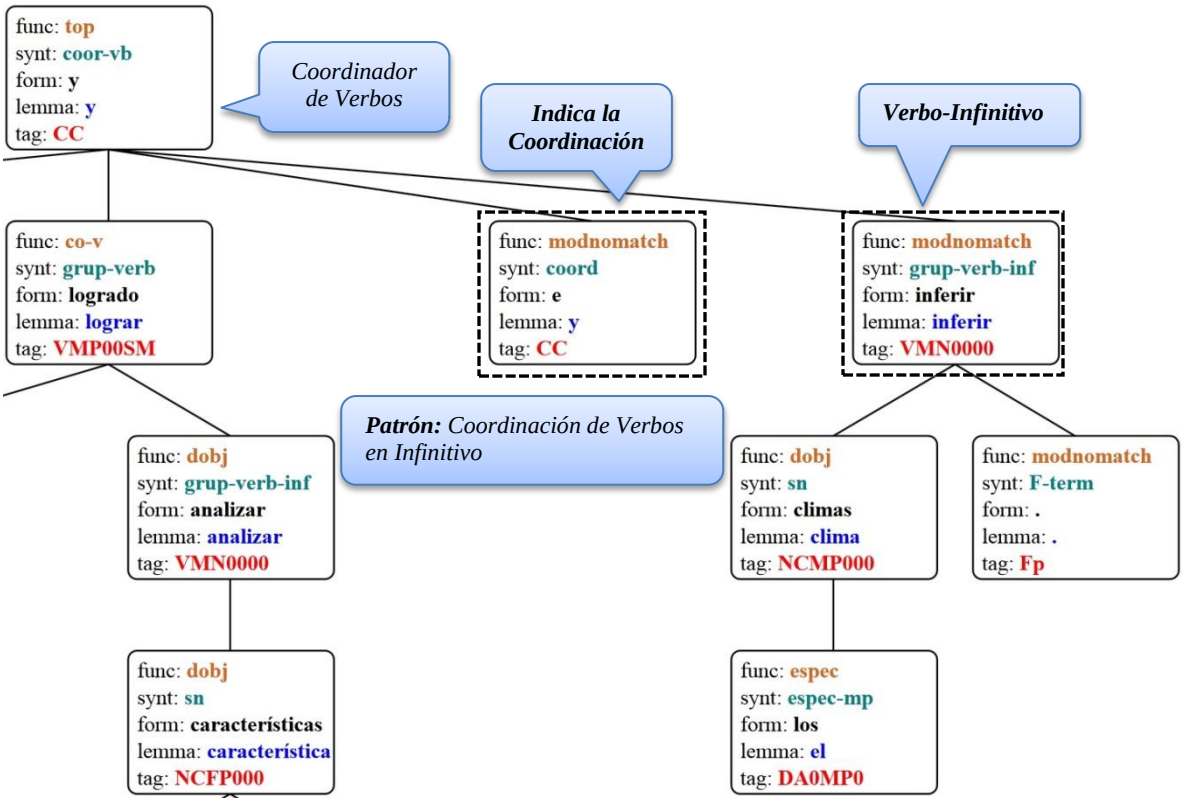

Figura 4.31 Patrón sintáctico "*Coordinación de Verbos en Infinitivo*" en el árbol de dependencias de la oración "*La Botánica estudia el polen fósil y ha logrado analizar las características de la vegetación e inferir los climas*".

4.5.14.3 El algoritmo

Como se puede observar en los patrones sintácticos “Perífrasis Verbal” y “Coordinación de Verbos en Infinitivo”, ambos presentan la similitud de tener un verbo en infinitivo, es por ello que ambos patrones se resumen en un patrón sintáctico común, este patrón es el que representa a la heurística “Verbo en infinitivo”.

La Figura 4.32 muestra el patrón sintáctico de la heurística “Verbo en Infinitivo”, se puede observar que la raíz es el verbo en infinitivo que podría tener como descendientes a un complemento simple o al patrón sintáctico “Coordinación de Sustantivos”.

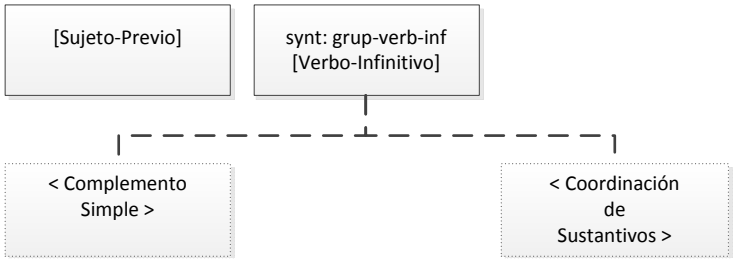

Figura 4.32 Diagrama del patrón sintáctico “*Verbo en Infinitivo*”.

La Figura 4.33 muestra un ejemplo del patrón sintáctico "Verbo en Infinitivo".

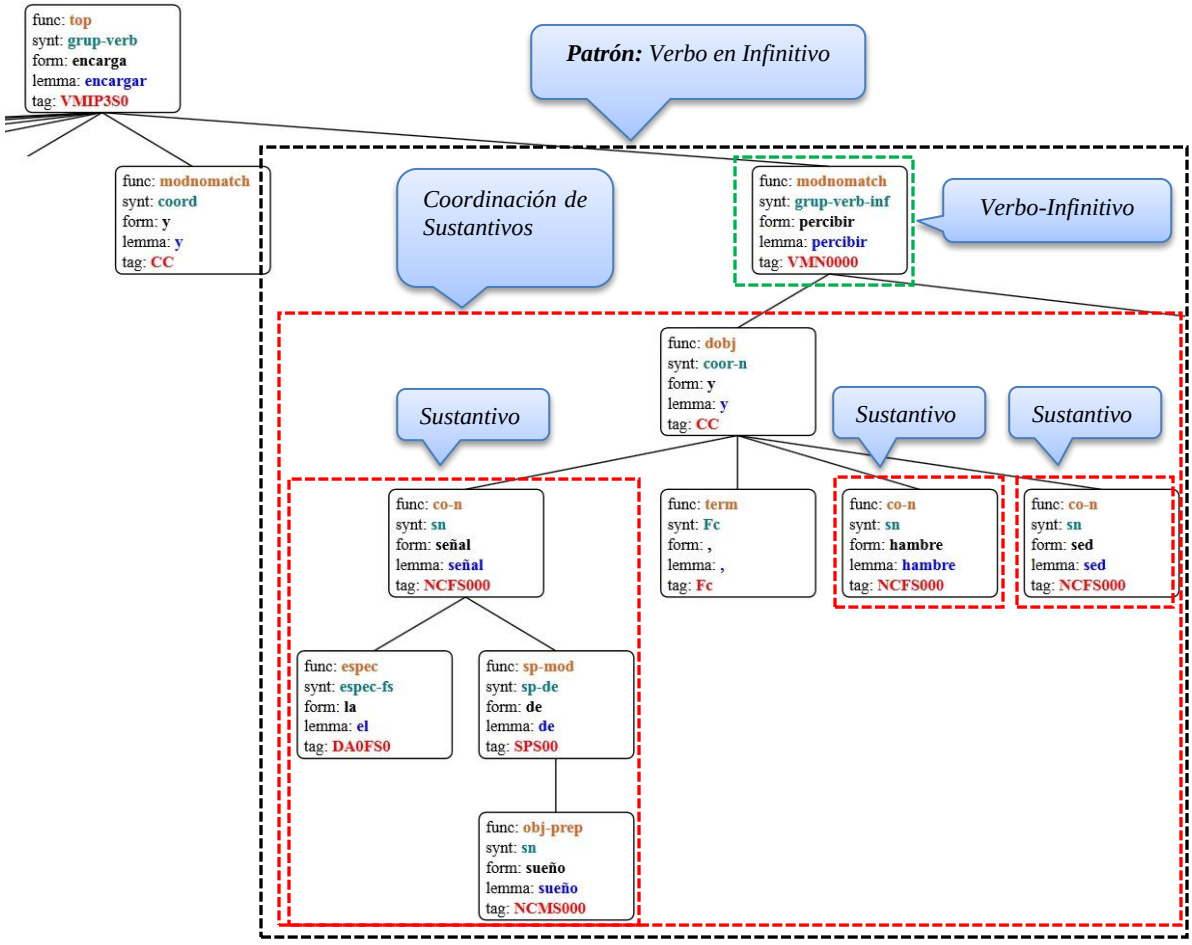

Figura 4.33 Patrón sintáctico "*Verbo en Infinitivo*" en el árbol de dependencias de la oración "*El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed*".

La forma de trabajar de la heurística es la siguiente: el sujeto se obtiene previamente, el verbo en infinitivo se toma para el hecho, el complemento se obtiene de los descendientes del verbo en infinitivo.

En la Tabla 4.19 se describe el algoritmo de manera detallada.

Tabla 4.19 Algoritmo de la heurística "*Verbo en Infinitivo*".

```
1   Verbo en Infinitivo:
2   Parámetros de entrada: Sujeto y un apuntador al nodo Verbo en
    Infinitivo, tiene la etiqueta {synt: grup-verb-inf}.
3   Asignar: Verbo = Valor de {form:} del nodo verbo en infinitivo.
4   Recorrer todos los hijos del nodo verbo en infinitivo y revisar:
5         Si tiene etiquetas de Complemento Simple, extraer ese nodo y
          sus descendientes como complemento.
6               Construir: Hecho = [Sujeto] + [Verbo] + [Complemento].
7         Si tiene las etiquetas {synt: coor-n / tag: CC}.
8               Llamar algoritmo de la heurística
                Coordinación de Sustantivos (Apuntador al hijo, Sujeto,
                Verbo).
9   Fin.
```

En la Tabla 4.20 se muestran los hechos extraídos de una oración, con el algoritmo de la heurística "Verbo en Infinitivo".

Tabla 4.20 Hechos extraídos con la heurística "*Verbo en Infinitivo*" de la oración "*El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed*".

No.	Sujeto	Verbo	Complemento
1	Hipotálamo	percibir	señal de sueño
2	Hipotálamo	percibir	hambre
3	Hipotálamo	percibir	sed

4.5.15 Heurística: Correferencia de Sujeto

"Las oraciones subordinadas adverbiales finales indican la finalidad o el propósito que se busca al realizar la acción del verbo principal. Se caracterizan porque su verbo, cuando está conjugado, siempre está en subjuntivo. Los nexos más usuales para introducirlos son *para*, *para que*, *a fin de que*, *con el fin de que*" (Munguía Zatarain, Munguía Zatarain, & Rocha

Romero, 2000). “Aunque las finales van siempre en subjuntivo, cuando tienen el mismo sujeto que la principal el verbo va en infinitivo” (Fuentes de la Corte, 2010).

Cuando la persona de la oración principal y la de la oración final son la misma empleamos la construcción (para/de/como) + (infinitivo). Ejemplos:

- Pedro estudia para terminar la carrera.
- Pedro estudia para **[Pedro]** terminar la carrera.

Decimos que son sujetos correferentes porque se refieren a la misma persona.

La Figura 4.34 muestra el patrón sintáctico, se puede ver al nodo que indica la correferencia de sujeto con la etiqueta {synt: grup-sp-inf}, como descendiente al patrón sintáctico “Verbo en Infinitivo”.

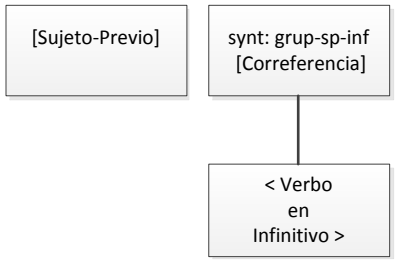

Figura 4.34 Diagrama del patrón sintáctico “Correferencia de Sujeto”.

La Figura 4.35 muestra un ejemplo del patrón sintáctico "Correferencia de Sujeto".

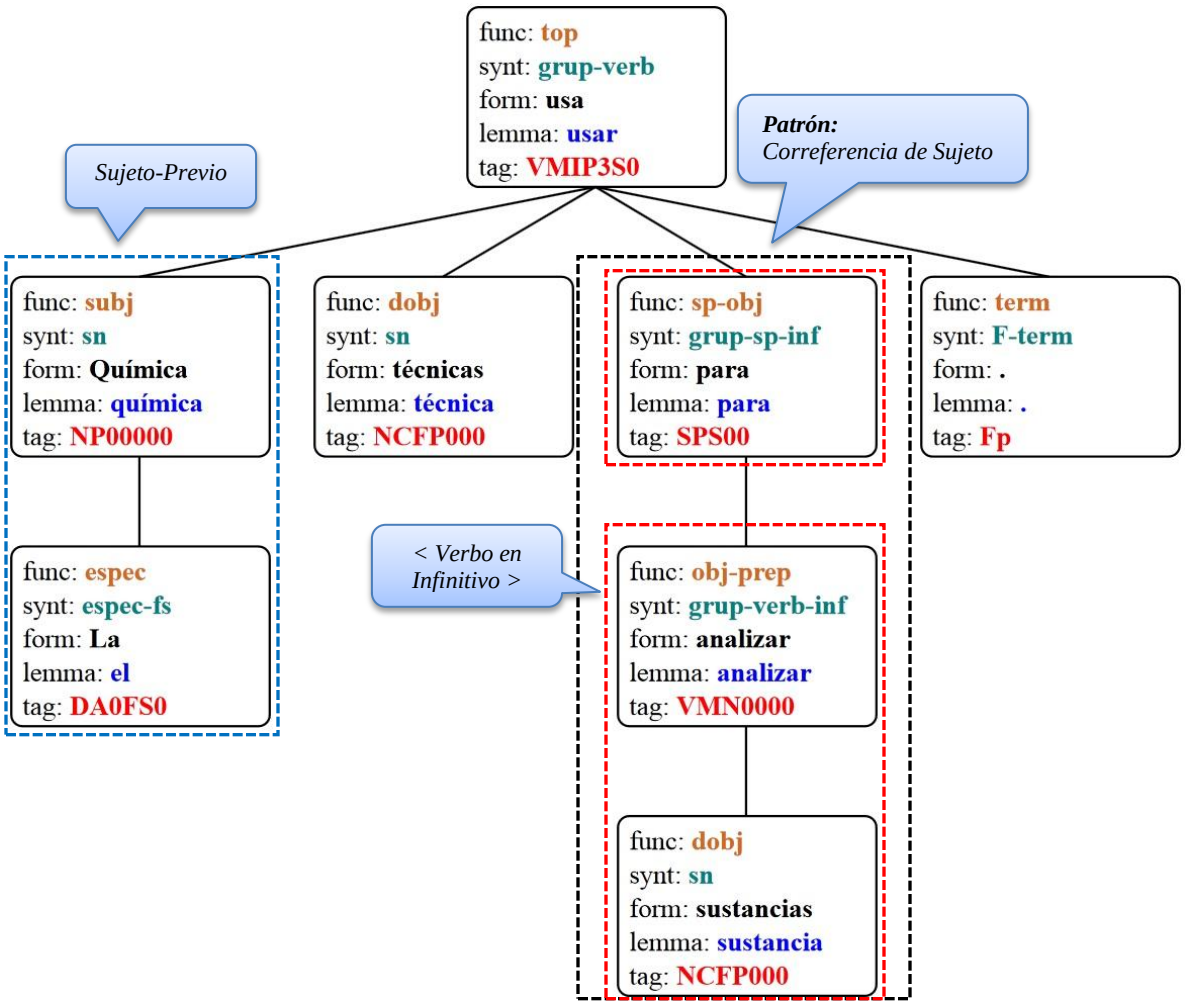

Figura 4.35 Patrón sintáctico "*Correferencia de Sujeto*" en el árbol de dependencias de la oración "*La Química usa técnicas para analizar sustancias*".

La forma de trabajar de la heurística es la siguiente: se obtiene el sujeto previamente, el verbo en infinitivo es el primer hijo del nodo etiquetado como referencia de sujeto, después únicamente se llama a la heurística "Verbo en infinitivo".

En la Tabla 4.21 se describe el algoritmo de manera detallada.

Tabla 4.21 Algoritmo de la heurística "*Correferencia de Sujeto*".

```
1   Correferencia de Sujeto:
2   Parámetros de entrada: Sujeto y un apuntador al nodo Correferencia,
    tiene la etiqueta {synt: grup-sp-inf}.
3   Asignar: Nodo verbo en infinitivo = Primer hijo del nodo
    Correferencia.
4   Llamar algoritmo de la heurística Verbo en Infinitivo (Apuntador a
    nodo verbo en infinitivo, Sujeto).
5   Fin.
```

En la Tabla 4.22 se muestra un hecho extraído de una oración, con el algoritmo de la heurística "Correferencia de Sujeto".

Tabla 4.22 Hechos extraídos con la heurística "*Correferencia de Sujeto*" de la oración "*La Química usa técnicas para analizar sustancias*".

No.	Sujeto	Verbo	Complemento
1	Química	analizar	sustancias

4.6 Almacenamiento de hechos

Los hechos se almacenan en una base de datos relacional compuesta de las siguientes tablas:

- Tabla de oraciones. Aquí se guardan todas oraciones que se les extrae sus hechos, una oración por registro. Se compone de dos campos: el identificador de la oración y la oración.
- Tabla de hechos. Aquí se guardan todos los que se extraen de las oraciones. Se compone de 5 campos: el identificador del hecho, sujeto, verbo, complemento y un campo para guardar el número de la oración a la que pertenecen los hechos.

La relación que se presenta entre las dos tablas es: una oración puede tener muchos hechos, pero un hecho pertenece únicamente a una oración.

5 DESARROLLO DEL SISTEMA

Se describe la construcción del corpus que se utiliza como estándar de oro para evaluar al sistema; se explica el comando y sus parámetros para ejecutar FreeLing; luego se habla sobre el diseño de la base de datos para guardar los hechos; inmediatamente se comentan las herramientas utilizadas para el desarrollo del sistema, se muestra y describe un diagrama de bloques para explicar su funcionamiento, a continuación se muestra y describe su interfaz gráfica de forma detallada.

5.1 Construcción del Corpus

Para evaluar el sistema se creó un corpus, formado por 166 hechos extraídos de 68 oraciones, y se utiliza como un estándar de oro (Gold Standard, en inglés) para la evaluación. Los datos de un estándar de oro representan lo que se define como "la respuesta correcta" (Hovy, Zhou, & Kwon, 2007). El corpus creado es llamado "FactSpCIC", porque es un corpus de hechos de textos en español y fue creado en la presente tesis del Centro de Investigación en Computación del IPN.

El conjunto de las 68 oraciones se formó de la siguiente manera, las primeras 22 fueron tomadas de (Herrera de la Cruz, 2010), de la 23 a 55 se tomaron de (SEP, 2010) y el resto de (SEPb, 2010), eligiendo las oraciones que tenían sujeto explícito, descartando las de sujeto tácito e interjecciones; la correferencia del sujeto se resolvió manualmente, se descartan las oraciones pregunta. De la 56 a la 68 se tomaron de una sola lección: "El sistema nervioso central y el sistema nervioso periférico".

Se eligieron libros de texto porque contienen muchas definiciones e información enunciativa; ya que han sido redactados para cumplir un propósito educativo, y por lo tanto contienen gran cantidad de hechos.

Para crear el estándar de oro, el conjunto de oraciones y la "Guía para un humano para extraer hechos" de (Aguilar-Galicia, Sidorov, & Ledeneva, 2012) se entregaron a dos personas (H1 y H2), y siguieron el procedimiento empleado por (Hovy, Zhou, & Kwon,

2007) para crear un Gold Standard de *nuggets*. En la Figura 5.1 se muestra el procedimiento adaptado a la extracción de hechos.

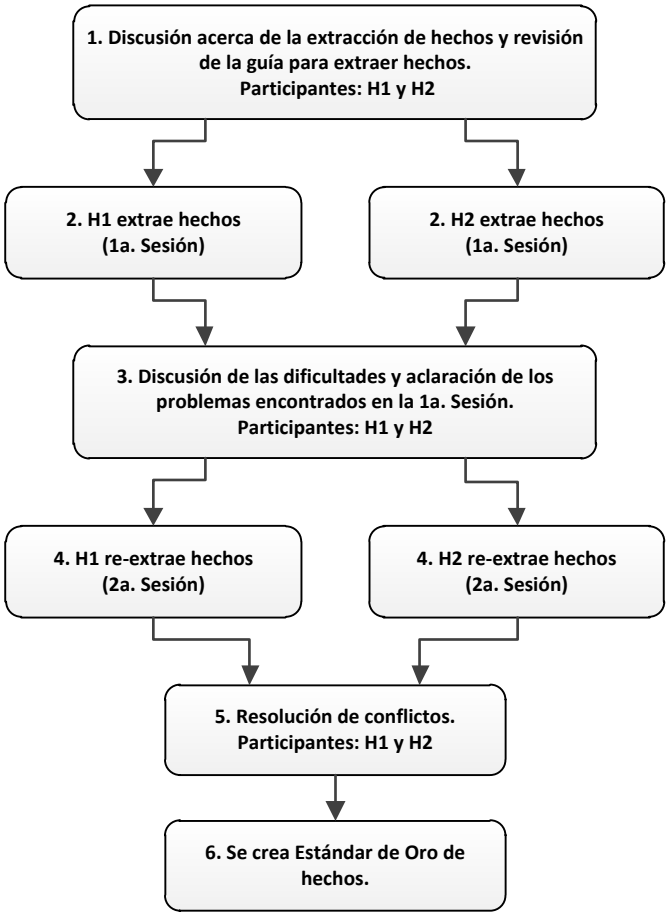

Figura 5.1 Procedimiento para crear Estándar de Oro de hechos.

