VNIVERSIDAD NACIONAL AVTÔNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE LETRAS HISPÁNICAS

METODOLOGÍA DE ELABORACIÓN PARA UN CORPUS INFORMÁTICO DE CONTEXTOS DEFINITORIOS

T E S I S

QUE, PARA OBTENER EL TÍTULO DE

LICENCIADO EN LENGUA Y LITERATURAS HISPÁNICAS,

PRESENTA:

VÍCTOR GERMÁN MIJANGOS DE LA CRUZ

ASESOR: DR. GERARDO EUGENIO SIERRA MARTÍNEZ

CIUDAD UNIVERSITARIA, 2011

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Cuando hay una tormenta los pajaritos se esconden,

pero las águilas vuelan más alto.

Mohandas Gandhi

Agradecimientos

Agradezco sobre todo a mis padres, a quienes dedicó esta tesis, porque ellos me han hecho el ser humano que soy ahora. Les agradezco por darme la vida y por hacer que obtener un título de licenciatura sea el primero de mis logros.

A mi familia. A mi hermana, que siempre ha estado ahí desde el día que nací; ha sido mi compañera de juegos, con ella he compartido mi vida, el juego más complicado que me ha tocado jugar. A mi tío Pedro quien ha sido un sustento para poder terminar todos mis proyectos. De él he aprendido muchas que no sería capaz de olvidar.

A mi tía Luisi y mis primos Juan Carlos, Cari, Diana y Toni, por sus visitas, su apoyo y todo lo que nos han dado. A mi tío y padrino Alejandro, a mi tía Paula y Adrián y todos sus hijos. A mi prima Silvia y mis sobrinos Caro y Cris, que me ha regalado muy buenos momentos en su presencia. Y gracias a todos mis tío y también padrino, Alejandro, a Blanquita, a Paula, a Adrián, a mi tío Antonio, Magda, Coquis y a mis primos, Arturo, Marcos, Carlos y los que me falten por nombrar.

Agradezco al asesor de esta tesis, el Dr. Gerardo Sierra Martínez, por darme el apoyo necesario para concluirla y por guiarme en el arduo trabajo de realizarla. A mis sinodales: la Mtra. Margarita Palacios Sierras, por sus maravillosas clases y todo el conocimiento y sabiduría que nos regala a todas aquellas personas que tenemos la oportunidad de conocerla. Al Mtro. Javier Cuétara Priede, quien despertó en mí el interés por esta área del conocimiento y por quien conocí al GIL, le doy las gracias por todo el apoyo y el conocimiento que compartió conmigo. Al Mtro. Carlos Méndez Cruz, quien es parte primordial de este proyecto y de esta tesis, ya que sin él no podría haber escrito esta tesis. Al Dr. Alfonso Medina Urrea, por su apoyo no sólo en la tesis, sino en todo el trabajo que he venido desarrollando dentro del GIL.

A todos mis amigos, quienes me han acompañado en todo momento. Primero, agradezco a Marco Antonio, compañero de la primaria y que con el tiempo se ha convertido en un miembro de la familia, al igual que Alma, Doña Naty y, la más pequeña, Mairin.

A la banda de la prepa. A Doris, quien ha sido una muestra viviente de que las personas pueden ser mejores, gracias por enseñarme tanto (muchas veces involuntariamente) y, sobre todo, gracias por seguir siendo mi amiga. A Viri, por escucharme tantas veces y por sus consejos, por su apoyo y por brindarme su amistad a lo largo de tantos años. A Yen-Len Siu, por ser un apoyo constante y por las buenas pláticas y los buenos momentos que hemos pasado juntos. Al buen Gilo y a Pol, con quienes pasé muy buenos momentos en esas reuniones semanales de hace tanto tiempo. A Rulistián, por su buena compañía y su música que ya ha mejorado con el tiempo. A Iván, Claudia, Dulce, Paola, Chucho, Pepe, todos ellos grandes amigos desde hace ya más de 6 años.