1. Se tiene una introducción inicial acerca de la tarea de extracción de hechos: su definición y sus características, en dónde se encuentran, cómo se forman. Se revisa la "Guía para un humano para extraer hechos" de (Aguilar-Galicia, Sidorov, & Ledeneva, 2012).
2. H1 y H2 extraen hechos de todas las oraciones, por separado. (Primera sesión de extracción).
3. Se reúnen H1 y H2 para aclarar las dudas acerca de las dificultades o problemas encontrados.
4. H1 y H2 vuelven a extraer hechos de todas las oraciones, por separado. (Segunda sesión de extracción).

5. Se reúnen H1 y H2 para resolución de conflictos, por ejemplo de comprensión o interpretación diferente de la información y por ello omitir algún hecho o incluir alguno que no lo es. De esta forma se acuerda que hechos forman parte del Estándar de Oro.
6. Se crea Estándar de Oro de hechos: Se reúnen y organizan los hechos por cada oración.

5.2 Configuración de FreeLing

FreeLing 2.2 está instalado en la computadora de desarrollo y ese ahí donde se ejecuta. Al ejecutarse genera los archivos de dependencias utilizados por el método propuesto. Los archivos son en formato de texto plano.

Para ejecutar FreeLing se hace con el siguiente comando:

analyzer.exe --outf "parsed" -f es.cfg <ArchivoEntrada.txt >ArchivoSalida.txt

Dónde:

- "**analyzer.exe**". Ejecutable del analizador sintáctico FreeLing.
- "**--outf parsed**". Parámetro constante, indica el tipo de archivo etiquetado resultante, en este caso un archivo de dependencias. Un archivo con etiquetas morfológicas solamente, se obtiene especificando el parámetro "tagged".
- "**-f**". Parámetro constante, indica el tipo de objeto que va a leer, en este caso un archivo (file, en inglés).
- "**es.cfg**". Parámetro constante, indica que se utiliza el archivo de configuración de etiquetado para el español. Ya que se pueden analizar oraciones en otros idiomas, por ejemplo para el inglés el archivo es "en.cfg".
- "**ArchivoEntrada.txt**". Parámetro variable, es el nombre de archivo de la oración a analizar, en formato de texto plano.
- "**ArchivoSalida.txt**". Parámetro variable, es el nombre de archivo de dependencias resultante de la oración analizada, en formato de texto plano.
- "<, >". Símbolos constantes para indicar el archivo de entrada y el de salida.

5.3 Representación de los datos

Los hechos se almacenan en una base de datos relacional compuesta de las siguientes tablas: tblOraciones y tblHechos. En la Figura 5.2 se muestra el diagrama entidad-relación de la base de datos, en donde se observan las tablas y la relación entre ellas, también se muestran los campos y las propiedades de estos.

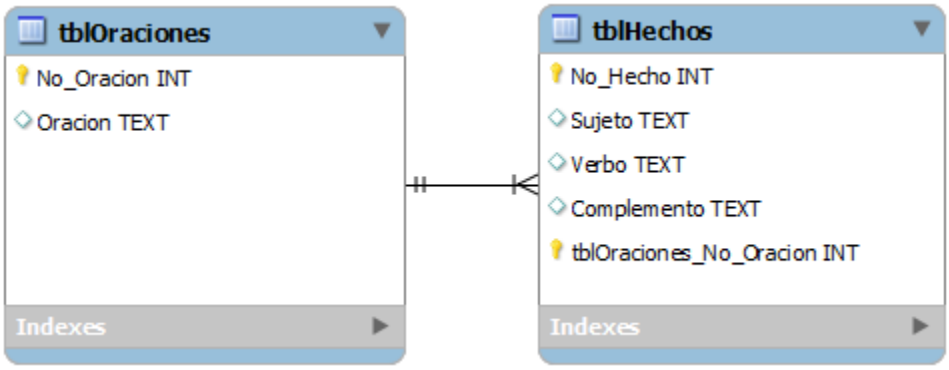

Figura 5.2 Diagrama entidad-relación de la base de datos de hechos.

Para ambas tablas se define una llave primaria, y para la tabla tblHechos una llave foránea con la cual se establece la relación de una oración con muchos hechos.

5.4 Desarrollo del sistema

5.4.1 Arquitectura de desarrollo y ejecución

El sistema está desarrollado como una aplicación de escritorio con el lenguaje de programación Java y la biblioteca Swing. El software utilizado para representar los datos es la base de datos MySQL.

A continuación se describe con más detalle la arquitectura de hardware y software utilizada para el desarrollo, y ejecución del sistema para su evaluación.

5.4.1.1 Hardware

- Procesador: Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz 2.30GHz.
- Memoria RAM: 4.00 GB.

- Disco duro de 640 GB.

5.4.1.2 Software

- **Microsoft Windows 7 Home Premium de 64 bits.** Sistema Operativo instalado en la computadora de desarrollo y ejecución del sistema.
- **Java Development Kit (JDK 1.7.0).** Conjunto de programas y bibliotecas para desarrollar, compilar y ejecutar programas en el lenguaje de programación Java. De distribución libre.
- **Biblioteca Swing.** Biblioteca gráfica para Java, ayuda a crear la interfaz gráfica de usuario, ya que contienes controles como cuadros de texto, etiquetas, áreas de texto, botones, tablas y otros. De distribución libre.
- **NetBeans IDE 7.0.1.** Entorno de desarrollo integrado, principalmente para el lenguaje de programación Java. De distribución libre.
- **MySQL Workbench 5.2.** Manejador de bases de datos, se utiliza para la representación y manejo de los datos. De distribución libre.
- **FreeLing-2.2.** Es un analizador sintáctico, se utiliza para crear los árboles de dependencia de las oraciones. De distribución libre.

5.4.2 Diagrama de bloques

En la Figura 5.3 se presenta el diagrama de bloques del sistema, este incluye las acciones que el usuario puede realizar a través de una interfaz gráfica, como son: Cargar el corpus, Crear archivos de dependencias, Extraer hechos para presentarlos en la interfaz o para guardarlos en la base de datos. También se muestra que los hechos pueden ser utilizados por otras áreas de PLN. El sistema creado en la investigación se le ha llamado: Fact Extraction System 2012 (FES 2012), un sistema de extracción de hechos.

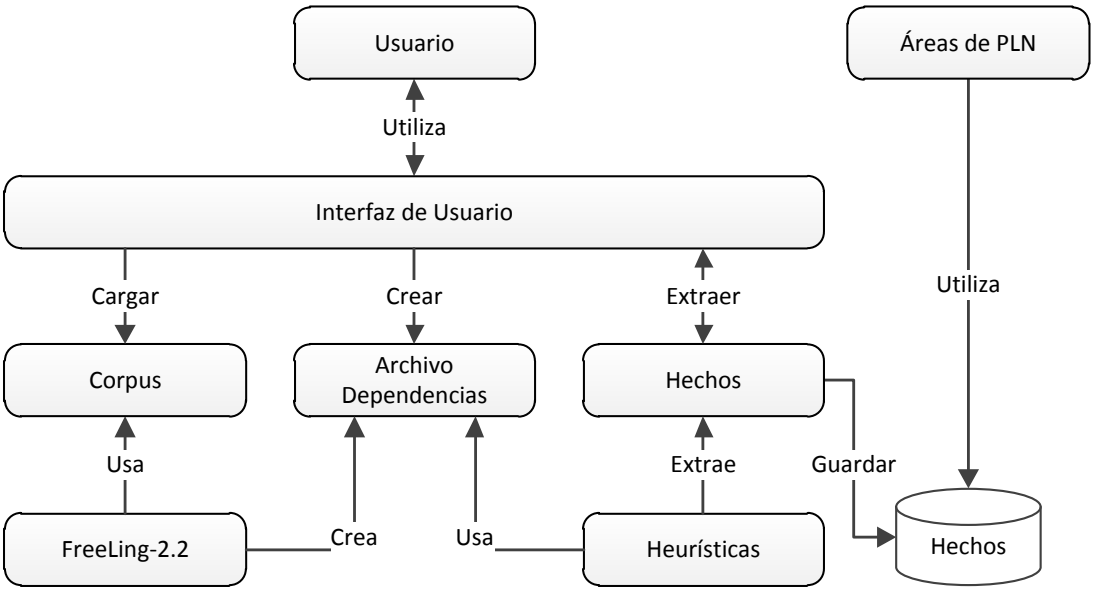

Figura 5.3 Diagrama de bloques del sistema.

Como se puede observar en el diagrama de bloques la extracción de hechos comienza al cargar el corpus, a partir de ahí se pueden crear archivos de dependencias de las oraciones del corpus a mediante FreeLing quien utiliza el corpus para crearlos.

Cuando ya se cuenta con archivos de dependencias, ahora se pueden extraer hechos de una oración o de todas las oraciones que componen el corpus mediante las heurísticas del método propuesto. Si el usuario elige extraer hechos sólo de una oración entonces los hechos se muestran en la interfaz, pero si elige extraer hechos de todo el corpus estos se guardan en la base de datos.

5.4.3 Interfaz del sistema

La interfaz principal del sistema es un formulario compuesto por tres secciones, tal como se puede observar en la Figura 5.4. Después de la figura se describen estas secciones.

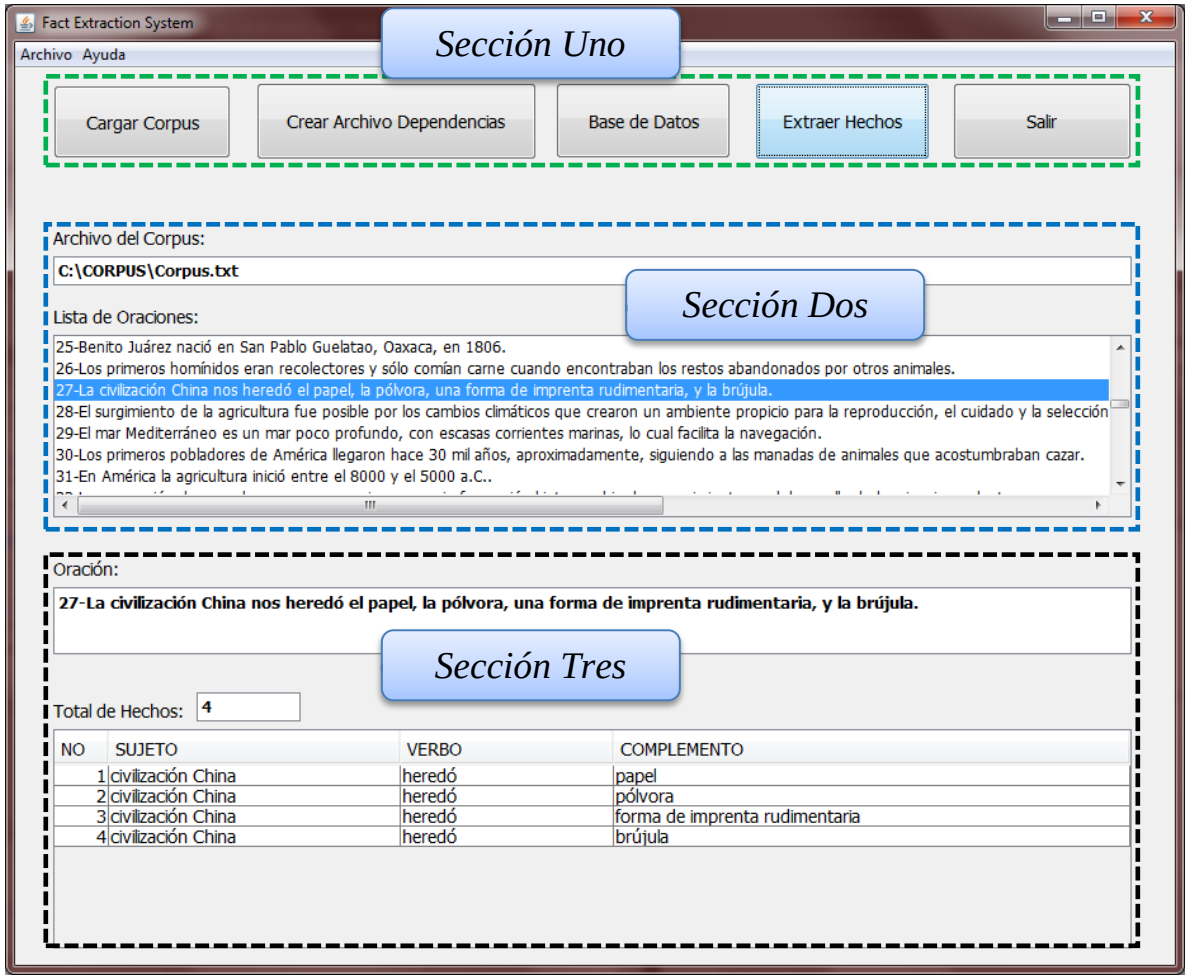

Figura 5.4 Interfaz principal del sistema.

5.4.3.1 Sección uno

Esta sección contiene el menú principal del sistema, aquí se encuentran las opciones con las que cuenta el sistema. La Figura 5.5 muestra exclusivamente la sección uno.

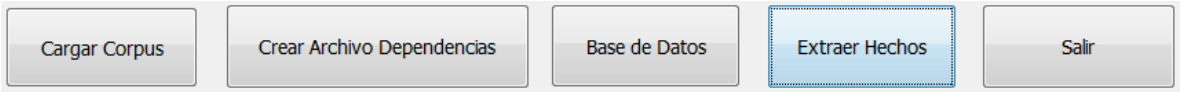

Figura 5.5 Sección uno de la interfaz principal del sistema.

- **Cargar Oraciones.** Al hacer clic se abre un cuadro de diálogo para seleccionar el archivo del corpus.

- **Crear Archivo Dependencias.** Al hacer clic se ejecuta FreeLing y este crea el archivo de dependencias de la oración seleccionada en el cuadro "*Lista de Oraciones*" de la sección dos.
- **Base de Datos.** Al hacer clic se extraen todos los hechos, de todas las oraciones del corpus y se guardan en la base de datos.
- **Extraer Hechos.** Al hacer clic se extraen los hechos únicamente de la oración seleccionada en el cuadro "*Lista de Oraciones*" de la sección dos, y son listados en la tabla "*Total de hechos*" de la sección tres. En el cuadro "*oración*" se muestra la oración seleccionada.
- **Salir.** Termina la ejecución del sistema.

5.4.3.2 Sección dos

Esta sección contiene elementos que son llenados al hacer clic en alguna opción del menú. La Figura 5.6 muestra exclusivamente la sección dos.

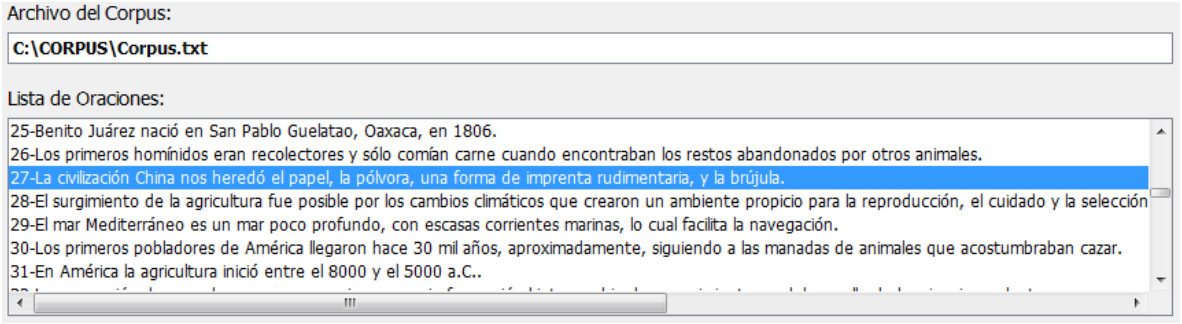

Figura 5.6 Sección dos de la interfaz principal del sistema.

- **Archivo del Corpus.** Este cuadro de texto muestra la ruta y el nombre del archivo del corpus cargado. Se llena al cargar el archivo del corpus.
- **Lista de Oraciones.** Este cuadro contiene en forma de lista todas las oraciones del corpus. Para seleccionar alguna oración se hace clic sobre ella. En la Figura 5.6 se observa que se ha se ha seleccionado la oración "*27-La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*". Se llena al cargar el archivo del corpus.

5.4.3.3 Sección tres

Esta sección contiene elementos que son llenados al hacer clic en alguna opción del menú. La Figura 5.7 muestra exclusivamente la sección tres.

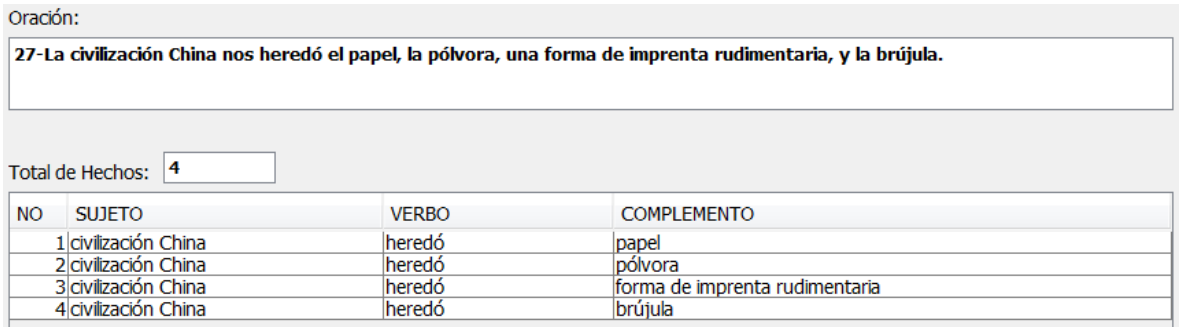

Figura 5.7 Sección tres de la interfaz principal del sistema.

- **Oración.** Este cuadro de texto muestra la oración seleccionada en el cuadro "*Lista de Oraciones*", pero al elegir la opción "*Extraer Hechos*".
- **Total de Hechos.** Este pequeño cuadro muestra la cantidad de hechos extraídos de la oración.
- **Tabla "Total de Hechos".** Esta tabla muestra en forma de lista y enumerados los hechos extraídos de la oración. Separando cada hecho por Sujeto, Verbo y Complemento.

6 EVALUACIÓN Y RESULTADOS

Se explica cómo se evalúa el sistema FES 2012, las métricas de evaluación, se presentan y describen los resultados obtenidos por el sistema, y una comparación con otros sistemas.

6.1 Método de evaluación utilizado

El método de evaluación consiste en la comparación de los hechos obtenidos por el sistema con respecto a un estándar de oro, en este caso el Corpus "FactSpCIC".

6.1.1 Definición del estándar de oro

FactSpCIC es un corpus compuesto de 166 hechos extraídos de 68 oraciones en el idioma español. El corpus está organizado como un conjunto de oraciones, donde cada una de ellas tiene un conjunto de hechos. Puede consultarse en el anexo A, en la sección "FactSpCIC: El estándar de oro".

6.1.2 Definición de la salida de FES 2012

Al ejecutar el sistema con el conjunto de oraciones del corpus, este produce un conjunto de hechos para cada una de las oraciones.

6.1.3 Medidas de evaluación

La extracción de información es una tarea que no es trivial. De hecho cuando se hace manualmente se ha encontrado que el grado de coincidencia en los resultados, comparando la producción de diferentes personas sobre la misma colección de noticias, está entre el 60 y 80% (Martí Antonín & Alonso Martín, Tecnologías del lenguaje, 2003). De igual forma la extracción de hechos no es trivial, pues la complejidad de la tarea depende del tipo de texto que se quiere procesar, por ejemplo.

En la Extracción de Información se emplean métricas que permiten determinar y comparar el grado de rendimiento que alcanza un sistema. La dos medidas básicas que se consideran son la *precisión* (calidad de la información extraída) y la *cobertura* (*Recall*, en inglés)

(cuánta de la información que debía ser extraída lo ha sido realmente). Estas métricas son las que se utilizan en la investigación, y se definen a continuación.

6.1.3.1 Precisión del sistema

La precisión (*P*) del sistema se mide dividiendo el número de hechos correctos obtenidos por el sistema entre el número total de hechos obtenidos por el sistema.

$$\boldsymbol{P} = \frac{\textit{Hechos correctos obtenidos por el sistema}}{\textit{Total de hechos obtenidos por el sistema}}$$

6.1.3.2 Recall

El recall (*R*) del sistema se mide dividiendo el número de hechos correctos obtenidos por el sistema entre el número total he hechos existentes en el texto, en este caso, los hechos en el estándar de oro.

$$\boldsymbol{R} = \frac{\textit{Hechos correctos obtenidos por el sistema}}{\textit{Total de hechos existentes en el corpus (extraídos por un humano)}}$$

6.1.3.3 F1

F1 es otra métrica utilizada para evaluar sistemas de Extracción de Información, donde *F1* es una medida que combina y balancea Precisión y Recall (Jurafsky & Martin, 2000), y está definida como sigue:

$$\boldsymbol{F1} = 2\frac{P\,R}{P + R}$$

En otras palabras, *F1* es la media armónica de precisión y recall.

6.2 Resultados de la evaluación

A continuación se describen los resultados que obtiene el sistema.

6.2.1 Oraciones procesadas

De las 68 oraciones del corpus, el sistema sólo extrae hechos de 59 oraciones ya que 9 de ellas no fueron etiquetadas correctamente por FreeLing. Estas nueve oraciones suman un total de 15 hechos, los que representa un 9.03% de los 166 totales del estándar de oro, que no se procesaron por el sistema, por lo tanto para la evaluación el estándar se reduce a 151 hechos.

Las nueve oraciones que no procesa el sistema son: "3, 5, 7, 11, 18, 40, 46, 55, 68", y se debe a que el sujeto de la oración no fue identificado, pues de acuerdo a los patrones sintácticos se espera que el sujeto se etiquete con {func: subj / synt: sn} ó {func: subj-pac}. Pero el sujeto identificado en estas oraciones en el estándar de oro, FreeLing no lo etiqueta como tal, en su lugar las etiquetas que coloca para el sujeto en cada oración son:

- *O-3: {func: cc / synt: sn}*
- *O-5: {func: dobj / synt: sn}*
- *O-7: {func: cc / synt: sn}*
- *O-11: {func: att / synt: sn}*
- *O-18: {func: dobj / synt: sn}*
- *O-40: {func: cc / synt: sn}*
- *O-46: {func: dobj / synt: sn}*
- *O-55: {func: cc / synt: sn}*
- *O-68: {func: dobj / synt: sn}*

Se puede observar que se ha etiquetado como complemento circunstancial, objeto directo y atributo nominal. La Figura 6.1 muestra como se ha etiquetado "*El ritmo*" que se identifica como sujeto en el estándar de oro, de la "*O-3: El ritmo es la alternancia de sílabas átonas con sílabas tónicas.*", pero FreeLing no lo reconoce así.

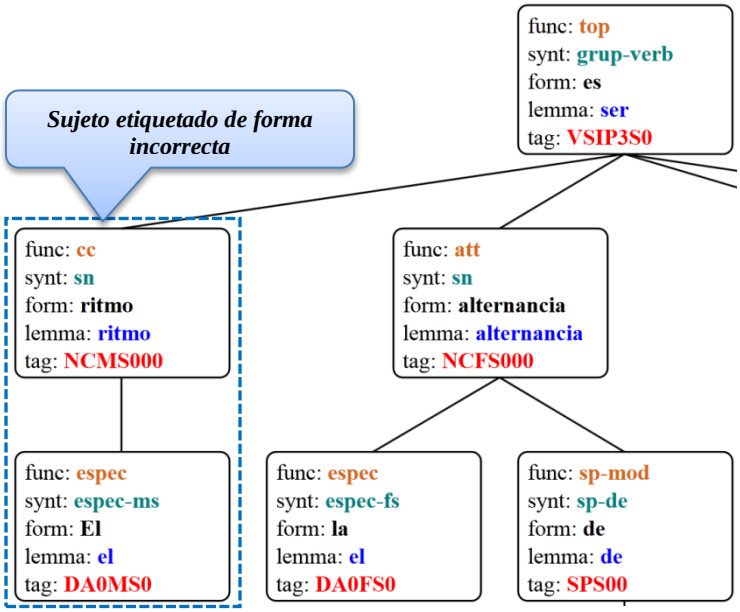

Figura 6.1 Sujeto etiquetado diferente a como se espera, de la oración "*El ritmo es la alternancia de sílabas átonas con sílabas tónicas.*".

6.2.2 Total de hechos obtenidos

Extrae un total de 182 hechos, de ellos 25 son correctos de acuerdo a la definición pero no son tomados en cuenta para la evaluación ya que son hechos con más detalle comparado con los del estándar de oro, pero finalmente se pueden igualar a un único hecho del corpus. Por ejemplo de "*O-25: Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806*", el estándar de oro contiene los siguientes hechos.

1. *(Benito_Juárez, **nació**, en San_Pablo_Guelatao, Oaxaca)*
2. *(Benito_Juárez, **nació**, en 1806)*

Y el sistema extrae:

1. *(Benito_Juárez, **nació**, en San_Pablo_Guelatao)*
2. *(Benito_Juárez, **nació**, Oaxaca)*
3. *(Benito_Juárez, **nació**, en 1806)*

Donde se puede observar que el hecho dos es parte del hecho uno, así que en la evaluación sólo se cuenta el uno y el tres. Por lo tanto sin considerar este tipo de hechos, tenemos un total de 157 obtenidos por el sistema.

6.2.3 Hechos correctos

De los 157 hechos obtenidos por el sistema, solamente 137 son correctos, es decir, por cada uno de los 137 le corresponde uno del estándar de oro. Por ejemplo de "*O-64: El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed*", el estándar contiene los siguientes hechos:

1. *(El Hipotálamo, **se encarga**, de algunas funciones corporales)*
2. *(El Hipotálamo, **regular**, la temperatura)*
3. *(El Hipotálamo, **percibir**, la señal de sueño)*
4. *(El Hipotálamo, **percibir**, hambre)*
5. *(El Hipotálamo, **percibir**, sed)*

Y el sistema extrae:

1. *(Hipotálamo, **encarga**, de funciones corporales)*
2. *(Hipotálamo, **regular**, temperatura)*
3. *(Hipotálamo, **percibir**, señal de sueño)*
4. *(El Hipotálamo, **percibir**, hambre)*
5. *(El Hipotálamo, **percibir**, sed)*

Se puede observar la correspondencia de hechos de uno a uno, del estándar a los obtenidos por el sistema.

6.2.4 Hechos incorrectos

De los 157 hechos obtenidos por el sistema, 20 de ellos son incorrectos ya que la información que enuncian no tiene sentido. Por ejemplo de "*O-62: El Tálamo se halla en el centro del encéfalo, recibe las señales enviadas por los sentidos y las reenvía a distintas áreas del cerebro para su procesamiento.*", el estándar contiene los siguientes hechos:

1. *(El Tálamo, **se halla**, en el centro del encéfalo)*
2. *(El Tálamo, **recibe**, las señales enviadas por los sentidos)*

Y el sistema extrae:

1. *(Tálamo, **halla**, en centro de encéfalo)*
2. *(Tálamo, **recibe**, señales enviadas)*
3. *(*) (Tálamo, **recibe**, por sentidos)*

El hecho tres que extrae el sistema es incorrecto porque no tiene sentido. Este tipo de hechos son los que se cuentan como incorrectos y son consecuencia de la forma en que FreeLing construye el árbol de dependencias.

6.2.5 Hechos no encontrados

De las oraciones si procesadas, existen algunos hechos en el estándar pero que el sistema no logra extraerlos. En total son 14. Por ejemplo de "*O-67: La Médula espinal es la prolongación del encéfalo, tiene forma de cordón y corre por dentro de la columna vertebral, que la protege*", el estándar contiene los siguientes hechos:

1. *(La Médula espinal, **es**, la prolongación del encéfalo)*
2. *(La Médula espinal, **tiene**, forma de cordón)*
3. *(La Médula espinal, **corre**, por dentro de la columna vertebral)*

Y el sistema extrae:

1. *(Médula espinal, **es**, prolongación de encéfalo)*
2. *(Médula espinal, **tiene**, forma de cordón)*

El sistema no encuentra el hecho tres del estándar, ya que FreeLing etiqueta de forma inesperada la coordinación de verbos en el árbol de dependencias. La Figura 6.2 muestra como el nodo del verbo "*corre*" no tiene descendientes, pero se esperaba que tuviera a "*por dentro de la columna vertebral*", por tal razón no se encuentra este hecho.

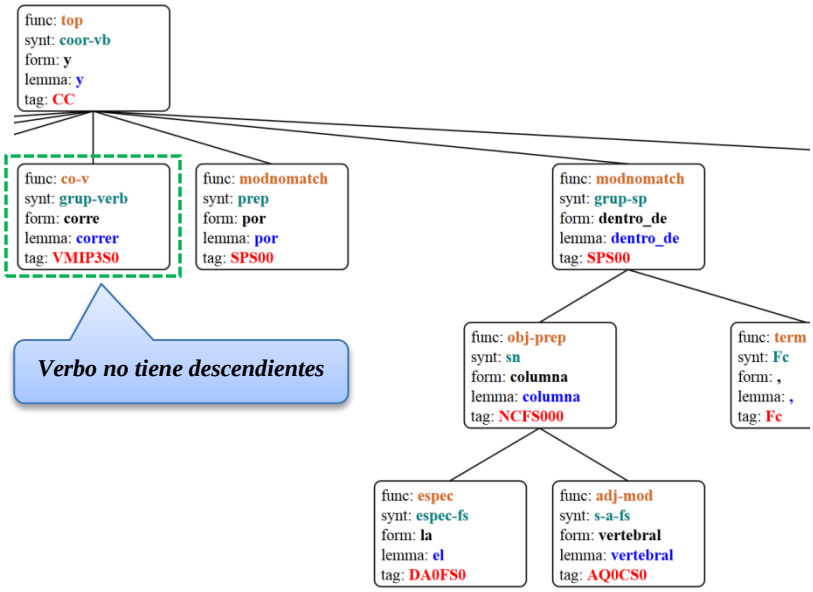

Figura 6.2 Coordinación de verbos etiquetada de forma inesperada.

6.2.6 Resultados detallados por oración

Las cantidades que se describen anteriormente se pueden consultar por oración y de forma detallada en la Tabla 6.1, en donde se puede ver que las oraciones no procesadas están marcadas con "NA – No Aplica".

Tabla 6.1 Resultados de la evaluación FactSpCIC vs FES 2012, de forma detallada.

No. O	FactSpCIC	FES 2012	Correctos	Incorrectos	No encontrados
1	1	2	1	1	
2	2	2	2	0	
3	NA	NA	NA	NA	NA
4	1	1	1	0	
5	NA	NA	NA	NA	NA
6	2	2	2	0	
7	NA	NA	NA	NA	NA
8	1	1	1	0	
9	1	1	1	0	
10	2	2	2	0	
11	NA	NA	NA	NA	NA
12	4	4	4	0	
13	1	2	1	1	
14	3	4	3	1	
15	1	1	1	0	
16	1	1	1	0	
17	2	2	2	0	
18	NA	NA	NA	NA	NA
19	2	2	2	0	
20	3	3	3	0	

No. O	FactSpCIC	FES 2012	Correctos	Incorrectos	No encontrados
21	10	5	4	1	6
22	2	2	2	0	
23	2	2	2	0	
24	1	1	1	0	
25	2	2	2	0	
26	4	4	3	1	1
27	4	4	4	0	
28	4	5	2	3	2
29	4	3	3	0	1
30	3	2	2	0	1
31	1	1	1	0	
32	3	4	3	1	
33	3	3	3	0	
34	5	5	5	0	
35	1	1	1	0	
36	4	5	3	2	1
37	2	2	2	0	
38	1	1	1	0	
39	1	1	1	0	
40	NA	NA	NA	NA	NA
41	1	2	1	1	
42	1	1	1	0	
43	2	2	2	0	
44	1	1	1	0	
45	4	4	4	0	
46	NA	NA	NA	NA	NA
47	3	3	3	0	
48	2	2	2	0	
49	2	2	2	0	
50	1	1	1	0	
51	2	2	2	0	
52	1	1	1	0	
53	3	3	3	0	
54	3	3	3	0	
55	NA	NA	NA	NA	NA
56	2	2	2	0	
57	4	5	4	1	
58	3	3	3	0	
59	6	6	6	0	
60	5	7	4	3	1
61	1	1	1	0	
62	3	4	3	1	
63	3	3	3	0	
64	5	5	5	0	
65	1	4	1	3	
66	5	5	5	0	
67	3	2	2	0	1
68	NA	NA	NA	NA	NA
Total	**151**	**157**	**137**	**20**	**14**

6.2.7 Precisión, Recall y F1.

Tomando en cuenta los cálculos anteriores se obtiene que el estándar de oro se compone de un total de 151 hechos y que el sistema extrae 157 hechos en total pero de ellos sólo 137 son correctos. La Figura 6.3 muestra estos datos.

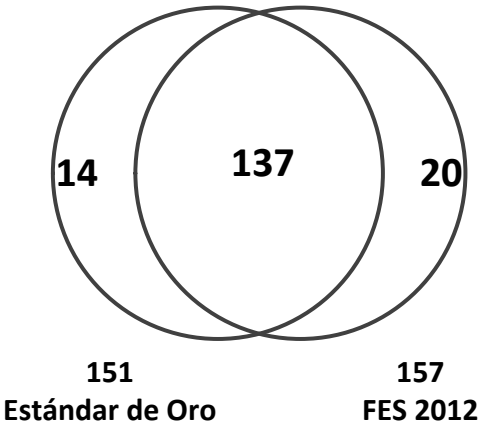

Figura 6.3 Resultados de la evaluación FactSpCIC vs FES 2012, de forma general.

De acuerdo a esto se calcula la precisión, recall y F1 del sistema, y se obtienen los resultados que se muestran en la Tabla 6.2.

Tabla 6.2 Resultados de la evaluación

Medidas	Resultado
Precisión del sistema	87%
Recall	91%
F1	88.9%

6.3 Comparación de resultados

El sistema FES 2012 se compara principalmente con el sistema de Herrera de la Cruz, comentado en el estado del arte, mencionando primero sus características y luego se muestra una tabla comparativa de los resultados de los dos.

Sistema de Herrera de la Cruz. La comparación de los resultados se hace contra el sistema propuesto por (Herrera de la Cruz, 2010): "Sistema de extracción automática de información semántica de los libros de texto estructurados". Este sistema extrae hechos de

oraciones de textos en español mediante un algoritmo basado en heurísticas, las cuales analizan árboles de dependencias generados por el analizador sintáctico "Connexor". Los hechos tienen la secuencia gramatical sujeto-verbo-complemento.

A continuación se muestra un tabla comparativa de los resultados del sistema propuesto por (Herrera de la Cruz, 2010) y los obtenidos por el sistema FES 2012 de la presente investigación. (Véase Tabla 6.3).

Tabla 6.3 Comparación de resultados del sistema de Herrera de la Cruz y FES 2012.

Medidas	Herrera de la Cruz	FES 2012
Precisión del sistema	80%	87%
Recall	73%	91%
F1	76%	88.9%

Otro sistema descrito en el estado del arte es el de (Hovy, Zhou, & Kwon, 2007), aquí se mencionan sus características y la precisión que alcanza. Es importante resaltar que este trabajo es con textos en el idioma inglés.

Sistema de Hovy et. al: Es un sistema que extrae porciones de texto, en el idioma inglés, más pequeñas que una oración y tienen independencia semántica, a las cuales les llaman nuggets (llamados hechos en esta tesis). Tiene varios niveles de granularidad, es decir, un nugget puede ser un sustantivo, sustantivo-verbo o un sustantivo-verbo-complemento. Los resultados obtenidos comparando contra los de un humano tienen una media geométrica de 0.7465.

6.4 Discusión de resultados

6.4.1 Costo computacional

Trabajar con análisis de dependencias implica un costo computacional alto debido a las características sintácticas detalladas, aunque este análisis podría mejorar la *precisión* y *recall* sobre análisis sintácticos superficiales, a costa de una extracción rápida (Gamallo & Garcia, 2012). Este costo se presentó en la investigación durante la creación de los árboles de dependencias por FreeLing.

6.4.2 Sobre las relaciones basadas en verbos

El sistema extrae relaciones binarias basadas en verbos que se encuentran de forma explícita en las oraciones, por ello no se podrían extraer otro tipo de relaciones que se encuentren implícitas. Por ejemplo de la oración "*Pedro, jugador de fútbol del club Leopardos, ganó el premio de mejor jugador del torneo*", extrae el siguiente hecho:

1. (Pedro, **ganó**, el premio de mejor jugador del torneo)

Pero podrían extraerse otras relaciones no verbales que se encuentran dentro de sintagmas nominales, por ejemplo:

1. (Pedro, **es**, un jugador de fútbol del club Leopardos)
2. (Mejor jugador del torneo, **es**, un premio)

Para poder extraer este tipo de hechos, el analizador sintáctico debería ser capaz de etiquetar la función sintáctica de oraciones explicativas, pues la frase "*jugador de fútbol del club Leopardos*" cumple esta función en la oración.

7 CONCLUSIONES Y TRABAJO FUTURO

Se presentan las conclusiones, aportaciones de la investigación, trabajo futuro, presentaciones y publicaciones durante el desarrollo de la tesis.

7.1 Conclusiones

La información semántica en esta tesis se define como hecho, y un conjunto de ellos pueden ser utilizados para otras tareas de PLN, por ejemplo medir la calidad en contenido de un documento de acuerdo al número de hechos contenidos en él respecto a su longitud. O lograr que la computadora guarde conocimiento y no solamente texto.

En este trabajo de tesis se desarrolló un sistema llamado "Fact Extraction System 2012 (FES 2012)", para extraer información semántica de forma automática mediante el análisis de estructuras sintácticas. La forma de trabajar del sistema en resumen es como sigue:

- Se carga un conjunto de oraciones.
- Se realiza un análisis sintáctico automático de cada oración generando un árbol de dependencias para cada una de ellas. Se utiliza FreeLing como analizador sintáctico.
- Utilizando los árboles de dependencias y un conjunto de algoritmos desarrollados en la investigación para extraer hechos, el sistema obtiene los hechos de cada oración.

Se creó un estándar de oro para poder evaluar los resultados obtenidos por FES 2012. El estándar de oro es un corpus llamado "FactSpCIC" que se compone de 166 hechos extraídos de 68 oraciones. El corpus fue creado manualmente por dos personas.

En la evaluación del sistema se obtuvo una *precisión* de 87% y un *recall* de 91%.

7.2 Aportaciones

Al finalizar la investigación se tienen aportaciones que se dividen en teóricas y técnicas, las cuales se describen a continuación.

7.2.1 Aportaciones científicas

- Desarrollo de un método para la extracción de hechos.
- Evaluación de los algoritmos para extraer hechos.

7.2.2 Aportaciones técnicas

- Manual para la extracción de hechos por un humano.
- El Corpus de hechos "FactSpCIC" que puede ser utilizado como un estándar de oro para comparar los resultados de otras investigaciones relacionadas con la tarea de extraer hechos.
- Los patrones sintácticos en los árboles de dependencias, que identifican los componentes de un hecho.
- Desarrollo e implementación de los algoritmos para extraer hechos. Pueden ser programados en cualquier lenguaje.
- Diseño de una base de datos relacional para guardar los hechos extraídos.
- El sistema Fact Extraction System 2012 (FES 2012) para extraer hechos de forma automática de las oraciones, basado en el análisis de estructuras sintácticas.

7.3 Trabajo futuro

- Hacer mejoras a las heurísticas desarrolladas para mejorar los resultados, un ejemplo sería, cuando existe la composición verbal (se) + (verbo) actualmente sólo se toma el verbo quizás debería tomarse también (se) para lograr dar más sentido al hecho.
- Investigar y elegir un formato estándar para almacenar los hechos para que sean de fácil recuperación por otras áreas de PLN o sistemas que ocupan el conocimiento.
- Analizar e implementar un posprocesamiento para las oraciones en donde FreeLing no ha identificado el *sujeto*, y así lograr extraer hechos de estas oraciones. Se buscará el sujeto partiendo de la hipótesis de que el verbo y el sujeto siempre concuerdan en género y número.

- Analizar si es posible crear un resumen extractivo de un texto con los hechos extraídos de él, considerando cuáles hechos son los más importantes para el resumen.

7.4 Presentaciones y publicaciones

Durante el período de estudio del programa de maestría se participó en algunos eventos científicos y la publicación de un artículo.

- Participación en el 8º Taller de Tecnologías del Lenguaje con el poster "*Automatic Fact Extraction based on Syntactic Structures*" en la Benemérita Universidad Autónoma de Puebla (BUAP). Ciudad de Puebla. Noviembre de 2011.

- Presentación de la ponencia "*Extracción automática de hechos de libros de texto basada en estructuras sintácticas*" en el Congreso Mexicano de Inteligencia Artificial (COMIA) 2012 en la Universidad Tecnológica de Xicotepec de Juárez. Xicotepec de Juárez, Puebla. Junio de 2012.

- Publicación: Aguilar-Galicia, Honorato; Sidorov, Grigori; Ledeneva, Yulia (2012). "Extracción automática de hechos de libros de texto basada en estructuras sintácticas". *Avances en Inteligencia Artificial*, (Vol. 55, pp 15-26), México, D.F., México: Instituto Politécnico Nacional.

- Participación en el 9º Taller de Tecnologías del Lenguaje con el poster "*Un algoritmo de extracción de hechos de textos*" en el Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE). Tonantzintla, Puebla. Octubre de 2012.

Anexo A. Corpus de prueba

Lista de oraciones del corpus

1. Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.
2. El libro de los muertos es otro de los escritos que se han encontrado en diversas tumbas egipcias.
3. El ritmo es la alternancia de sílabas átonas con sílabas tónicas.
4. Los datos de las fuentes consultadas deben registrarse en fichas bibliográficas.
5. Redactar un borrador en el cual se deben considerar las partes que conforman un trabajo de investigación.
6. Esta sección debe ser breve e interesante.
7. Las conclusiones sintetizan los argumentos presentados en el desarrollo del trabajo.
8. La bibliografía es el listado alfabético de todas las fuentes consultadas para la elaboración del trabajo.
9. La bibliografía se estructura con los datos de las fichas bibliográficas de esos textos.
10. Las repeticiones sirven para acentuar las emociones y sentimientos del hablante lírico.
11. Es muy común el empleo de llaves para estructurar un cuadro sinóptico.
12. El epíteto heroico es la expresión que menciona una característica o cualidad del personaje u objeto nombrado.
13. El vocativo épico es un enunciado exclamativo intercalado en una oración.
14. El vocativo épico atrae la atención del lector o del oyente y lo ubica en la trama del relato.
15. Los documentos escritos antiguos pertenecen a la época de la dinastía Shang.
16. Los métodos modernos de investigación han permitido estudiar al hombre prehistórico.
17. La Química usa técnicas para analizar sustancias.
18. Se usa el guión menor para formar adjetivos compuestos.
19. Los documentos escritos más antiguos fueron encontrados en Mesopotamia y Egipto.
20. La Botánica estudia el polen fósil y ha logrado analizar las características de la vegetación e inferir los climas.