A los amigos que conocí en la carrera y que compartieron conmigo esa pasión por la lingüística y/o la literatura. A Angélica, por los buenos momentos que hemos pasado, por su paciencia y sus consejos. A Deneb, a Ernesto y a Kero, grandes y capaces compañeros. A mis profesores, especialmente a Bulmaro Reyes Coria, quien me inspiro para comenzar mi camino por la lingüística.

A los amigos del GIL. A Irasema, quien tan bien compañera de la universidad, a ella le agradezco todo el apoyo en mi carrera académica y personal, gracias por ser esa gran compañera de clases, de trabajo y, sobre todo, por ser una gran amiga. A Alejandro, porque me ha apoyado en todo lo que tenga que ver con el GIL y lo que no, gracias por esas buenas noches de jueves, viernes o cualquier otro día y gracias a su hermano, Fernando, por darme de beber cuando tenía sed y gracias también a Bucéfalo. Gracias a la banda del 12, Teresita y Pavel, a ella por su amistad, su compañía, sus comentarios que siempre me han caído en el momento preciso, y a él por su amistad, por los viajes que hemos realizado juntos y las fiestas que hemos compartido. A Azury y Josh, por ser grandes amigos, por las comidas que hemos compartido, las fiestas, los momentos, por ser una pareja tan fabulosa. A Brenda, por los buenos momentos que hemos pasado, por sus buenas y largas pláticas que ha compartido conmigo. A Bandita, a quien espero ver tan pronto acabe este ciclo. A José Luis, con quien ha pasado los mejores festejos y los viajes más divertidos. A Claudia, por su gusto compartido por los dulces. A Paulina, por ser esa amiga con la que comparto, aunque distinta, la extrañeza. A aquellos que se han ido a continuar con esta pasión a otro país, el estimado Tavito con quien comparto el disgusto por los hippies, a Alemol. A los

maestros y doctores que han compartido conmigo el GIL. A Rodrigo, por creer en mí y hacer nacer en mí ese interés, más certero, en el PLN. A César y Olga, por su buen apoyo, por resolverme todas aquellas dudas que tenía sobre mi trabajo dentro del GIL, por lo buenos momentos que he pasado con ellos dentro y fuera del grupo. A Fernanda, Iria y Juan Manuel, porque de cada uno de ellos he aprendido tanto, gracias por compartir ese conocimiento conmigo y por el apoyo que me han brindado. A Ita, Yanin, Jessy, Cecy, Alisa, Lizzy, Adriana Valerio, Adriana Ballesteros, Ariadna, Adrián, Juan Miguel, César Eduardo. Una larga lista, ojalá no me falte nadie.

A aquellos amigos que no entran en las listas anteriores. A Lilián, de quien he aprendido mucho y con quien he pasado muy buenos momentos. A Dominique, a quien espero poder visitar algún día. A Bárbara, Val, Edgar. A los compañeros de capoeira, el profesor Negaço, Juan, Javier, Zoneka, Mara, Magui, Limpo y quienes me falten por nombrar.

A aquellas criaturas que me han acompañado de una manera especial, a la Gorda, a la Güera y a la Chiquita y, también, a Austria y Sinnombre.

A todos ellos, gracias.

Metodología de elaboración para un corpus informático de contextos definitorios

Contenido

Índice de ilustraciones

Índice de tablas

1. Introducción

1.1. Antecedentes del tema

Este trabajo surge en el marco del proyecto *Extracción automática de definiciones en textos de especialidad*, bajo el apoyo del Comité Técnico de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), y se inserta en el marco de investigación del Grupo de Ingeniería Lingüística (GIL) de la UNAM. La investigación se ha enfocado en la extracción automática de contextos definitorios que se ha venido desarrollando en el GIL y que ha elaborado una vasta producción de artículos y otros medios informativos especializados. A este respecto, se puede mencionar las tesis que han surgido dentro del GIL con respecto al tema, entre algunas de ellas se encuentran *Metodología de Análisis Lingüístico de definiciones en contextos definitorios* (Aguilar 2009), *Análisis lingüístico de contextos definitorios en textos de especialidad* (Alarcón 2003), *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios* (Alarcón 2009), *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática* (Hernández 2009) y *Análisis de relaciones léxicas en definiciones analíticas, extensionales y funcionales* (Sánchez 2009).