21. La arqueología usa nuevas técnicas para excavar y localizar y estudiar los restos materiales y huellas y señales que el hombre ha dejado en el pasado para reconstruir y comprender su vida en todos los aspectos posibles.
22. La Prehistoria abarca desde la aparición de la humanidad hasta la invención de la escritura.
23. El agua es indispensable para la vida.
24. La numeración arábiga procede de India.
25. Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806.
26. Los primeros homínidos eran recolectores y sólo comían carne cuando encontraban los restos abandonados por otros animales.
27. La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula.
28. El surgimiento de la agricultura fue posible por los cambios climáticos que crearon un ambiente propicio para la reproducción, el cuidado y la selección de plantas.
29. El mar Mediterráneo es un mar poco profundo, con escasas corrientes marinas, lo cual facilita la navegación.
30. Los primeros pobladores de América llegaron hace 30 mil años, aproximadamente, siguiendo a las manadas de animales que acostumbraban cazar.
31. En América la agricultura inició entre el 8000 y el 5000 a.C..
32. La agrupación de seres humanos en un mismo espacio favoreció el intercambio de conocimientos y el desarrollo de las ciencias y el arte.
33. El mamut era un animal de gran tamaño al que se cazaba mediante diversas técnicas.
34. Los mamuts migraron de África hace 3.5 millones de años y llegaron a vivir en Europa, Asia y América.
35. Las civilizaciones agrícolas también desarrollaron la ciencia.
36. Los cretenses eran un pueblo pacífico de navegantes que estuvo en contacto con Egipto y Medio Oriente.
37. Los primeros griegos se organizaron en grupos que tenían lazos familiares.
38. Esparta era gobernada por reyes.
39. En Atenas los gobernantes eran elegidos por el voto de los ciudadanos.
40. El término democracia significa gobierno del pueblo.

41. La democracia ateniense se basaba en la participación de todos los ciudadanos en la vida política.
42. La cultura griega alcanzó su esplendor en el siglo V a.C..
43. La civilización helenística llegó a su fin en el siglo I a.C., cuando Roma consumó la conquista de Egipto.
44. La historia de la civilización romana se divide en tres periodos.
45. Roma fue gobernada por siete reyes, etruscos y latinos, en diferentes periodos.
46. Durante la república comenzó la expansión de los romanos.
47. El último periodo de la civilización romana fue el imperio, que abarcó desde el año 27 a.C. hasta el año 476 d.C..
48. El primer emperador de Roma fue el político y militar Octavio Augusto.
49. Roma no imponía ideas políticas o credos en sus territorios.
50. Los habitantes de la antigua Roma se ocupaban en diversos trabajos.
51. Los mesopotámicos nos legaron la rueda y la escritura.
52. El pueblo griego nos dejó como herencia la democracia.
53. La palabra Mesoamérica fue creada por un antropólogo en el siglo xx para definir el lugar en el que florecieron las culturas más desarrolladas del México antiguo.
54. El preclásico duró aproximadamente 2700 años, ya que inició en el 2500 a.C. y concluyó hacia el 200 d.C..
55. El periodo clásico abarcó del 200 al 900 d.C..
56. Para su estudio, el sistema nervioso se divide en sistema nervioso central y sistema nervioso periférico.
57. El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas.
58. El encéfalo se encuentra dentro del cráneo y consta de varios órganos, cada uno de éstos realiza distintas funciones.
59. El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada.

60. El cerebro almacena enormes cantidades de información, realiza millones de actividades todos los días y es capaz de llevar a cabo varias acciones al mismo tiempo, como interpretar lo que los ojos ven, pensar y controlar muchos de los movimientos del cuerpo.
61. El cerebro es un órgano tan complejo que no se conoce al detalle su funcionamiento completo.
62. El Tálamo se halla en el centro del encéfalo, recibe las señales enviadas por los sentidos y las reenvía a distintas áreas del cerebro para su procesamiento.
63. El Cerebelo es el segundo órgano más grande del encéfalo, sirve para mantener el equilibrio y controlar los movimientos finos.
64. El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed.
65. El Hipotálamo es el responsable de las manifestaciones emocionales (como la amistad, el cariño y el amor).
66. El Bulbo raquídeo es el encargado de transmitir mensajes entre el cerebro y el cuerpo, y controla funciones básicas como el latido del corazón, la digestión y la respiración.
67. La Médula espinal es la prolongación del encéfalo, tiene forma de cordón y corre por dentro de la columna vertebral, que la protege.
68. De la médula espinal nacen los nervios periféricos, que permiten movimientos voluntarios e involuntarios, sensaciones y reflejos.

FactSpCIC: El estándar de oro

En esta sección se presenta el corpus utilizado como estándar de oro (Gold Standard) para evaluar el sistema. Se organiza de la siguiente forma: la oración esta enumerada y en seguida se muestra una tabla que contiene sus respectivos hechos obtenidos por las personas encargadas de construir el estándar de oro. Los hechos están redactados tal y como las personas lo hicieron, y al tercer componente de la tripleta del hecho ([*objeto/Complemento*]) se le llama solamente [*Complemento*]. Son un total de 68 oraciones y 166 hechos.

1. Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.

No.	Sujeto	Verbo	Complemento
1	Los egipcios	se caracterizaron	por sus creencias relacionadas con la muerte

2. El libro de los muertos es otro de los escritos que se han encontrado en diversas tumbas egipcias.

No.	Sujeto	Verbo	Complemento
1	libro de los muertos	es	otro de los escritos
2	escritos	se han encontrado	en diversas tumbas egipcias

3. El ritmo es la alternancia de sílabas átonas con sílabas tónicas.

No.	Sujeto	Verbo	Complemento
1	El ritmo	es	la alternancia de sílabas átonas con sílabas tónicas

4. Los datos de las fuentes consultadas deben registrarse en fichas bibliográficas.

No.	Sujeto	Verbo	Complemento
1	Los datos de las fuentes consultadas	deben registrarse	en fichas bibliográficas

5. Redactar un borrador en el cual se deben considerar las partes que conforman un trabajo de investigación.

No.	Sujeto	Verbo	Complemento
1	Redactar borrador	se deben considerar	las partes
2	Partes	conforman	un trabajo de investigación

6. Esta sección debe ser breve e interesante.

No.	Sujeto	Verbo	Complemento
1	Esta sección	debe ser	breve
2	Esta sección	debe ser	interesante

7. Las conclusiones sintetizan los argumentos presentados en el desarrollo del trabajo.

No.	Sujeto	Verbo	Complemento
1	Las conclusiones	sintetizan	los argumentos presentados
2	Los argumentos	presentados	en el desarrollo del trabajo

8. La bibliografía es el listado alfabético de todas las fuentes consultadas para la elaboración del trabajo.

No.	Sujeto	Verbo	Complemento
1	Las bibliografía	es	el listado alfabético de todas las fuentes consultadas para la elaboración del trabajo

9. La bibliografía se estructura con los datos de las fichas bibliográficas de esos textos.

No.	Sujeto	Verbo	Complemento
1	Las bibliografía	se estructura	con los datos de las fichas bibliográficas de esos textos.

10. Las repeticiones sirven para acentuar las emociones y sentimientos del hablante lírico.

No.	Sujeto	Verbo	Complemento
1	Las repeticiones	acentuar	las emociones
2	Las repeticiones	acentuar	sentimientos del hablante lírico

11. Es muy común el empleo de llaves para estructurar un cuadro sinóptico.

No.	Sujeto	Verbo	Complemento
1	El empleo de llaves	estructurar	un cuadro sinóptico

12. El epíteto heroico es la expresión que menciona una característica o cualidad del personaje u objeto nombrado.

No.	Sujeto	Verbo	Complemento
1	El epíteto heroico	es	la expresión
2	La expresión	menciona	una característica
3	La expresión	menciona	cualidad del personaje
4	La expresión	menciona	objeto nombrado

13. El vocativo épico es un enunciado exclamativo intercalado en una oración.

No.	Sujeto	Verbo	Complemento
1	El vocativo épico	es	un enunciado exclamativo en una oración

14. El vocativo épico atrae la atención del lector o del oyente y lo ubica en la trama del relato.

No.	Sujeto	Verbo	Complemento
1	El vocativo épico	atrae	la atención del lector
2	El vocativo épico	atrae	la atención del oyente
3	El vocativo épico	ubica	en la trama del relato

15. Los documentos escritos antiguos pertenecen a la época de la dinastía Shang.

No.	Sujeto	Verbo	Complemento
1	Los documentos escritos antiguos	pertenecen	a la época de la dinastía Shang

16. Los métodos modernos de investigación han permitido estudiar al hombre prehistórico.

No.	Sujeto	Verbo	Complemento
1	Los métodos modernos de investigación	han permitido estudiar	al hombre prehistórico

17. La Química usa técnicas para analizar sustancias.

No.	Sujeto	Verbo	Complemento
1	La química	usa	técnicas
2	La química	analizar	sustancias

18. Se usa el guión menor para formar adjetivos compuestos.

No.	Sujeto	Verbo	Complemento
1	El guión menor	formar	adjetivos compuestos

19. Los documentos escritos más antiguos fueron encontrados en Mesopotamia y Egipto.

No.	Sujeto	Verbo	Complemento
1	Los documentos escritos más antiguos	fueron encontrados	en Mesopotamia
2	Los documentos escritos más antiguos	fueron encontrados	en Egipto

20. La Botánica estudia el polen fósil y ha logrado analizar las características de la vegetación e inferir los climas.

No.	Sujeto	Verbo	Complemento
1	La Botánica	estudia	el polen fósil
2	La Botánica	ha logrado analizar	las características de la vegetación
3	La Botánica	ha logrado inferir	los climas

21. La arqueología usa nuevas técnicas para excavar y localizar y estudiar los restos materiales y huellas y señales que el hombre ha dejado en el pasado para reconstruir y comprender su vida en todos los aspectos posibles.

No.	Sujeto	Verbo	Complemento
1	La arqueología	usa	técnicas nuevas
2	La arqueología	excavar	restos materiales
3	La arqueología	localizar	restos materiales
4	La arqueología	estudiar	restos materiales
5	La arqueología	excavar	huellas
6	La arqueología	localizar	huellas
7	La arqueología	estudiar	huellas
8	La arqueología	excavar	señales
9	La arqueología	localizar	señales
10	La arqueología	estudiar	señales

22. La Prehistoria abarca desde la aparición de la humanidad hasta la invención de la escritura.

No.	Sujeto	Verbo	Complemento
1	La prehistoria	abarca	desde la aparición de la humanidad
2	La prehistoria	abarca	hasta la invención de la escritura

23. El agua es indispensable para la vida.

No.	Sujeto	Verbo	Complemento
1	El agua	es	indispensable
2	El agua	es	para la vida

24. La numeración arábiga procede de India.

No.	Sujeto	Verbo	Complemento
1	La numeración arábiga	procede	de India

25. Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806.

No.	Sujeto	Verbo	Complemento
1	Benito Juárez	nació	en San Pablo Guelatao, Oaxaca
2	Benito Juárez	nació	en 1806

26. Los primeros homínidos eran recolectores y sólo comían carne cuando encontraban los restos abandonados por otros animales.

No.	Sujeto	Verbo	Complemento
1	Los primeros homínidos	eran	recolectores
2	Los primeros homínidos	comían	carne
3	Los primeros homínidos	encontraban	restos abandonados
4	Restos	abandonados	por otros animales

27. La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula.

No.	Sujeto	Verbo	Complemento
1	La civilización china	nos heredó	el papel
2	La civilización china	nos heredó	la pólvora
3	La civilización china	nos heredó	una forma de imprenta rudimentaria
4	La civilización china	nos heredó	la brújula

28. El surgimiento de la agricultura fue posible por los cambios climáticos que crearon un ambiente propicio para la reproducción, el cuidado y la selección de plantas.

No.	Sujeto	Verbo	Complemento
1	El surgimiento de la agricultura	fue	posible por los cambios climáticos
2	Los cambios climáticos	crearon	un ambiente propicio para la reproducción de plantas
3	Los cambios climáticos	crearon	un ambiente propicio para el cuidado de plantas
4	Los cambios climáticos	crearon	un ambiente propicio para la selección de plantas

29. El mar Mediterráneo es un mar poco profundo, con escasas corrientes marinas, lo cual facilita la navegación.

No.	Sujeto	Verbo	Complemento
1	El mar Mediterráneo	es	un mar
2	El mar Mediterráneo	es	poco profundo
3	El mar Mediterráneo	es	con escasas corrientes marinas
4	Escasas corrientes marinas	facilita	la navegación

30. Los primeros pobladores de América llegaron hace 30 mil años, aproximadamente, siguiendo a las manadas de animales que acostumbraban cazar.

No.	Sujeto	Verbo	Complemento
1	Los primeros pobladores de América	llegaron	hace 30 mil años, aproximadamente
2	Los primeros pobladores de América	llegaron siguiendo	a las manadas de animales
3	manadas de animales	acostumbraban	cazar

31. En América la agricultura inició entre el 8000 y el 5000 a.C..

No.	Sujeto	Verbo	Complemento
1	En América la agricultura	inició	entre el 8000 y el 5000 a.C.

32. La agrupación de seres humanos en un mismo espacio favoreció el intercambio de conocimientos y el desarrollo de las ciencias y el arte.

No.	Sujeto	Verbo	Complemento
1	La agrupación de seres humanos en un mismo espacio	favoreció	el intercambio de conocimientos
2	La agrupación de seres humanos en un mismo espacio	favoreció	el desarrollo de las ciencias
3	La agrupación de seres humanos en un mismo espacio	favoreció	el desarrollo del arte

33. El mamut era un animal de gran tamaño al que se cazaba mediante diversas técnicas.

No.	Sujeto	Verbo	Complemento
1	El mamut	era	un animal
2	El mamut	era	de gran tamaño
3	Animal de gran tamaño	se cazaba	mediante diversas técnicas

34. Los mamuts migraron de África hace 3.5 millones de años y llegaron a vivir en Europa, Asia y América.

No.	Sujeto	Verbo	Complemento
1	Los mamuts	migraron	de África
2	Los mamuts	migraron	hace 3.5 millones de años
3	Los mamuts	llegaron a vivir	en Europa
4	Los mamuts	llegaron a vivir	en Asia
5	Los mamuts	llegaron a vivir	en América

35. Las civilizaciones agrícolas también desarrollaron la ciencia.

No.	Sujeto	Verbo	Complemento
1	Las civilizaciones agrícolas	desarrollaron	la ciencia

36. Los cretenses eran un pueblo pacífico de navegantes que estuvo en contacto con Egipto y Medio Oriente.

No.	Sujeto	Verbo	Complemento
1	Los cretenses	eran	un pueblo pacífico
2	Los cretenses	eran	un pueblo de navegantes
3	Pueblo pacífico de navegantes	estuvo	en contacto con Egipto
4	Pueblo pacífico de navegantes	estuvo	en contacto con Medio Oriente

37. Los primeros griegos se organizaron en grupos que tenían lazos familiares.

No.	Sujeto	Verbo	Complemento
1	Los primeros griegos	se organizaron	en grupos
2	Grupos	tenían	lazos familiares

38. Esparta era gobernada por reyes.

No.	Sujeto	Verbo	Complemento
1	Esparta	era gobernada	por reyes

39. En Atenas los gobernantes eran elegidos por el voto de los ciudadanos.

No.	Sujeto	Verbo	Complemento
1	En Atenas los gobernantes	eran	elegidos por el voto de los ciudadanos

40. El término democracia significa gobierno del pueblo.

No.	Sujeto	Verbo	Complemento
1	El término democracia	significa	gobierno del pueblo

41. La democracia ateniense se basaba en la participación de todos los ciudadanos en la vida política.

No.	Sujeto	Verbo	Complemento
1	La democracia ateniense	se basaba	en la participación de todos los ciudadanos en la vida política

42. La cultura griega alcanzó su esplendor en el siglo V a.C..

No.	Sujeto	Verbo	Complemento
1	La cultura griega	alcanzó	su esplendor en el siglo V a.C.

43. La civilización helenística llegó a su fin en el siglo I a.C., cuando Roma consumó la conquista de Egipto.

No.	Sujeto	Verbo	Complemento
1	La civilización helenística	llegó	a su fin en el siglo I a.C.
2	Roma	consumó	la conquista de Egipto

44. La historia de la civilización romana se divide en tres periodos.

No.	Sujeto	Verbo	Complemento
1	La historia de la civilización romana	se divide	en tres periodos

45. Roma fue gobernada por siete reyes, etruscos y latinos, en diferentes periodos.

No.	Sujeto	Verbo	Complemento
1	Roma	fue gobernada	por siete reyes
2	Roma	fue gobernada	por Etruscos
3	Roma	fue gobernada	por latinos
4	Roma	fue gobernada	en diferentes periodos

46. Durante la república comenzó la expansión de los romanos.

No.	Sujeto	Verbo	Complemento
1	La expansión de los romanos	comenzó	durante la república

47. El último periodo de la civilización romana fue el imperio, que abarcó desde el año 27 a.C. hasta el año 476 d.C..

No.	Sujeto	Verbo	Complemento
1	El último periodo de la civilización romana	fue	el imperio
2	El imperio	abarcó	desde el año 27 a.C.
3	El imperio	abarcó	hasta el año 476 d.C.

48. El primer emperador de Roma fue el político y militar Octavio Augusto.

No.	Sujeto	Verbo	Complemento
1	El primer emperador de Roma	fue	el político
2	El primer emperador de Roma	fue	el militar Octavio Augusto

49. Roma no imponía ideas políticas o credos en sus territorios.

No.	Sujeto	Verbo	Complemento
1	Roma	no imponía	ideas políticas en sus territorios
2	Roma	no imponía	credos en sus territorios

50. Los habitantes de la antigua Roma se ocupaban en diversos trabajos.

No.	Sujeto	Verbo	Complemento
1	Los habitantes de la antigua Roma	se ocupaban	en diversos trabajos

51. Los mesopotámicos nos legaron la rueda y la escritura.

No.	Sujeto	Verbo	Complemento
1	Los mesopotámicos	nos legaron	la rueda
2	Los mesopotámicos	nos legaron	la escritura

52. El pueblo griego nos dejó como herencia la democracia.

No.	Sujeto	Verbo	Complemento
1	El pueblo griego	nos dejó	como herencia la democracia

53. La palabra Mesoamérica fue creada por un antropólogo en el siglo xx para definir el lugar en el que florecieron las culturas más desarrolladas del México antiguo.

No.	Sujeto	Verbo	Complemento
1	La palabra Mesoamérica	fue creada	por un antropólogo
2	La palabra Mesoamérica	fue creada	en el siglo xx
3	La palabra Mesoamérica	definir	el lugar en el que florecieron las culturas más desarrolladas del México antiguo

54. El preclásico duró aproximadamente 2700 años, ya que inició en el 2500 a.C. y concluyó hacia el 200 d.C..

No.	Sujeto	Verbo	Complemento
1	El preclásico	duró	aproximadamente 2700 años
2	El preclásico	inició	en el 2500 a.C.
3	El preclásico	concluyó	hacia el 200 d.C.

55. El periodo clásico abarcó del 200 al 900 d.C..

No.	Sujeto	Verbo	Complemento
1	El periodo clásico	abarcó	del 200 al 900 d.C.

56. Para su estudio, el sistema nervioso se divide en sistema nervioso central y sistema nervioso periférico.

No.	Sujeto	Verbo	Complemento
1	El sistema nervioso	se divide	en sistema nervioso central
2	El sistema nervioso	se divide	sistema nervioso periférico

57. El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas.

No.	Sujeto	Verbo	Complemento
1	El sistema nervioso periférico	lo conforman	los nervios
2	Los nervios	nacen	del cerebro
3	Los nervios	nacen	de la médula espinal
4	Los nervios	llegan	a todas las partes del cuerpo por medio de fibras nerviosas

58. El encéfalo se encuentra dentro del cráneo y consta de varios órganos, cada uno de éstos realiza distintas funciones.

No.	Sujeto	Verbo	Complemento
1	El encéfalo	se encuentra	dentro del cráneo
2	El encéfalo	consta	de varios órganos
3	Órganos	realizan	distintas funciones

59. El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada.

No.	Sujeto	Verbo	Complemento
1	El Cerebro	es	el órgano más grande del encéfalo
2	El Cerebro	está dividido	en dos mitades
3	El Cerebro	está dividido	hemisferios
4	El Cerebro	presenta	hendiduras
5	El Cerebro	presenta	pliegues
6	Pliegues	le dan	el aspecto de una nuez pelada

60. El cerebro almacena enormes cantidades de información, realiza millones de actividades todos los días y es capaz de llevar a cabo varias acciones al mismo tiempo, como interpretar lo que los ojos ven, pensar y controlar muchos de los movimientos del cuerpo.

No.	Sujeto	Verbo	Complemento
1	El Cerebro	almacena	enormes cantidades de información
2	El Cerebro	realiza	millones de actividades todos los días
3	El Cerebro	es	capaz de llevar a cabo varias acciones al mismo tiempo
4	El Cerebro	interpretar	lo que los ojos ven
5	El Cerebro	controlar	muchos de los movimientos del cuerpo

61. El cerebro es un órgano tan complejo que no se conoce al detalle su funcionamiento completo.

No.	Sujeto	Verbo	Complemento
1	El Cerebro	es	un órgano tan complejo que no se conoce al detalle su funcionamiento completo

62. El Tálamo se halla en el centro del encéfalo, recibe las señales enviadas por los sentidos y las reenvía a distintas áreas del cerebro para su procesamiento.

No.	Sujeto	Verbo	Complemento
1	El Tálamo	se halla	en el centro del encéfalo
2	El Tálamo	recibe	las señales enviadas por los sentidos
3	El Tálamo	las reenvía	a distintas áreas del cerebro para su procesamiento

63. El Cerebelo es el segundo órgano más grande del encéfalo, sirve para mantener el equilibrio y controlar los movimientos finos.

No.	Sujeto	Verbo	Complemento
1	El Cerebelo	es	el segundo órgano más grande del encéfalo
2	El Cerebelo	mantener	el equilibrio
3	El Cerebelo	controlar	los movimientos finos

64. El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed.

No.	Sujeto	Verbo	Complemento
1	El Hipotálamo	se encarga	de algunas funciones corporales
2	El Hipotálamo	regular	la temperatura
3	El Hipotálamo	percibir	la señal de sueño
4	El Hipotálamo	percibir	hambre
5	El Hipotálamo	percibir	sed

65. El Hipotálamo es el responsable de las manifestaciones emocionales (como la amistad, el cariño y el amor).

No.	Sujeto	Verbo	Complemento
1	El Hipotálamo	es	el responsable de las manifestaciones emocionales (como la amistad, el cariño y el amor)

66. El Bulbo raquídeo es el encargado de transmitir mensajes entre el cerebro y el cuerpo, y controla funciones básicas como el latido del corazón, la digestión y la respiración.

No.	Sujeto	Verbo	Complemento
1	El Bulbo raquídeo	es	el encargado de transmitir mensajes entre el cerebro y el cuerpo
2	El Bulbo raquídeo	controla	funciones básicas
3	El Bulbo raquídeo	controla	el latido del corazón
4	El Bulbo raquídeo	controla	la digestión
5	El Bulbo raquídeo	controla	la respiración

67. La Médula espinal es la prolongación del encéfalo, tiene forma de cordón y corre por dentro de la columna vertebral, que la protege.

No.	Sujeto	Verbo	Complemento
1	La Médula espinal	es	la prolongación del encéfalo
2	La Médula espinal	tiene	forma de cordón
3	La Médula espinal	corre	por dentro de la columna vertebral

68. De la médula espinal nacen los nervios periféricos, que permiten movimientos voluntarios e involuntarios, sensaciones y reflejos.

No.	Sujeto	Verbo	Complemento
1	De la médula espinal	nacen	los nervios periféricos
2	Los nervios periféricos	permiten	movimientos voluntarios
3	Los nervios periféricos	permiten	movimientos involuntarios
4	Los nervios periféricos	permiten	sensaciones
5	Los nervios periféricos	permiten	reflejos

Hechos extraídos por FES 2012

En la Tabla A.1 se muestran los hechos extraídos por el sistema de cada una de las oraciones del corpus. En la columna "Tipo" se marcan los hechos con las siguientes letras: "C - Correctos", "I - Incorrectos"; con "EX" los hechos que no se consideran para la evaluación porque aunque son correctos, en el estándar de oro se presentan agrupados en un solo hecho.

Tabla A.1 Hechos extraídos del corpus de prueba, por FES 2012.

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
1	**Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.**			**2**
1	egipcios	caracterizaron	por creencias relacionadas	**C**
2	egipcios	caracterizaron	con muerte	**I**
2	**El libro de los muertos es otro de los escritos que se han encontrado en diversas tumbas egipcias.**			**3**
1	libro de muertos	es	otro de escritos	**C**
2	libro de muertos	es	otro	**EX**
3	otro de escritos	encontrado	en tumbas diversas egipcias	**C**
3	**El ritmo es la alternancia de sílabas átonas con sílabas tónicas.**			**0**
4	**Los datos de las fuentes consultadas deben registrarse en fichas bibliográficas.**			**1**
1	datos de fuentes consultadas	registrar	en fichas bibliográficas	**C**
5	**Redactar un borrador en el cual se deben considerar las partes que conforman un trabajo de investigación.**			**0**
6	**Esta sección debe ser breve e interesante.**			**2**
1	sección	ser	breve	**C**
2	sección	ser	interesante	**C**
7	**Las conclusiones sintetizan los argumentos presentados en el desarrollo del trabajo.**			**0**
8	**La bibliografía es el listado alfabético de todas las fuentes consultadas para la elaboración del trabajo.**			**3**
1	bibliografía	es	listado alfabético	**C**
2	bibliografía	es	listado de fuentes consultadas	**EX**
3	bibliografía	es	para elaboración de trabajo	**EX**
9	**La bibliografía se estructura con los datos de las fichas bibliográficas de esos textos.**			**1**
1	bibliografía	estructura	con datos de fichas bibliográficas de textos	**C**
10	**Las repeticiones sirven para acentuar las emociones y sentimientos del hablante lírico.**			**2**
1	repeticiones	acentuar	emociones	**C**
2	repeticiones	acentuar	sentimientos de hablante lírico	**C**
11	**Es muy común el empleo de llaves para estructurar un cuadro sinóptico.**			**0**
12	**El epíteto heroico es la expresión que menciona una característica o cualidad del personaje u objeto nombrado.**			**4**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
1	epíteto heroico	es	expresión	**C**
2	expresión	menciona	característica	**C**
3	expresión	menciona	cualidad de personaje	**C**
4	expresión	menciona	objeto nombrado	**C**
13	**El vocativo épico es un enunciado exclamativo intercalado en una oración.**			**2**
1	vocativo épico	es	enunciado exclamativo intercalado	**C**
2	vocativo épico	es	en oración	**I**
14	**El vocativo épico atrae la atención del lector o del oyente y lo ubica en la trama del relato.**			**4**
1	vocativo épico	atrae	atención de lector	**C**
2	vocativo épico	atrae	de oyente	**C**
3	vocativo épico	ubica	en trama	**C**
4	vocativo épico	ubica	de relato	**I**
15	**Los documentos escritos antiguos pertenecen a la época de la dinastía Shang.**			**1**
1	documentos escritos antiguos	pertenecen	a época de dinastía Shang	**C**
16	**Los métodos modernos de investigación han permitido estudiar al hombre prehistórico.**			**1**
1	métodos modernos de investigación	estudiar	a hombre prehistórico	**C**
17	**La Química usa técnicas para analizar sustancias.**			**2**
1	Química	usa	técnicas	**C**
2	Química	analizar	sustancias	**C**
18	**Se usa el guión menor para formar adjetivos compuestos.**			**0**
19	**Los documentos escritos más antiguos fueron encontrados en Mesopotamia y Egipto.**			**2**
1	documentos escritos antiguos más	encontrados	en Mesopotamia	**C**
2	documentos escritos antiguos más	encontrados	Egipto	**C**
20	**La Botánica estudia el polen fósil y ha logrado analizar las características de la vegetación e inferir los climas.**			**3**
1	Botánica	estudia	polen fósil	**C**
2	Botánica	analizar	características de vegetación	**C**
3	Botánica	inferir	climas	**C**
21	**La arqueología usa nuevas técnicas para excavar y localizar y estudiar los restos materiales y huellas y señales que el hombre ha dejado en el pasado para reconstruir y comprender su vida en todos los aspectos posibles.**			**7**
1	arqueología	usa	técnicas nuevas	**C**
2	arqueología	estudiar	restos materiales	**C**
3	arqueología	estudiar	huellas	**C**
4	arqueología	estudiar	señales	**C**
5	señales	dejado	en pasado	**I**
6	arqueología	comprender	vida	**EX**
7	arqueología	comprender	en aspectos posibles	**EX**
22	**La Prehistoria abarca desde la aparición de la humanidad hasta la invención de la escritura.**			**2**
1	Prehistoria	abarca	desde aparición de humanidad	**C**
2	Prehistoria	abarca	hasta invención de escritura	**C**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
23	**El agua es indispensable para la vida.**			**2**
1	agua	es	indispensable	**C**
2	agua	es	para vida	**C**
24	**La numeración arábiga procede de India.**			**1**
1	numeración arábiga	procede	de India	**C**
25	**Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806.**			**3**
1	Benito_Juárez	nació	en San_Pablo_Guelatao	**C**
2	Benito_Juárez	nació	Oaxaca	**EX**
3	Benito_Juárez	nació	en 1806	**C**
26	**Los primeros homínidos eran recolectores y sólo comían carne cuando encontraban los restos abandonados por otros animales.**			**4**
1	homínidos primeros	eran	recolectores	**C**
2	homínidos primeros	comían	carne	**C**
3	homínidos primeros	encontraban	restos abandonados	**C**
4	homínidos primeros	encontraban	por animales	**I**
27	**La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula.**			**4**
1	civilización China	heredó	papel	**C**
2	civilización China	heredó	pólvora	**C**
3	civilización China	heredó	forma de imprenta rudimentaria	**C**
4	civilización China	heredó	brújula	**C**
28	**El surgimiento de la agricultura fue posible por los cambios climáticos que crearon un ambiente propicio para la reproducción, el cuidado y la selección de plantas.**			**6**
1	surgimiento de agricultura	fue	posible	**C**
2	surgimiento de agricultura	fue	por cambios climáticos	**EX**
3	por cambios climáticos	crearon	ambiente propicio	**C**
4	por cambios climáticos	crearon	para reproducción	**I**
5	por cambios climáticos	crearon	cuidado	**I**
6	por cambios climáticos	crearon	selección de plantas	**I**
29	**El mar Mediterráneo es un mar poco profundo, con escasas corrientes marinas, lo cual facilita la navegación.**			**3**
1	mar Mediterráneo	es	mar profundo poco	**C**
2	mar Mediterráneo	es	con corrientes escasas marinas	**C**
3	con corrientes escasas marinas	facilita	navegación	**C**
30	**Los primeros pobladores de América llegaron hace 30 mil años, aproximadamente, siguiendo a las manadas de animales que acostumbraban cazar.**			**2**
1	pobladores primeros de América	llegaron	hace años 30_mil	**C**
2	pobladores primeros de América	llegaron	siguiendo a manadas de animales	**C**
31	**En América la agricultura inició entre el 8000 y el 5000 a.C.**			**3**
1	agricultura	inició	En América	**EX**
2	agricultura	inició	entre 8000 y 5000	**C**
3	agricultura	inició	a.C.	**EX**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
32	**La agrupación de seres humanos en un mismo espacio favoreció el intercambio de conocimientos y el desarrollo de las ciencias y el arte.**			**4**
1	agrupación de seres humanos	favoreció	en espacio mismo	**I**
2	agrupación de seres humanos	favoreció	intercambio de conocimientos	**C**
3	agrupación de seres humanos	favoreció	desarrollo de ciencias	**C**
4	agrupación de seres humanos	favoreció	arte	**C**
33	**El mamut era un animal de gran tamaño al que se cazaba mediante diversas técnicas.**			**3**
1	mamut	era	animal de tamaño gran	**C**
2	mamut	era	animal	**C**
3	animal de tamaño gran	cazaba	mediante técnicas diversas	**C**
34	**Los mamuts migraron de África hace 3.5 millones de años y llegaron a vivir en Europa, Asia y América.**			**5**
1	mamuts	migraron	de África	**C**
2	mamuts	migraron	hace años 3.5_millones de	**C**
3	mamuts	vivir	en Europa	**C**
4	mamuts	vivir	Asia	**C**
5	mamuts	vivir	América	**C**
35	**Las civilizaciones agrícolas también desarrollaron la ciencia.**			**1**
1	civilizaciones agrícolas	desarrollaron	ciencia	**C**
36	**Los cretenses eran un pueblo pacífico de navegantes que estuvo en contacto con Egipto y Medio Oriente.**			**5**
1	cretenses	eran	pueblo pacífico	**C**
2	cretenses	eran	pueblo de navegantes	**C**
3	pueblo pacífico de navegantes	estuvo	en contacto	**C**
4	pueblo pacífico de navegantes	estuvo	con Egipto	**I**
5	pueblo pacífico de navegantes	estuvo	Medio_Oriente	**I**
37	**Los primeros griegos se organizaron en grupos que tenían lazos familiares.**			**2**
1	griegos primeros	organizaron	en grupos	**C**
2	en grupos	tenían	lazos familiares	**C**
38	**Esparta era gobernada por reyes.**			**1**
1	Esparta	gobernada	por reyes	**C**
39	**En Atenas los gobernantes eran elegidos por el voto de los ciudadanos.**			**2**
1	gobernantes	elegidos	En Atenas	**EX**
2	gobernantes	elegidos	por voto de ciudadanos	**C**
40	**El término democracia significa gobierno del pueblo.**			**0**
41	**La democracia ateniense se basaba en la participación de todos los ciudadanos en la vida política.**			**2**
1	democracia ateniense	basaba	en participación de ciudadanos	**C**
2	democracia ateniense	basaba	en vida política	**I**
42	**La cultura griega alcanzó su esplendor en el siglo v a.C.**			**2**
1	cultura griega	alcanzó	esplendor	**C**
2	cultura griega	alcanzó	en siglo_V_a.C.	**EX**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
43	**La civilización helenística llegó a su fin en el siglo I a.C., cuando Roma consumó la conquista de Egipto.**			**3**
1	civilización helenística	llegó	a fin	**C**
2	civilización helenística	llegó	en siglo_I_a.C.	**EX**
3	Roma	consumó	conquista de Egipto	**C**
44	**La historia de la civilización romana se divide en tres periodos.**			**1**
1	historia de civilización romana	divide	en periodos tres	**C**
45	**Roma fue gobernada por siete reyes, etruscos y latinos, en diferentes periodos.**			**4**
1	Roma	gobernada	por reyes siete	**C**
2	Roma	gobernada	etruscos	**C**
3	Roma	gobernada	latinos	**C**
4	Roma	gobernada	en periodos diferentes	**C**
46	**Durante la república comenzó la expansión de los romanos.**			**0**
47	**El último periodo de la civilización romana fue el imperio, que abarcó desde el año 27 a.C. hasta el año 476 d.C.**			**3**
1	periodo último de civilización romana	fue	imperio	**C**
2	imperio	abarcó	desde año_27_a.C.	**C**
3	imperio	abarcó	hasta año_476_d.C.	**C**
48	**El primer emperador de Roma fue el político y militar Octavio Augusto.**			**2**
1	emperador primer de Roma	fue	Octavio_Augusto político	**C**
2	emperador primer de Roma	fue	Octavio_Augusto militar	**C**
49	**Roma no imponía ideas políticas o credos en sus territorios.**			**3**
1	Roma	imponía	ideas políticas	**C**
2	Roma	imponía	credos	**C**
3	Roma	imponía	en territorios	**EX**
50	**Los habitantes de la antigua Roma se ocupaban en diversos trabajos.**			**1**
1	habitantes de Roma antigua	ocupaban	en trabajos diversos	**C**
51	**Los mesopotámicos nos legaron la rueda y la escritura.**			**2**
1	mesopotámicos	legaron	rueda	**C**
2	mesopotámicos	legaron	escritura	**C**
52	**El pueblo griego nos dejó como herencia la democracia.**			**2**
1	pueblo griego	dejó	como herencia	**C**
2	pueblo griego	dejó	democracia	**EX**
53	**La palabra Mesoamérica fue creada por un antropólogo en el siglo xx para definir el lugar en el que florecieron las culturas más desarrolladas del México antiguo.**			**3**
1	palabra Mesoamérica	creada	por antropólogo	**C**
2	palabra Mesoamérica	creada	en siglo xx	**C**
3	palabra Mesoamérica	definir	lugar	**C**
54	**El preclásico duró aproximadamente 2700 años, ya que inició en el 2500 a.C. y concluyó hacia el 200 d.C.**			**5**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
1	preclásico	duró	años 2700	**C**
2	preclásico	inició	en 2500	**C**
3	preclásico	inició	a.C.	**EX**
4	preclásico	concluyó	hacia 200	**C**
5	preclásico	concluyó	d.C.	**EX**
55	**El periodo clásico abarcó del 200 al 900 d.C.**			**0**
56	**Para su estudio, el sistema nervioso se divide en sistema nervioso central y sistema nervioso periférico.**			**3**
1	sistema nervioso	divide	Para estudio	**EX**
2	sistema nervioso	divide	en sistema nervioso central	**C**
3	sistema nervioso	divide	sistema nervioso periférico	**C**
57	**El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas.**			**6**
1	sistema nervioso periférico	conforman	nervios	**C**
2	nervios	nacen	de cerebro	**C**
3	nervios	nacen	de médula espinal	**C**
4	nervios	llegan	a partes	**C**
5	nervios	llegan	de cuerpo	**I**
6	nervios	llegan	por_medio_de fibras nerviosas	**EX**
58	**El encéfalo se encuentra dentro del cráneo y consta de varios órganos, cada uno de éstos realiza distintas funciones.**			**3**
1	encéfalo	encuentra	dentro_de cráneo	**C**
2	encéfalo	consta	de órganos	**C**
3	encéfalo	realiza	funciones distintas	**C**
59	**El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada.**			**7**
1	Cerebro	es	órgano grande más	**C**
2	Cerebro	es	órgano de encéfalo	**EX**
3	Cerebro	dividido	en mitades dos	**C**
4	Cerebro	dividido	hemisferios	**C**
5	Cerebro	presenta	hendiduras	**C**
6	Cerebro	presenta	pliegues	**C**
7	pliegues	dan	aspecto de nuez pelada	**C**
60	**El cerebro almacena enormes cantidades de información, realiza millones de actividades todos los días y es capaz de llevar a cabo varias acciones al mismo tiempo, como interpretar lo que los ojos ven, pensar y controlar muchos de los movimientos del cuerpo.**			**8**
1	cerebro	almacena	cantidades enormes	**C**
2	cerebro	almacena	de información	**I**
3	cerebro	realiza	millones de actividades	**C**
4	cerebro	realiza	días	**I**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
5	cerebro	es	capaz	**C**
6	cerebro	llevar_a_cabo	acciones	**EX**
7	cerebro	llevar_a_cabo	a tiempo mismo	**I**
8	cerebro	controlar	muchos de movimientos de cuerpo	**C**
61	**El cerebro es un órgano tan complejo que no se conoce al detalle su funcionamiento completo.**			**1**
1	cerebro	es	órgano complejo tan	**C**
62	**El Tálamo se halla en el centro del encéfalo, recibe las señales enviadas por los sentidos y las reenvía a distintas áreas del cerebro para su procesamiento.**			**5**
1	Tálamo	halla	en centro de encéfalo	**C**
2	Tálamo	recibe	señales enviadas	**C**
3	Tálamo	recibe	por sentidos	**I**
4	Tálamo	reenvía	a áreas distintas de cerebro	**C**
5	Tálamo	reenvía	para procesamiento	**EX**
63	**El Cerebelo es el segundo órgano más grande del encéfalo, sirve para mantener el equilibrio y controlar los movimientos finos.**			**5**
1	Cerebelo	es	órgano segundo	**C**
2	Cerebelo	es	órgano grande más	**EX**
3	Cerebelo	es	órgano de encéfalo	**EX**
4	Cerebelo	mantener	equilibrio	**C**
5	Cerebelo	controlar	movimientos finos	**C**
64	**El Hipotálamo se encarga de algunas funciones corporales, como regular la temperatura y percibir la señal de sueño, hambre y sed.**			**5**
1	Hipotálamo	encarga	de funciones corporales	**C**
2	Hipotálamo	regular	temperatura	**C**
3	Hipotálamo	percibir	señal de sueño	**C**
4	Hipotálamo	percibir	hambre	**C**
5	Hipotálamo	percibir	sed	**C**
65	**El Hipotálamo es el responsable de las manifestaciones emocionales (como la amistad, el cariño y el amor).**			**4**
1	Hipotálamo	es	responsable de manifestaciones emocionales	**C**
2	Hipotálamo	es	como amistad	**I**
3	Hipotálamo	es	cariño	**I**
4	Hipotálamo	es	amor	**I**
66	**El Bulbo raquídeo es el encargado de transmitir mensajes entre el cerebro y el cuerpo, y controla funciones básicas como el latido del corazón, la digestión y la respiración.**			**7**
1	Bulbo raquídeo	es	encargado	**C**
2	Bulbo raquídeo	transmitir	mensajes	**EX**
3	Bulbo raquídeo	transmitir	entre cerebro y cuerpo	**EX**
4	Bulbo raquídeo	controla	funciones básicas	**C**
5	Bulbo raquídeo	controla	como latido de corazón	**C**
6	Bulbo raquídeo	controla	digestión	**C**

No O	Oración			Total
No H	**Sujeto**	**Verbo**	**Complemento**	**Tipo**
7	Bulbo raquídeo	controla	respiración	**C**
67	**La Médula espinal es la prolongación del encéfalo, tiene forma de cordón y corre por dentro de la columna vertebral, que la protege.**			**2**
1	Médula espinal	es	prolongación de encéfalo	**C**
2	Médula espinal	tiene	forma de cordón	**C**
68	**De la médula espinal nacen los nervios periféricos, que permiten movimientos voluntarios e involuntarios, sensaciones y reflejos.**			**0**

Árboles de dependencias de algunas oraciones del corpus

Se muestran algunos árboles de dependencias creados por FreeLing, de algunas oraciones del corpus. FreeLing guarda estos árboles en un archivo de texto plano.

56. Para su estudio, el sistema nervioso se divide en sistema nervioso central y sistema nervioso periférico.

```
Grup-verb/top/(divide dividir VMIP3S0 -) [
  morfema-verbal/es/(se se P0000000 -)
  grup-sp/ador/(Para para SPS00 -) [
    sn/obj-prep/(estudio estudio NCMS000 -) [
      espec-ms/espec/(su su DP3CS0 -)
    ]
    Fc/term/(, , Fc -)
  ]
  sn/subj/(sistema sistema NCMS000 -) [
    espec-ms/espec/(el el DA0MS0 -)
    s-a-ms/adj-mod/(nervioso nervioso AQ0MS0 -)
  ]
  grup-sp/cc/(en en SPS00 -) [
    sn/obj-prep/(sistema sistema NCMS000 -) [
      s-a-ms/adj-mod/(nervioso nervioso AQ0MS0 -) [
        s-a-ms/modnomatch/(central central AQ0CS0 -)
      ]
    ]
  ]
  coor-n/dobj/(y y CC -) [
    sn/co-n/(sistema sistema NCMS000 -) [
      s-a-ms/adj-mod/(nervioso nervioso AQ0MS0 -) [
        s-a-ms/modnomatch/(periférico periférico AQ0MS0 -)
      ]
    ]
  ]
  F-term/term/(. . Fp -)
]
```

57. El sistema nervioso periférico lo conforman los nervios que nacen del cerebro y de la médula espinal y llegan a todas las partes del cuerpo por medio de fibras nerviosas.

```
coor-vb/top/(y y CC -) [
  grup-verb/co-v/(conforman conformar VMIP3P0 -) [
    patons/dobj/(lo lo PP3CNA00 -)
    sn/subj/(sistema sistema NCMS000 -) [
      espec-ms/espec/(El el DA0MS0 -)
      s-a-ms/adj-mod/(nervioso nervioso AQ0MS0 -)
      modnomatch/s-a-ms/(periférico periférico AQ0MS0 -)
    ]
    sn/modnomatch/(nervios nervio NCMP000 -) [
      espec-mp/espec/(los el DA0MP0 -)
      subord-rel/subord-mod/(que que PR0CN000 -) [
        grup-verb/vsubord/(nacen nacer VMIP3P0 -) [
          coor-sp/sp-obj/(y y CC -) [
            sp-de/co-sp/(de de SPS00 -) [
              sn/obj-prep/(cerebro cerebro NCMS000 -) [
                espec-ms/espec/(el el DA0MS0 -)
              ]
            ]
            sp-de/co-sp/(de de SPS00 -) [
              sn/obj-prep/(médula médula NCFS000 -) [
                espec-fs/espec/(la el DA0FS0 -)
                s-a-fs/adj-mod/(espinal espinal AQ0CS0 -)
              ]
            ]
          ]
        ]
      ]
    ]
  ]
  grup-verb/co-v/(llegan llegar VMIP3P0 -) [
    grup-sp/sp-obj/(a a SPS00 -) [
      sn/obj-prep/(partes parte NCFP000 -) [
        espec-fp/espec/(todas todo DI0FP0 -) [
          j-fp/espec/(las el DA0FP0 -)
        ]
      ]
    ]
    sp-de/sp-obj/(de de SPS00 -) [
      sn/obj-prep/(cuerpo cuerpo NCMS000 -) [
        espec-ms/espec/(el el DA0MS0 -)
      ]
    ]
    grup-sp/sp-obj/(por_medio_de por_medio_de SPS00 -) [
      sn/obj-prep/(fibras fibra NCFP000 -) [
        s-a-fp/adj-mod/(nerviosas nervioso AQ0FP0 -)
      ]
    ]
  ]
  F-term/modnomatch/(. . Fp -)
]
```

58. El encéfalo se encuentra dentro del cráneo y consta de varios órganos, cada uno de éstos realiza distintas funciones.

```
coor-vb/top/(y y CC -) [
  grup-verb/co-v/(encuentra encontrar VMIP3S0 -) [
    morfema-verbal/es/(se se P0000000 -)
    sn/subj/(encéfalo encéfalo NCMS000 -) [
      espec-ms/espec/(El el DA0MS0 -)
    ]
    grup-sp/sp-obj/(dentro_de dentro_de SPS00 -) [
      sn/obj-prep/(cráneo cráneo NCMS000 -) [
        espec-ms/espec/(el el DA0MS0 -)
      ]
    ]
  ]
  grup-verb/co-v/(consta constar VMIP3S0 -) [
    sp-de/cc/(de de SPS00 -) [
      sn/obj-prep/(órganos órgano NCMP000 -) [
        espec-mp/espec/(varios varios DI0MP0 -)
      ]
    ]
  ]
  Fc/modnomatch/(, , Fc -)
  grup-verb/co-v/(realiza realizar VMIP3S0 -) [
    sn/subj/(uno uno PI0MS000 -) [
      indef-ms/espec/(cada cada DI0CS0 -)
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(éstos este PD0MP000 -)
      ]
    ]
    sn/dobj/(funciones función NCFP000 -) [
      s-a-fp/adj-mod/(distintas distinto AQ0FP0 -)
    ]
    F-term/term/(. . Fp -)
  ]
]
```

59. El Cerebro es el órgano más grande del encéfalo, está dividido en dos mitades o hemisferios y presenta hendiduras y pliegues que le dan el aspecto de una nuez pelada.

```
coor-vb/top/(y y CC -) [
  grup-verb/co-v/(es ser VSIP3S0 -) [
    sn/subj/(Cerebro cerebro NP00000 -) [
      espec-ms/espec/(El el DA0MS0 -)
    ]
    sn/att/(órgano órgano NCMS000 -) [
      espec-ms/espec/(el el DA0MS0 -)
      s-adj/adj-mod/(grande grande AQ0CS0 -) [
        sadv/espec/(más más RG -)
      ]
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(encéfalo encéfalo NCMS000 -) [
          espec-ms/espec/(el el DA0MS0 -)
        ]
      ]
    ]
  ]
  Fc/modnomatch/(, , Fc -)
  grup-verb/co-v/(dividido dividir VMP00SM -) [
    vaux/aux/(está estar VAIP3S0 -)
    grup-sp/cc/(en en SPS00 -) [
      sn/obj-prep/(mitades mitad NCFP000 -) [
        numero-nopart/sn-mod/(dos 2 Z -)
      ]
    ]
    coor-n/dobj/(o o CC -) [
      sn/co-n/(hemisferios hemisferio NCMP000 -)
    ]
  ]
  grup-verb/co-v/(presenta presentar VMIP3S0 -) [
    coor-n/dobj/(y y CC -) [
      sn/co-n/(hendiduras hendidura NCFP000 -)
      sn/co-n/(pliegues pliegue NCMP000 -) [
        subord-rel/subord-mod/(que que PR0CN000 -) [
          grup-verb/vsubord/(dan dar VMIP3P0 -) [
            patons/iobj/(le le PP3CSD00 -)
            sn/dobj/(aspecto aspecto NCMS000 -) [
              espec-ms/espec/(el el DA0MS0 -)
              sp-de/sp-mod/(de de SPS00 -) [
                sn/obj-prep/(nuez nuez NCFS000 -) [
                  espec-fs/espec/(una uno DI0FS0 -)
                  s-a-fs/adj-mod/(pelada pelar VMP00SF -)
                ]
              ]
            ]
          ]
        ]
      ]
    ]
  ]
  F-term/modnomatch/(. . Fp -)
]
```

Anexo B. Guía para extraer hechos de forma manual

Esta guía es una segunda versión de la publicada por (Aguilar-Galicia, Sidorov, & Ledeneva, 2012). El objetivo de esta guía es ayudar a una persona a identificar y extraer los hechos de una oración: el objeto de análisis. Se empieza por describir lo qué es un texto, un párrafo, y una oración. Después se define lo qué es un hecho, se muestran algunos ejemplos y exponen una serie de pasos para extraer hechos.

Ubicación del objeto de análisis: la oración

- **Texto**. El diccionario de la Real Academia Española (RAE) define texto como "*4. m. Todo lo que se dice en el cuerpo de la obra manuscrita o impresa, a diferencia de lo que en ella va por separado; como las portadas, las notas, los índices, etc.*", donde normalmente el cuerpo de la obra se presenta por temas o subtema, tratando un tópico en particular; y su distribución es con un título seguido de párrafos.
- **Párrafo**. El diccionario de la Real Academia Española (RAE) indica que párrafo es "*1. m. Gram. Cada una de las divisiones de un escrito señaladas por letra mayúscula al principio de línea y punto y aparte al final del fragmento de escritura.*" Se puede decir que un título es un párrafo de una oración y que un párrafo está compuesto por un conjunto de oraciones
- **Oración**. (Gartz, 2011) clasifica a "la oración como aquella estructura lingüística que, en la lengua oral, se pronuncia entre dos pausas [pausa = fase de silencio]. Y en el texto escrito toma los puntos como límite de la oración". Para esta guía los signos de interrogación y exclamación también son un límite.

La característica principal de una oración es que enuncia o dice algo acerca de alguien (Mora, 2004), por ejemplo en la oración: "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*"; se habla de "*la civilización China*" y de la civilización China se dice que "*nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*".

Definición de Hecho

Como se observa según lo expuesto anteriormente, la información en un texto se conforma de párrafos, cada párrafo por un conjunto de oraciones, y siguiendo este patrón se tiene que

la oración esta compuestas por "unidades de texto más pequeñas que ella misma, que se pueden obtener a través de la descomposición de la oración en una colección de frases. Donde cada frase tiene información independiente que puede ser usada como una unidad independiente" (Hovy, Zhou, & Kwon, 2007).

Estas frases se encuentran fusionadas en la oración para enunciar algo de manera más amplia, pero al separarse de la oración tienen sentido completo, es decir, tienen información semántica por ellas mismas. Una oración tiene sentido completo si contiene sujeto y predicado (Fuentes de la Corte, 2010).

La definición de "*Hecho*" que se utiliza en esta guía es la siguiente: *Un **hecho** es la unidad mínima de texto que se puede extraer de una oración, tiene independencia semántica, únicamente un verbo y su forma es una tripleta conformada así*:

Hecho = [*Sujeto*] + [*Verbo*] + [*Objeto/Complemento*]

Ejemplo. De la oración: "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*", se pueden identificar los hechos que se muestran en la siguiente tabla.

No.	Sujeto	Verbo	Objeto/Complemento
1	La civilización China	heredó	el papel
2	La civilización China	heredó	la pólvora
3	La civilización China	heredó	una forma de imprenta rudimentaria
4	La civilización China	heredó	la brújula

Observaciones:

- Cada hecho tiene independencia semántica, es decir, ninguno necesita a otro para tener sentido completo o informar algo.
- Todos tienen un solo verbo.
- Todos cumplen la tripleta que define un hecho.
- Una oración puede contener varios hechos.

En el resto de la guía al tercer componente de la tripleta del hecho ([*objeto/Complemento*]) se le llama solamente [*Complemento*] para simplificar la escritura de la tripleta o para referirse a este componente del hecho.

Otros ejemplos

Cuando las personas hablan o escriben fusionan los hechos en una oración de forma automática, porque así funciona el idioma.

En la oración: "*Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806"*, se identifican dos hechos:

No.	Sujeto	Verbo	Complemento
1	Benito Juárez	nació	en San Pablo Guelatao, Oaxaca
2	Benito Juárez	nació	en 1806

En la oración "*Los primeros homínidos eran recolectores y sólo comían carne cuando encontraban los restos abandonados por otros animales*", se identifican cuatro hechos:

No.	Sujeto	Verbo	Complemento
1	Los primeros homínidos	eran	recolectores
2	Los primeros homínidos	comían	carne
3	Los primeros homínidos	encontraban	restos abandonados
4	Restos	abandonados	por otros animales

Se observa que los hechos son unidades de texto de las oraciones; cada uno nos enuncia algo de forma independiente; todos tienen un solo verbo, un sujeto y un complemento que juntos forman a una pequeña oración con sentido completo. En la oración "*Los primeros homínidos…*" se observa la fusión de varios hechos en la oración, para enunciar algo de una manera más amplia.

Entonces se tiene que identificar hechos, consiste en desmenuzar una oración extrayendo los tres componentes de un hecho; y después unirlos para crear el hecho.

Extraer hechos

Para la extracción de hechos cada oración se procesa de forma independiente. A continuación se describen una serie de pasos a seguir para extraer los hechos de la oración.

1. **Localizar el sujeto en la oración**. Para hacerlo se contesta a una de las siguientes preguntas:
    a. ¿De qué o de quién se habla?
    b. ¿Quién o qué realiza la acción?
2. **Localizar el predicado en la oración**. Para hacerlo se contesta a una de las siguientes preguntas:
    a. ¿Qué se dice, de quien se habla o de lo que se habla?
    b. ¿Qué se dice, de quien o lo que realiza la acción?
3. **Identificar el verbo principal en la oración**. Se encuentra al inicio del predicado, y puede presentarse cómo un solo verbo o en forma de perífrasis verbal.
4. **Complemento**. El complemento para el primer hecho se obtiene restando del predicado obtenido en el paso 2, el verbo obtenido en el paso 3.
5. **Construir Hecho**. [sujeto obtenido en paso 1] + [verbo obtenido en paso 3] + [complemento obtenido en paso 4].
6. **Más hechos**. Revisar el complemento del paso 4 buscando si tiene alguna de las siguientes características. De acuerdo a ellas se pueden extraer más hechos. Esto reducirá el complemento del primer hecho.
    a. Conjunción Copulativa.
    b. Conjunción Disyuntiva.
    c. Pronombre Relativo.
    d. Preposición: *desde* y *hasta*.
    e. Preposición: *en*.
7. Fin

Ejemplo, Oración: "*La numeración arábiga procede de India.*"

1. Se habla de *la numeración arábiga*.
    a. Sujeto = {La numeración arábiga}
2. De la numeración arábiga se dice que *procede de India*.
    a. Predicado = {procede de India}
3. Verbo principal = {procede}
4. Complemento = {de India}.
5. Hecho = [La numeración arábiga] + [procede] + [de India].
6. No aplica.
7. Fin.

Conjunción copulativa

Conjunciones copulativas son las "que coordinan dos o más palabras las cuales desempeñan una misma función. También pueden unir oraciones. Las conjunciones copulativas son *y*, *e*, *ni*" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Ejemplo:

- El domingo compré discos de música hindú, turca ***y*** rusa.

Si la oración contiene conjunción copulativa, se obtiene un hecho por cada término coordinado por la conjunción. Para el ejemplo anterior los hechos son:

- El domingo compré discos de música hindú
- El domingo compré discos de música turca
- El domingo compré discos de música rusa

El hecho se forma así:

- [Sujeto] + [Verbo] + [Complemento_1]
- [Sujeto] + [Verbo] + [Complemento_2]
- [Sujeto] + [Verbo] + [Complemento_n]

El Sujeto y Verbo son el mismo para cada hecho, sólo cambia el complemento.

Conjunción disyuntiva

Las conjunciones disyuntivas "son conjunciones que enlazan palabras u oraciones para expresar posibilidades alternativas, distintas o contradictorias. Las conjunciones disyuntivas son *o, u*" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Ejemplo:

- Pedro se hospedará en una pensión ***u*** hotel cualquiera.

Cuando la oración tiene conjunción disyuntiva, se obtiene un hecho por cada término coordinado por la conjunción. Para el ejemplo anterior los hechos son:

- Pedro se hospedará en una pensión
- Pedro se hospedará en un hotel cualquiera

El hecho se forma así:

- [Sujeto] + [Verbo] + [Complemento_1]
- [Sujeto] + [Verbo] + [Complemento_2]
- [Sujeto] + [Verbo] + [Complemento_n]

El Sujeto y Verbo son el mismo para cada hecho, sólo cambia el complemento.

Pronombre relativo

"Los pronombres relativos hacen referencia a alguien o a algo que se ha mencionado antes en el discurso o que ya es conocido por los interlocutores. Los pronombres relativos, funcionan, en la mayor parte de los casos, como elementos de subordinación de oraciones. Los pronombres relativos son: *que, quien, quienes, cual, cuales, cuanto, cuantos, cuanta, cuantas*" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Ejemplo:

- Pedro conoció a un estudiante ***que*** sabe hablar chino.

Como el pronombre relativo hace referencia a alguien o a algo que se ha mencionado antes, entonces se busca al sujeto (sustantivo, pronombre personal) para un siguiente hecho en la parte inmediata que antecede al pronombre relativo. El verbo de este hecho se encuentra localizado después del pronombre relativo. Los hechos del ejemplo, son:

- Pedro conoció a un estudiante – (Primer hecho)
- Estudiante sabe hablar Chino – (Hecho obtenido por el pronombre relativo)

La estructura del hecho es:

[Sujeto nuevo] + [Verbo nuevo] + [Complemento ubicado después del pronombre relativo]

Preposición: *desde* y *hasta*

"*Desde*. Denota inicio de una acción en el tiempo o en el espacio. *Hasta*. Expresa el fin de algo o límite de lugar, de número o de tiempo" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Ejemplo:

- La primavera comprende desde el mes de marzo hasta el mes de junio.

Si una oración contiene estas preposiciones, se formará un hecho por cada una de ellas. Para el ejemplo se tienen los siguientes hechos:

- La primavera comprende desde el mes de marzo
- La primavera comprende hasta el mes de junio

Se forman dos hechos con la siguiente estructura:

- [Sujeto] + [Verbo] + [Preposición *desde*] + [Complemento]
- [Sujeto] + [Verbo] + [Preposición *hasta*] + [Complemento]

Preposición: *en*

La preposición *en* "indica tiempo, expresa lugar, señala modo, significa ocupación o actividad, indica medio o instrumento, forma locuciones adverbiales" (Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000). Ejemplo:

- Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806.

Cuando una oración contiene una o más de una preposición *en*, se forma un hecho por cada una de ellas. Los hechos del ejemplo son:

- Benito Juárez nació ***en*** San Pablo Guelatao, Oaxaca
- Benito Juárez nació ***en*** 1806

Los hechos se forman con la siguiente estructura:

[Sujeto] + [Verbo] + [Preposición *en*] + [Complemento]

Perífrasis verbal

(Munguía Zatarain, Munguía Zatarain, & Rocha Romero, 2000) Dice que "Las perífrasis verbales son construcciones que se forman con dos o más verbos que, en ocasiones, pueden estar unidos por una palabra de enlace. El primer verbo se conjuga y el segundo se expresa

por medio de una forma no personal, es decir, por un infinitivo, un gerundio o un participio, aunque también es posible encontrarlo conjugado".

Las perífrasis normalmente tienen la siguiente forma: (verbo auxiliar) + (preposición o conjunción) + (infinitivo, gerundio o participio).

En esta guía cuando se encuentra perífrasis verbal en una oración, se toma la última parte de la forma, es decir, se toma el: "infinitivo, gerundio o participio".

Anexo C. Etiquetas sintácticas empleadas por FreeLing

A continuación se listan en tablas las "Etiquetas sintácticas de dependencias" y las "Etiquetas sintácticas superficiales" para español que utiliza (FreeLing). La descripción de estas etiquetas es parte de la documentación de FreeLing y se encuentran en los archivos "ca+esLABELINGtags" y "esCHUNKtags" respectivamente, localizados en la carpeta "doc\grammars\", creada al instalar FreeLing.

Etiquetas sintácticas de dependencias

Etiquetas, para marcar relaciones sintácticas de dependencias, empleadas por FreeLing para el idioma español. Esta etiqueta se muestra en los nodos de los árboles de dependencias así: {func: <Etiqueta func>}.

Tabla C.1 Etiquetas sintácticas de dependencias para español, empleadas por FreeLing.

No	Etiqueta func	Descripción
1	adj-mod	adjectival modifier (adjectives modifying a noun)
2	ador	sentence adjunct (sentence elaboration introduced by discourse markers such as "but", "thanks to", "eventhough", "for this reason", etc.)
3	agent	agent in passive sentences
4	att	attribute of predicate whose head is a copulative verb ("John is tall")
5	aux	auxiliary verbs
6	cc	adjunct (tipically specifying time, place, manner, etc. of the verb)
7	co-adj	coordinated adjective
8	co-adv	coordinated adverb
9	co-ger	coordinated gerund
10	co-inf	coordinated infinitive
11	co-n	coordinated noun
12	co-part	coordinated participle
13	co-sp	coordinated prepositional phrase
14	co-subord	coordinated clause
15	co-v	coordinated verb or sentence
16	dconj	subordinating conjunction of a verb - periphrasis ("I think that he will come")
17	dep-adv	adverbs - verbless sentences
18	dep-ger	gerund clauses - verbless sentences
19	dep-inf	infinitive clauses - verbless sentences
20	dep-noun	nouns - verbless sentences
21	dep-part	participle clauses - verbless sentences
22	dep-prep	prepostitional phrases - verbless sentences
23	dep-subord	finite clauses - verbless sentences
24	dep	clitic pronouns
25	dobj	direct object
26	dprep	verb + preposition - periphrasis
27	dverb	verb + verb - periphrasis

No	Etiqueta func	Descripción
28	es	"se" passive, impersonal, pronominal morpheme, reflexive pronouns
29	espec	nominal and verbal determiners
30	iobj	indirect object
31	obj-prep	prepositional object in a PP ("in the house")
32	pred	attribute of (non-copulative) predicative verbs ("they remained silent")
33	prepos	objects whose head is prepositional
34	sn-mod	nominal modifier (noun phrase modifying another)
35	sp-mod	prepositional modifier (prepositional phrases modifying a noun or adjectival phrase)
36	sp-obj	prepositional object ("I believe in ghosts", "He pointed towards the crowd")
37	subj-pac	patient - passive
38	subj	subject
39	subord-mod	relative clauses modifying a nominal head.
40	term	punctuation
41	top	sentence head (highest head)
42	vsubord	verb subordinated to an interrogative pronoun or to subordinating conjunction ("I don't know who washed the dishes", "I think that he called her")

Etiquetas sintácticas superficiales

Etiquetas sintácticas superficiales, empleadas por FreeLing para el idioma español. Esta etiqueta se muestra en los nodos de los árboles de dependencias así: {synt: <Etiqueta synt>}.

Tabla C.2 Etiquetas sintácticas superficiales para español, empleadas por FreeLing.

No	Etiqueta synt	Descripción
1	a-fp	adjective, adjective feminine plural
2	a-fs	adjective, adjective feminine singular
3	a-mp	adjective, adjective masculine plural
4	a-ms	adjective, adjective masculine singular
5	s-a-fp	adjective, adjective phrase feminine plural
6	s-a-fs	adjective, adjective phrase feminine singular
7	s-a-mp	adjective, adjective phrase masculine plural
8	s-a-ms	adjective, adjective phrase masculine singular
9	s-adj	adjective, adjective phrase
10	sadv	adverb, adverb phrase
11	cuantif	adverb, adverb which expresses quantification
12	adv-interrog	adverb, interrogative adverb
13	neg	adverb, negation
14	coord	conjunction, coordinating conjunction
15	conj-subord	conjunction, subordinating conjunction
16	data	date
17	grup-complex-spec-fp	determiner, complex determiner feminine plural
18	grup-complex-	determiner, complex determiner feminine singular

No	Etiqueta synt	Descripción
	spec-fs	
19	grup-complex-spec-mp	determiner, complex determiner masculine plural
20	grup-complex-spec-ms	determiner, complex determiner masculine singular
21	j-fp	determiner, definite determiner feminine plural
22	j-fs	determiner, definite determiner feminine singluar
23	j-mp	determiner, definite determiner masculine plural
24	j-ms	determiner, definite determiner masculine singular
25	espec-ms-E	determiner, definite / indefinite / demonstrative determiners masculin singular whose head is a noun feminine singular
26	dem-fp	determiner, demonstrative determiner feminine plural
27	dem-fs	determiner, demonstrative determiner feminine singular
28	dem-mp	determiner, demonstrative determiner masculine plural
29	dem-ms	determiner, demonstrative determiner masculine singular
30	espec-fp	determiner, determiner feminine plural
31	espec-fs	determiner, determiner feminine singular
32	espec-mp	determiner, determiner masculine plural
33	espec-ms	determiner, determiner masculine singular
34	exc-fp	determiner, exclamative determiner feminine plural
35	exc-fs	determiner, exclamative determiner feminine singular
36	exc-mp	determiner, exclamative determiner masculine plural
37	exc-ms	determiner, exclamative determiner masculine singular
38	indef-fp	determiner, indefinite determiner feminine plural
39	indef-fs	determiner, indefinite determiner feminine singular
40	indef-ms	determiner, indefinite determiner feminine singular
41	indef-mp	determiner, indefinite determiner masculine plural
42	int-fp	determiner, interrogative determiner feminine plural
43	int-fs	determiner, interrogative determiner feminine singular
44	int-mp	determiner, interrogative determiner masculine plural
45	int-ms	determiner, interrogative determiner masculine singular
46	quant-fp	determiner, numeral determiner feminine plural
47	quant-fs	determiner, numeral determiner feminine singular
48	quant-mp	determiner, numeral determiner masculine plural
49	quant-ms	determiner, numeral determiner masculine singular
50	pos-fp	determiner, possessive determiner femenine plural
51	pos-fs	determiner, possessive determiner femenine singular
52	pos-mp	determiner, possessive determiner masculin plural
53	pos-ms	determiner, possessive determiner masculin singular
54	interjeccio	interjection
55	grup-nom-fp	noun, nominal chunk feminine plural
56	grup-nom-fs	noun, nominal chunk feminine singular
57	grup-nom-mp	noun, nominal chunk masculine plural
58	grup-nom-ms	noun, nominal chunk masculine singular
59	grup-nom	noun, nominal chunk neuter
60	n-fp	noun, noun feminine plural
61	nom-tmp-fp	noun, noun feminine plural - time expression
62	nom-tmp-fs	noun, noun feminine singluar - time expression
63	n-fs	noun, noun feminine singular
64	nom-fs-E	noun, noun feminine singular that goes with masculine singular determiners (ej. el agua)
65	n-mp	noun, noun masculine plural
66	nom-tmp-mp	noun, noun masculine plural - time expression
67	n-ms	noun, noun masculine singular
68	nom-tmp-ms	noun, noun masculine singular - time expression

No	Etiqueta synt	Descripción
69	sn	noun, noun phrase
70	sn-tmp	noun, noun phrase - time expression
71	w-fp	noun, proper noun feminine plural
72	w-fs	noun, proper noun feminine singular
73	w-mp	noun, proper noun masculine plural
74	w-ms	noun, proper noun masculine singular
75	num-fp	number, number feminine plural
76	num-fs	number, number feminine singluar
77	num-mp	number, number masculine plural
78	num-ms	number, number masculine singular
79	numero-nopart	number, number nonpartitive (ej. un centenar de personas)
80	numero-part	number, number partitive (ej. un centenar)
81	numero	number
82	morf-pron	particle, pronoun morpheme 'es'
83	morfema-verbal	particle, verbal morpheme 'es'
84	prep	preposition
85	grup-sp	preposition, prepositional chunk
86	sp-de	preposition, prepositional phrase 'de'
87	grup-sp-inf	preposition, prepositional phrase whose daughter is an infinitive clause
88	prepc-ms	preposition, preposition + definite determiner masculine singular
89	prel-adv	pronoun, adverbial relative pronoun
90	paton-fp	pronoun, clitic 3rd person feminine plural
91	paton-fs	pronoun, clitic 3rd person feminine singular
92	paton-mp	pronoun, clitic 3rd person masculine plural
93	paton-ms	pronoun, clitic 3rd person masculine singular
94	paton-p	pronoun, clitic 3rd person neuter plural
95	paton-s	pronoun, clitic 3rd person neuter singluar
96	patons	pronoun, clitic 3rd person
97	relatiu	pronoun, complex relative pronoun
98	pdem-fp	pronoun, demonstrative pronoun feminine plural
99	pdem-fs	pronoun, demonstrative pronoun feminine singular
100	pdem-mp	pronoun, demonstrative pronoun masculine plural
101	pdem-ms	pronoun, demonstrative pronoun masculine singular
102	pindef-fp	pronoun, indefinite pronoun feminine plural
103	pindef-fs	pronoun, indefinite pronoun feminine singular
104	pindef-mp	pronoun, indefinite pronoun masculine plural
105	pindef-ms	pronoun, indefinite pronoun masculine singular
106	pinterrog-fp	pronoun, interrogative pronoun feminine plural
107	pinterrog-fs	pronoun, interrogative pronoun feminine singular
108	pinterrog-mp	pronoun, interrogative pronoun masculine plural
109	pinterrog-ms	pronoun, interrogative pronoun masculine singular
110	pinterrog-s	pronoun, interrogative pronoun neuter singular
111	pinterrog	pronoun, interrogative pronoun 'quÃ©'
112	pinterrog-p	pronoun, interrogative pronoun 'quiÃ©nes'
113	pnum-fp	pronoun, numeral pronoun feminine plural
114	pnum-fs	pronoun, numeral pronoun feminine singular
115	pnum-mp	pronoun, numeral pronoun masculine plural
116	pnum-ms	pronoun, numeral pronoun masculine singular
117	psubj-s	pronoun, personal pronoun 1st / 2nd person singular
118	psubj-fp	pronoun, personal pronoun 3rd person feminine plural
119	psubj-fs	pronoun, personal pronoun 3rd person feminine singular
120	psubj-mp	pronoun, personal pronoun 3rd person masculine plural
121	psubj-ms	pronoun, personal pronoun 3rd person masculine singular
122	paton	pronoun, personal pronoun 3rd person - non direct object ('lo': la casa es grande --> la casa lo es)

No	Etiqueta synt	Descripción
123	psubj	pronoun, personal pronoun
124	ptonic	pronoun, personal pronouns 'mi' & 'si'
125	pposs-fp	pronoun, possessive pronoun femenine plural
126	pposs-fs	pronoun, possessive pronoun femenine singular
127	pposs-mp	pronoun, possessive pronoun masculine plural
128	pposs-ms	pronoun, possessive pronoun masculine singular
129	pposs-ns	pronoun, possessive pronoun neuter singular
130	cuyo-fp	pronoun, possessive relative pronoun feminine plural
131	cuyo-fs	pronoun, possessive relative pronoun feminine singular
132	cuyo-mp	pronoun, possessive relative pronoun masculine plural
133	cuyo-ms	pronoun, possessive relative pronoun masculine singular
134	pron	pronoun
135	pron-fp	pronoun, pronoun feminine plural
136	pron-fs	pronoun, pronoun feminine singular
137	pron-mp	pronoun, pronoun masculine plural
138	pron-ms	pronoun, pronoun masculine singular
139	pron-ns	pronoun, pronoun neuter singular
140	quien-p	pronoun, relative pronoun plural 'quien'
141	cual-p	pronoun, relative pronoun 'qual' plural
142	cual-s	pronoun, relative pronoun 'qual' singular
143	prel	pronoun, relative pronoun 'que'
144	quien-s	pronoun, relative pronoun singular 'quien'
145	F-no-c	punctuation
146	F-term	punctuation, sentence terminators
147	vaux	verb, auxiliary verb
148	vaux	verb, auxiliary verb
149	geraux	verb, gerund as auxiliary verb
150	forma-ger	verb, gerund
151	ger	verb, gerund
152	gerundi	verb, gerund - periphrasis
153	geraux-ser	verb, gerund 'ser' as auxiliary verb
154	infaux	verb, infinitive as auxiliary verb
155	forma-inf	verb, infinitive
156	infinitiu	verb, infinitive - prepositional head & periphrasis
157	infaux-ser	verb, infinitive 'ser' as auxiliary verb
158	inf	verb, infinitive - verbal head
159	grup-verb-inf	verb, infinitive whose head is prepositional
160	parti-aux	verb, participle as auxiliary verb
161	parti-fp	verb, participle feminine plural
162	parti-fs	verb, participle feminine singular
163	parti-mp	verb, participle masculine plural
164	parti-ms	verb, participle masculine singular
165	parti	verb, participle
166	parti-ser	verb, participle 'ser' as auxiliary verb
167	parti-flex	verb, participle with gender and number
168	ger-pas	verb, passive gerund
169	inf-pas	verb, passive infinitive
170	vser	verb, passive verb 'ser'
171	verb-pass	verb, passive verb
172	verb	verb
173	grup-verb	verb, verbal chunk
174	v-hacer-3p	verb, verb 'fer' singular - time expression

Anexo D. Etiquetas morfológicas empleadas por FreeLing

Etiquetas morfológicas empleadas por FreeLing para el idioma español. Esta etiqueta se muestra en los nodos de los árboles de dependencias así: {tag: <Valor de tag>}.

La información que se presenta aquí fue tomada textualmente del sitio web de (FreeLing), de la página "Introducción a las etiquetas Eagles v. 2.0" (EagV2.0). Se muestra aquí completamente, con la intención de que la tesis cuente con la información necesaria para su entendimiento.

El analizador morfológico para el castellano utiliza un conjunto de etiquetas para representar la información morfológica de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo "Expert Advisory Group on Language Engineering Standards" (EAGLES) para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Así pues está previsto que recojan los accidentes gramaticales existentes en las lenguas europeas. Es por esto que dependiendo de la lengua hay atributos que pueden no especificarse. Si un atributo no se especifica significa que o bien expresa un tipo de información que no existe en la lengua o que la información no se considera relevante. La infra especificación de un atributo se marca con el 0.

A continuación presentamos las etiquetas que el analizador morfológico utiliza para el castellano en formato de tabla y algunos ejemplos de cada categoría.

Para cada categoría se presentan los atributos, valores y códigos que puede tomar:

ETIQUETAS			
Posición	**Atributo**	**Valor**	**Código**
Columna 1	*Columna 2*	*Columna 3*	*Columna 4*

En la *columna 1* encontramos un número que hace referencia al orden y posición en que aparecen los atributos. La *columna 2* hace referencia a los atributos, el número de los cuales varía dependiendo de la categoría. En la *columna 3* encontramos los valores que puede tomar cada atributo y, finalmente, la *columna 4* representa los códigos que se han

establecido para su representación. Las etiquetas en sí sólo son los códigos (columna 4) y se sabe a qué atributo pertenecen por la posición (columna 1) en la que se encuentran.

ETIQUETAS POR CATEGORÍA:

1. ADJETIVOS
2. ADVERBIOS
3. DETERMINANTES
4. NOMBRES
5. VERBOS
6. PRONOMBRES
7. CONJUNCIONES
8. INTERJECCIONES
9. PREPOSICIONES
10. SIGNOS DE PUNTUACIÓN
11. NUMERALES
12. FECHAS y HORAS

1. ADJETIVOS

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
		-	0
3	Grado	-	0
		Aumentativo	A
		Diminutivo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participio	P

1.1. Adjetivos calificativos

- El lema de los adjetivos calificativos será siempre la forma masculina singular (bonito) o la forma singular si el adjetivo es de género común (alegre). Para los adjetivos invariables, es decir, aquellos que tanto para el singular como para el plural presentan la misma forma, el lema y la forma han de coincidir.
- El valor del último dígito será normalmente 0. Tan sólo aquellos adjetivos que tengan función de participio tendrán una P. En el caso que la forma del adjetivo coincida con la del participio, solo tomará la etiqueta de participio.
- El atributo grado se especificará para aquellos adjetivos que tengan grado (comparativos, superlativos) o sufijación apreciativa (aumentativos, despectivos, etc.). Se distinguirán porque el tercer dígito de la etiqueta será A (aumentativo) o D (diminutivo), C (comparativo) o S (superlativo) mientras que para el resto de adjetivos este valor será 0.

Ejemplos:

Forma	Lema	Etiqueta
alegres	alegre	AQ0CP0
alegre	alegre	AQ0CS0
bonita	bonito	AQ0FS0
grandazo	grande	AQAMS0
pésimo	malo	AQSMP0
pequeñitas	pequeño	AQDFP0
antiarrugas	antiarrugas	AQ0CN0
desnuda	desnudo	AQ0FSP

1.2. Adjetivos Ordinales

- Los adjetivos de tipo ordinal tendrán como lema el masculino singular siendo también la forma masculina singular plena (*primero*, *tercero*) el lema de las formas apocopadas (*primer*, *tercer*).

Ejemplos:

Forma	Lema	Etiqueta
primer	primero	AO0MS0
primera	primero	AO0FS0
primeras	primero	AO0FP0
primero	primero	AO0MS0
primeros	primero	AO0MP0

2. ADVERBIOS

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N

- Para los adverbios, de momento, tan sólo indicamos que es de tipo general o de tipo negativo.
- La etiqueta de adverbio negativo (RN) está reservada exclusivamente para el adverbio no.

- Esta etiqueta sirve tanto para los adverbios y para las locuciones adverbiales.
- El lema de los adverbios acabados en -mente es la misma forma adverbial acabada en -mente, es decir, el lema de rápidamente será rápidamente.

Ejemplos:

Forma	Lema	Etiqueta
despacio	despacio	RG
ahora	ahora	RG
siempre	siempre	RG
hábilmente	hábilmente	RG
posteriormente	posteriormente	RG
a_cuatro_patas	a_cuatro_patas	RG
a_granel	a_granel	RG
no	no	RN

3. DETERMINANTES

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Poseedor	Singular	S
		Plural	P

El atributo Persona tendrá por defecto el valor 0, con excepción de los determinantes posesivos que podrán tomar el valor 1, 2 y 3.

El atributo Poseedor sólo se especificará para los determinantes posesivos. Si el referente es singular, el valor será S, si es plural el valor será P. Cuando el referente sea de tercera persona el valor de este atributo será 0 ya que es imposible distinguir los referentes de él-de ellos.

Ejemplos:

3.1. Determinantes Demostrativos

Forma	Lema	Etiqueta
aquel	aquel	DD0MS0
aquella	aquel	DD0FS0
aquellas	aquel	DD0FP0
aquellos	aquel	DD0MP0
esa	ese	DD0FS0
esas	ese	DD0FP0
ese	ese	DD0MS0
esos	ese	DD0MP0
esta	este	DD0FS0
estas	este	DD0FP0
este	este	DD0MS0
estos	este	DD0MP0
tal	tal	DD0CS0
tales	tal	DD0CP0
semejante	semejante	DD0CS0
semejantes	semejante	DD0CP0

3.2. Determinantes Posesivos

Forma	Lema	Etiqueta
mi	mi	DP1CSS
mis	mi	DP1CPS
tu	tu	DP2CSS
tus	tu	DP2CPS
su	su	DP3CS0
sus	su	DP3CP0

Forma	Lema	Etiqueta
nuestra	nuestro	DP1FSP
nuestras	nuestro	DP1FPP
nuestro	nuestro	DP1MSP
nuestros	nuestro	DP1MPP
vuestra	vuestro	DP2FSP
vuestras	vuestro	DP2FPP
vuestro	vuestro	DP2MSP
vuestros	vuestro	DP2MPP

3.3. Determinantes Interrogativos

Forma	Lema	Etiqueta
cuánta	cuánto	DT0FS0
cuántas	cuánto	DT0FP0
cuánto	cuánto	DT0MS0
cuántos	cuánto	DT0MP0
qué	qué	DT0CN0

3.4. Determinantes Exclamativos

Forma	Lema	Etiqueta
qué	qué	DE0CN0

3.5. Determinantes Indefinidos

Forma	Lema	Etiqueta
alguna	alguno	DI0FS0
algunas	alguno	DI0FP0
alguno	alguno	DI0MS0
algún	alguno	DI0MS0
algunos	alguno	DI0MP0
bastante	bastante	DI0CS0

Forma	Lema	Etiqueta
bastantes	bastante	DI0CP0
cada	cada	DI0CS0
ninguna	ninguno	DI0FS0
ningunas	ninguno	DI0FP0
ninguno	ninguno	DI0MS0
ningún	ninguno	DI0MS0
ningunos	ninguno	DI0MP0
otra	otro	DI0FS0
otras	otro	DI0FP0
otro	otro	DI0MS0
otros	otro	DI0MP0
sendas	sendos	DI0FP0
sendos	sendos	DI0MP0
tantas	tanto	DI0FP0
tanta	tanto	DI0FS0
tantos	tanto	DI0MP0
tanto	tanto	DI0MS0
todas	todo	DI0FP0
toda	todo	DI0FS0
todos	todo	DI0MP0
todo	todo	DI0MS0
unas	uno	DI0FP0
una	uno	DI0FS0
unos	uno	DI0MP0
un	uno	DI0MS0
varias	varios	DI0FP0
varios	varios	DI0MP0

3.6. Artículos

- Sólo tratamos como artículo las formas del artículo definido. No contemplamos la categoría de artículo indefinido (un) puesto que hemos decidido tratarlas como determinantes indefinidos (véase 3.5) o numerales (véase 3.7).

Ejemplos:

Forma	Lema	Etiqueta
el	el	DA0MS0
los	el	DA0MP0
lo	el	DA0NS0
la	el	DA0FS0
las	el	DA0FP0

4. NOMBRES

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

- Los nombres tienen como lema la forma singular, tanto si es de género femenino como masculino o neutro. Para los nombres invariables, es decir, aquellos que tanto

para el singular como para el plural presentan la misma forma (tesis), el lema y la forma coincidirán.

- Los nombres propios tendrán la etiqueta NP00000 si no se usa clasificación semántica, o bien, el valor correspondiente en los dígitos 5-6.
- El atributo grado se especificará para aquellos nombres que tengan sufijación apreciativa (aumentativos, despectivos, etc.). Para el resto de nombres este valor será 0.

Ejemplos:

Forma	Lema	Etiqueta
chico	chico	NCMS000
chicas	chico	NCFP000
gatito	gato	NCMS00D
oyente	oyente	NCCS000
oyentes	oyente	NCCP000
cortapapeles	cortapapeles	NCMN000
tesis	tesis	NCFN000
Barcelona	barcelona	NP000G0
COI	coi	NP000O0
Pedro	pedro	NP000P0

5. VERBOS

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

- El lema del verbo ha de ser siempre el infinitivo.
- Etiquetamos las formas del verbo haber como auxiliares (VA) cuando actúan como tal, y como verbo principal (VM) en los existenciales (hay dinero, cuando haya dinero), las del verbo ser como semiauxiliares (VS) y las restantes como principales (VM).

- El atributo de Género sólo afecta a los participios, para el resto de formas este atributo no se especifica (0).
- Para las formas de infinitivo y gerundio no se especifican los atributos de Tiempo, Persona, Número y Género, por lo que su valor será 0.

Forma	Lema	Etiqueta
cantada	cantar	VMP00S***F***
cantadas	cantar	VMP00P***F***
cantado	cantar	VMP00S***M***
cantados	cantar	VMP00P***M***

Ejemplos:

Tiempo	VERBOS PRINCIPALES				VERBO SEMIAUXILIAR		
	Forma	Lema	Etiqueta		Forma	Lema	Etiqueta
PRESENTE DE INDICATIVO	canto	cantar	VMIP1S0		soy	ser	VSIP1S0
	cantas	cantar	VMIP2S0		eres	ser	VSIP2S0
	canta	cantar	VMIP3S0		es	ser	VSIP3S0
	cantamos	cantar	VMIP1P0		somos	ser	VSIP1P0
	cantáis	cantar	VMIP2P0		sois	ser	VSIP2P0
	cantan	cantar	VMIP3P0		son	ser	VSIP3P0
PRETÉRITO IMPERFECTO	cantaba	cantar	VMII1S0		era	ser	VSII1S0
	cantabas	cantar	VMII2S0		eras	ser	VSII2S0
	cantaba	cantar	VMII3S0		era	ser	VSII3S0
	cantábamos	cantar	VMII1P0		éramos	ser	VSII1P0
	cantabais	cantar	VMII2P0		erais	ser	VSII2P0
	cantaban	cantar	VMII3P0		eran	ser	VSII3P0
PRETÉRITO PERFECTO SIMPLE	canté	cantar	VMIS1S0		fui	ser	VSIS1S0
	cantaste	cantar	VMIS2S0		fuiste	ser	VSIS2S0
	cantó	cantar	VMIS3S0		fue	ser	VSIS3S0
	cantamos	cantar	VMIS1P0		fuimos	ser	VSIS1P0
	cantasteis	cantar	VMIS2P0		fuisteis	ser	VSIS2P0
	cantaron	cantar	VMIS3P0		fueron	ser	VSIS3P0
FUTURO DE INDICATIVO	cantaré	cantar	VMIF1S0		seré	ser	VSIF1S0
	cantarás	cantar	VMIF2S0		serás	ser	VSIF2S0
	cantará	cantar	VMIF3S0		será	ser	VSIF3S0
	cantaremos	cantar	VMIF1P0		seremos	ser	VSIF1P0
	cantaréis	cantar	VMIF2P0		seréis	ser	VSIF2P0

Tiempo	VERBOS PRINCIPALES			VERBO SEMIAUXILIAR		
	Forma	**Lema**	**Etiqueta**	**Forma**	**Lema**	**Etiqueta**
	cantarán	cantar	VMIF3P0	serán	ser	VSIF3P0
CONDICIONAL	cantaría	cantar	VMIC1S0	sería	ser	VSIC1S0
	cantarías	cantar	VMIC2S0	serías	ser	VSIC2S0
	cantaría	cantar	VMIC3S0	sería	ser	VSIC3S0
	cantaríamos	cantar	VMIC1P0	seríamos	ser	VSIC1P0
	cantaríais	cantar	VMIC2P0	seríais	ser	VSIC2P0
	cantarían	cantar	VMIC3P0	serían	ser	VSIC3P0
PRESENTE DE SUBJUNTIVO	cante	cantar	VMSP1S0	sea	ser	VSSP1S0
	cantes	cantar	VMSP2S0	seas	ser	VSSP2S0
	cante	cantar	VMSP3S0	sea	ser	VSSP3S0
	cantemos	cantar	VMSP1P0	seamos	ser	VSSP1P0
	cantéis	cantar	VMSP2P0	seáis	ser	VSSP2P0
	canten	cantar	VMSP3P0	sean	ser	VSSP3P0
PRETÉRITO IMPERFECTO	cantara	cantar	VMSI1S0	fuera	ser	VSSI1S0
	cantaras	cantar	VMSI2S0	fueras	ser	VSSI2S0
	cantara	cantar	VMSI3S0	fuera	ser	VSSI3S0
	cantáramos	cantar	VMSI1P0	fuéramos	ser	VSSI1P0
	cantarais	cantar	VMSI2P0	fuerais	ser	VSSI2P0
	cantaran	cantar	VMSI3P0	fueran	ser	VSSI3P0
	cantase	cantar	VMSI1S0	fuese	ser	VSSI1S0
	cantases	cantar	VMSI2S0	fueses	ser	VSSI2S0
	cantase	cantar	VMSI3S0	fuese	ser	VSSI3S0
	cantásemos	cantar	VMSI1P0	fuésemos	ser	VSSI1P0
	cantaseis	cantar	VMSI2P0	fueseis	ser	VSSI2P0
	cantasen	cantar	VMSI3P0	fuesen	ser	VSSI3P0
FUTURO DE SUBJUNTIVO	cantare	cantar	VMSF1S0	fuere	ser	VSSF1S0
	cantares	cantar	VMSF2S0	fueres	ser	VSSF2S0
	cantare	cantar	VMSF3S0	fuere	ser	VSSF3S0
	cantáremos	cantar	VMSF1P0	fuéremos	ser	VSSF1P0
	cantareis	cantar	VMSF2P0	fuereis	ser	VSSF2P0
	cantaren	cantar	VMSF3P0	fueren	ser	VSSF3P0
GERUNDIO	cantando	cantar	VMG0000	siendo	ser	VSG0000
IMPERATIVO	canta	cantar	VMM02S0	sé	ser	VSM02S0
	cante	cantar	VMM03S0	sea	ser	VSM03S0
	cantemos	cantar	VMM01P0	seamos	ser	VSM01P0
	cantad	cantar	VMM02P0	sed	ser	VSM02P0
	canten	cantar	VMM03P0	sean	ser	VSM03P0
INFINITIVO	cantar	cantar	VMN0000	ser	ser	VSN0000
PARTICIPIO	cantada	cantar	VMP00SF	sido	ser	VSP00SM

Tiempo	VERBOS PRINCIPALES				VERBO SEMIAUXILIAR		
	Forma	Lema	Etiqueta		Forma	Lema	Etiqueta
	cantado	cantar	VMP00SM				
	cantadas	cantar	VMP00PF				
	cantados	cantar	VMP00PM				

6. PRONOMBRES

PRONOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		ImpersonalMInvariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

- El atributo Persona se especificará para los pronombres personales y posesivos, para el resto de formas el valor será 0.
- El atributo Caso es específico para los pronombres personales, para el resto será 0.

- El atributo Poseedor sólo se usará con los pronombres posesivos para marcar el número del poseedor: singular para el mío, el tuyo, plural para el nuestro y el vuestro. Para los pronombres en que el poseedor es una tercera persona (el suyo), como no se podrá distinguir si es de singular o plural (si se refiere a él o a ellos), este atributo tomará valor 0.
- El atributo Politeness (cortesía) se especificará para los pronombres personales usted, ustedes y vos.
- El lema será la forma masculina del pronombre con las mismas características de caso y persona.

6.1. Pronombres Personales

Forma	Lema	Etiqueta
yo	yo	PP1CSN00
me	me	PP1CS000
mí	mí	PP1CSO00
nos	me	PP1CP000
nosotras	nosotros	PP1FP000
nosotros	nosotros	PP1MP000
conmigo	conmigo	PP1CSO00
te	te	PP2CS000
ti	tí	PP2CSO00
tú	tú	PP2CSN00
os	te	PP2CP000
usted	usted	PP2CS00P
ustedes	usted	PP2CP00P
vos	tú	PP3CN00P
vosotras	vosotros	PP2FP000
vosotros	vosotros	PP2MP000
contigo	contigo	PP2CNO00
él	él	PP3MS000
ella	él	PP3FS000
ellas	ellos	PP3FP000
ello	ello	PP3NS000
ellos	ellos	PP3MP000
la	lo	PP3FSA00
las	lo	PP3FPA00
lo	lo	PP3MSA00

Forma	Lema	Etiqueta
lo	lo	PP3CNA00
los	lo	PP3MPA00
le	le	PP3CSD00
les	le	PP3CPD00
se	se	PP3CN000
sí	sí	PP3CNO00
consigo	consigo	PP3CNO00

6.2. Pronombres Demostrativos

Forma	Lema	Etiqueta
aquéllas	aquel	PD0FP000
aquélla	aquel	PD0FS000
aquéllos	aquel	PD0MP000
aquél	aquel	PD0MS000
aquellas	aquel	PD0FP000
aquella	aquel	PD0FS000
aquellos	aquel	PD0MP000
aquel	aquel	PD0MS000
aquello	aquel	PD0NS000
ésas	ese	PD0FP000
ésa	ese	PD0FS000
esas	ese	PD0FP000
esa	ese	PD0FS000
esos	ese	PD0MP000
ese	ese	PD0MS000
ésos	ese	PD0MP000
ése	ese	PD0MS000
eso	ese	PD0NS000
esotra	esotro	PD0FS000
esotro	esotro	PD0MS000
esta	este	PD0FS000
éstas	este	PD0FP000
ésta	este	PD0FS000
estas	este	PD0FP000
esta	este	PD0FS000
estos	este	PD0MP000
este	este	PD0MS000
éstos	este	PD0MP000
éste	este	PD0MS000

Forma	Lema	Etiqueta
esto	este	PD0NS000
estotra	estotro	PD0FS000
estotro	estotro	PD0MS000
tal	tal	PD0CS000
tales	tal	PD0CP000

6.3. Pronombres Posesivos

- Los pronombres posesivos (mío, tuyo, suyo, etc.) se comportan como adjetivos. Se usa la etiqueta PX ya que permite expresar el número del poseedor, lo que no sería posible creando una nueva subcategoría de adjetivos. Los PX aparecerán siempre detrás de un determinante (el mío, la suya, ...)

Forma	Lema	Etiqueta
mía	mío	PX1FS0S0
mías	mío	PX1FP0S0
mío	mío	PX1MS0S0
míos	mío	PX1MP0S0
mío	mío	PX1NS0S0
nuestra	nuestro	PX1FS0P0
nuestras	nuestro	PX1FP0P0
nuestro	nuestro	PX1MS0P0
nuestros	nuestro	PX1MP0P0
nuestro	nuestro	PX1NS0P0
suya	suyo	PX3FS000
suyas	suyo	PX3FP000
suyo	suyo	PX3MS000
suyos	suyo	PX3MP000
suyo	suyo	PX3NS000
tuya	tuyo	PX2FS0S0
tuyas	tuyo	PX2FP0S0
tuyo	tuyo	PX2MS0S0
tuyos	tuyo	PX2MP0S0
tuyo	tuyo	PX2NS0S0
vuestra	vuestro	PX2FS0P0
vuestras	vuestro	PX2FP0P0
vuestro	vuestro	PX2MS0P0
vuestros	vuestro	PX2MP0P0
vuestro	vuestro	PX2NS0P0

6.4. Pronombres Indefinidos

Forma	Lema	Etiqueta
algo	algo	PI0CS000
alguien	alguien	PI0CS000
alguna	alguno	PI0FS000
algunas	alguno	PI0FP000
alguno	alguno	PI0MS000
algunos	alguno	PI0MP000
bastante	bastante	PI0MS000
bastantes	bastante	PI0MP000
cualesquiera	cualquiera	PI0CP000
cualquiera	cualquiera	PI0CS000
demás	demás	PI0CP000
misma	mismo	PI0FS000
mismas	mismo	PI0FP000
mismo	mismo	PI0MS000
mismos	mismo	PI0MP000
mucha	mucho	PI0FS000
muchas	mucho	PI0FP000
mucho	mucho	PI0MS000
muchos	mucho	PI0MP000
nada	nada	PI0CS000
nadie	nadie	PI0CS000
ninguna	ninguno	PI0FS000
ningunas	ninguno	PI0FP000
ninguno	ninguno	PI0MS000
ningunos	ninguno	PI0MP000
otra	otro	PI0FS000
otras	otro	PI0FP000
otro	otro	PI0MS000
otros	otro	PI0MP000
poca	poco	PI0FS000
pocas	poco	PI0FP000
poco	poco	PI0MS000
pocos	poco	PI0MP000
quienquier	quienquiera	PI0CS000
quienesquiera	quienquiera	PI0CP000
quienquiera	quienquiera	PI0CS000
tanta	tanto	PI0FS000
tantas	tanto	PI0FP000

Forma	Lema	Etiqueta
tanto	tanto	PI0MS000
tantos	tanto	PI0MP000
toda	todo	PI0FS000
todas	todo	PI0FP000
todo	todo	PI0MS000
todos	todo	PI0MP000
una	uno	PI0FS000
unas	uno	PI0FP000
uno	uno	PI0MS000
unos	uno	PI0MP000
varias	varios	PI0FP000
varios	varios	PI0MP000

6.5. Pronombres Interrogativos

Forma	Lema	Etiqueta
adónde	adónde	PT000000
cómo	cómo	PT000000
cuál	cuál	PT0CS000
cuáles	cuál	PT0CP000
cuándo	cuándo	PT000000
cuánta	cuánto	PT0FS000
cuántas	cuánto	PT0FP000
cuánto	cuánto	PT0MS000
cuántos	cuánto	PT0MP000
dónde	dónde	PT000000
qué	qué	PT0CS000
quién	quién	PT0CS000
quiénes	quién	PT0CP000

6.6. Pronombres Relativos

Forma	Lema	Etiqueta
como	como	PR000000
donde	donde	PR000000
adonde	adonde	PR000000
cuando	cuando	PR000000
cual	cual	PR0CS000
cuales	cual	PR0CP000
cuanta	cuanto	PR0FS000

Forma	Lema	Etiqueta
cuantas	cuanto	PR0FP000
cuanto	cuanto	PR0MS000
cuantos	cuanto	PR0MP000
cuya	cuyo	PR0FS000
cuyas	cuyo	PR0FP000
cuyo	cuyo	PR0MS000
cuyos	cuyo	PR0MP000
que	que	PR0CN000
quien	quien	PR0CS000
quienes	quien	PR0CP000

6.8. Pronombres Exclamativos

Forma	Lema	Etiqueta
qué	qué	PE000000

7. CONJUNCIONES

CONJUNCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

7.1. Conjunción Coordinada

Forma	Lema	Etiqueta
e	e	CC
empero	empero	CC
mas	mas	CC
ni	ni	CC
o	o	CC
ora	ora	CC
pero	pero	CC
sino	sino	CC
siquiera	siquiera	CC
u	u	CC
y	y	CC

7.2. Conjunción Subordinada

Forma	Lema	Etiqueta
aunque	aunque	CS
como	como	CS
conque	conque	CS
cuando	cuando	CS
donde	donde	CS
entonces	entonces	CS
ergo	ergo	CS
incluso	incluso	CS
luego	luego	CS
mientras	mientras	CS
porque	porque	CS
pues	pues	CS
que	que	CS
sea	sea	CS
si	si	CS
ya	ya	CS

8. INTERJECCIONES

INTERJECCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interjección	I

Ejemplos:

Forma	Lema	Etiqueta
ah	ah	I
eh	eh	I
ejem	ejem	I
ele	ele	I

9. PREPOSICIONES

PREPOSICIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

- Los atributos de género y número tan sólo se especifican para las preposiciones contraídas al y del.
- El analizador actual separa las contracciones en sus componentes, por lo que se obtienen las categorías de la preposición más el artículo por separado.
- Estas etiquetas también se usan para las locuciones preposicionales.

Ejemplos:

Forma	Lema	Etiqueta
al	al	SPCMS
del	del	SPCMS
a	a	SPS00
ante	ante	SPS00
bajo	bajo	SPS00
cabe	cabe	SPS00
con	con	SPS00
a_partir_de	a_partir_de	SPS00
a_causa_del	a_causa_del	SPCMS

10. SIGNOS DE PUNTUACIÓN

SIGNOS DE PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

Forma	Lema	Etiqueta
¡	¡	Faa
!	!	Fat
,	,	Fc
[	[	Fca
]	]	Fct
:	:	Fd
"	"	Fe
-	-	Fg
/	/	Fh
¿	¿	Fia
?	?	Fit
{	{	Fla
}	}	Flt
.	.	Fp
(	(	Fpa
)	)	Fpt
«	«	Fra
»	»	Frc
...	...	Fs
%	%	Ft
;	;	Fx
_	_	Fz
+	+	Fz
=	=	Fz

11. CIFRAS Y NUMERALES

CIFRAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z
2	Tipo	partitivo	d
		Moneda	m
		porcentaje	p
		unidad	u

- Las cifras y numerales se etiquetarán con Z. Bajo esta etiqueta encontraremos: números, direcciones, números de teléfono, tanteos, etc.
- Los numerales partitivos tendrán tipo d. (una docena, un millón, un centenar,...)
- Las cantidades monetarias recibirán la etiqueta Zm, tendrán como lema la cantidad (en cifras) y el nombre de la unidad monetaria en singular.
- Las fracciones i porcentajes recibirán la etiqueta Zp. El lema normalizará la proporcion
- Las magnitudes físicas recibiran la etiqueta Zu. El lema normalizará la unidad de medida y la magnitud

Ejemplos:

Forma	Lema	Etiqueta
239	239	Z
doscientos_veinte	220	Z
un_millón	1000000	Zd
una_docena	12	Zd
tres_de_cada_cuatro	03-abr	Zp
seis_octavas_partes	06-ago	Zp
ochenta por ciento	80/100	Zp
74_%	74/100	Zp
2000_dólares	$_USD:2000	Zm
30_Km_por_hora	SP_km/h:30	Zu
ocho_gramos_/_litro	DN_g/l:8	Zu

12. FECHAS Y HORAS

FECHAS Y HORAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Fecha/Hora	W

Ejemplos:

Forma	Lema	Etiqueta
viernes_26_de_septiembre_de_1992	[V:26:09:1992:??.??]	W
las_tres_de_la_tarde_del_26_de _septiembre_de_1992	[??:26:09:1992:03.00:pm]	W
sábado por la tarde	[S:??:??:??:??.??:pm]	
marzo_de_1954	[??:??:03:1954:??.??:??]	W
siglo_XIX	[s:xix]	W
año_1987	[??:??:??:1987:??.??:??:??]	W
cinco_de_la_mañana	[??:??/??/??:05.00:am]	W

BIBLIOGRAFÍA

Aguilar-Galicia, H., Sidorov, G., & Ledeneva, Y. (2012). Extracción automática de hechos de libros de texto basada en estructuras sintácticas. En G. Sidorov (Ed.), *Avances en Inteligencia Artificial* (Vol. 55, págs. 15-26). México, D.F., México: Instituto Politécnico Nacional.

EAGLES. (s.f.). *Expert Advisory Group on Language Engineering Standards.* Recuperado el 8 de Noviembre de 2012, de http://www.ilc.cnr.it/EAGLES96/home.html

EagV2.0. (s.f.). *Introducción a las etiquetas EAGLES (v. 2.0).* Recuperado el 8 de Noviembre de 2012, de Sitio Web de FreeLing: http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

FreeLing, 2.2, 3.0. (s.f.). Recuperado el 8 de Noviembre de 2012, de http://nlp.lsi.upc.edu/freeling/index.php

Fuentes de la Corte, J. L. (2010). *Gramática Moderna de la legua española.* México, D.F., México: Limusa.

Galicia Haro, S. N., & Gelbukh, A. (2007). *Investigación en análisis sintáctico para el español.* México, D.F., México: Instituto Politécnico Nacional.

Gamallo, P., & Garcia, M. (Abril de 2012). Dependency-Based Open Information Extraction. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (págs. 10-18). Avignon, France: Association for Computational Linguistics.

Gartz, I. (2011). *Análisis de las estructuras del español* (Primera ed.). México, D.F., México: Trillas.

Gelbukh, A., & Sidorov, G. (2010). *Procesamiento automático del español, con enfoque en recursos léxicos grandes* (Segunda ed.). México, D.F., México: Instituto Politécnico Nacional.

Giammatteo, M., & Albano, H. (2009). *Lengua. Léxico, gramática y texto* (Primera ed.). Buenos Aires, Argentina: Biblos.

Herrera de la Cruz, J. A. (2010). *Sistema de extracción automática de información semántica de los libros de textos estructurados.* Centro de Investigación en Computación. México, D.F.: Instituto Politécnico Nacional.

Hovy, E., Zhou, L., & Kwon, N. (2007). A Semi-Automatic Evaluation Scheme: Automated Nuggetization for Manual Annotation. *Proceedings of NAACL HLT 2007. Companion Volume*, pp. 217–220. Rochester, NY: Association for Computational Linguistics.

Jackson, P., & Moulinier, I. (2007). *Natural Language Processing for Online Applications Text Retrieval, Extraction and Categorization* (Second revised ed.). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Jain, A., & Pennacchiotti, M. (2010). *Open information extraction from web search query logs.* Sunnyvale, CA: Yahoo! Labs.

Joosse, W. (2007). *User Trainable Fact Extraction.* Masters thesis, Universiteit Twente de ondernemende universiteit.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing.* USA: Prentice Hall.

Lex, E., & Horn, C. (2012, April 16). Measuring the Quality of Web Content using Factual Information. *ACM*.

Martí Antonín, M. A., & Alonso Martín, J. A. (2003). *Tecnologías del lenguaje* (Primera ed.). España: UOC.

Martí Antonín, M. A., & Alonso Martín, J. A. (2003). *Tecnologías del lenguaje* (Primera ed.). España: UOC.

Mora, A. (2004). *Las partes de la oración* (Segunda ed.). México, D.F., México: Trillas.

Munguía Zatarain, I., Munguía Zatarain, M. E., & Rocha Romero, G. (2000). *Gramática Lengua Española. Reglas y ejercicios.* (Primera ed.). México, D.F., México: Larousse.

Padró, L. (2011). *Analizadores Multilingües en FreeLing.* Centro de Investigación TALP / Universitat Politècnica de Catalunya, Lenguajes y Sistemas Informáticos.

RAE. (s.f.). *Diccionario de la lengua española*, Vigésima segunda edición. Recuperado el 21 de Septiembre de 2011, de Real Academina Española: www.rae.es

SEP. (2010). *Historia. Sexto grado.* (Primera ed.). México, D.F., México: Secretaría de Educación Pública.

SEPb. (2010). *Ciencias Naturales. Sexto grado* (Primera ed.). México, D.F., México: Secretaría de Educación Pública.

Simon, H. A. (1999). *The Sciences of the Artificial* (Third ed.). Massachusetts: The MIT Press.

Sinclair, J. (1996). *EAGLES Preliminary Recommendations on Text Typology*. EAGLES Document EAG-TCWG-CTYP/P.

Turing Center University of Washington. (s.f.). *Open Information Extraction*. (ReVerb, Productor) Recuperado el 30 de Octubre de 2012, de http://openie.cs.washington.edu/