El principal motivo que impulsa la realización de esta tesis es la necesidad de agrupar aquellos contextos definitorios que han surgido en las investigaciones del GIL enfocadas en extracción automática de definiciones y, a partir de ellas, reunir dichos contextos definitorios en un corpus que permita que cualquier usuario pueda tener acceso a ellas a través de Internet de manera fácil y práctica. Además, se pretende elaborar una herramienta útil para los investigadores en diversas áreas del conocimiento, puesto que:

> Un corpus de contextos definitorios (CCDs), más allá de ser concebido como un mero repositorio de documentos, es una herramienta valiosa para la terminología y la lexicografía, ya que puede facilitar el proceso de extracción de unidades tales como términos y definiciones (Sierra, Alarcón y Aguilar, y otros 2006).

Al mismo tiempo, se busca la elaboración de un método para la construcción de corpus informáticos, no sólo por medio de la recopilación y revisión de la literatura que se ha

escrito al respecto, sino a partir de métodos que se han venido utilizando dentro del GIL para la elaboración de sus propios corpus lingüísticos informatizados.

1.2. Exposición del proyecto

Este proyecto se enfoca tanto en la descripción de los métodos de análisis y extracción de contextos definitorios como en la descripción de la metodología utilizada para elaborar un corpus informático que se conforme de éstos.

El enfoque desde el cual se realiza esta tesis pertenece a la ingeniería lingüística. Se utilizarán, por tanto, herramientas informáticas para el análisis lingüístico. En este caso, se describirán las herramientas utilizadas para la extracción de contextos definitorios, herramientas que si bien son computacionales utilizan bases de análisis lingüístico para la identificación de patrones verbales recurrentes en definiciones (Aguilar 2009). De igual forma, se utilizarán los métodos de la lingüística de corpus, utilizados actualmente en el GIL, que proponen el uso del estándar XML para el etiquetado. Se emplearán, de igual forma, metodologías de análisis de definiciones y, en específico, de contextos definitorios, expuestos principalmente en Aguilar (2009).

1.2.1. Planteamiento del problema

En el GIL se había llevado a cabo la construcción de un corpus de contextos definitorios (CORCODE) que contaba ya con una interfaz web de consulta (Sierra, Alarcón y Aguilar, y otros 2006). Este corpus contaba con un número pequeño de muestras, no superaban los 200 contextos definitorios; por tanto, se decidió reelaborar el corpus.

La reelaboración del corpus de contextos definitorios presenta, en primera instancia, la cuestión de cómo elaborar dicho corpus con base en los objetivos que se plantean (véase 4.1.1). De igual forma, es de gran importancia detallar la metodología que se siguió para la elaboración del corpus, desde la descripción de las herramientas utilizadas en el proceso de su construcción, hasta la descripción de su estructura y de su arquitectura a nivel computacional.

Asimismo deben tomarse en cuenta los problemas que pueden llegar a presentarse, puesto que en el análisis de los contextos definitorios se presentaron distintas variables:

a) Elementos constitutivos de los contextos definitorios que no se tienen considerados en el marco teórico abordado.
b) Contextos definitorios no prototípicos que dificultan la delimitación de los mismos y la identificación de sus elementos constituyentes.
c) Estructuras que dificultan el proceso de etiquetado del corpus, debido principalmente a que no se han abordado en la teoría que utilizamos.
d) Los diferentes tipos de necesidad que pueden presentarse en los usuarios hacia los que va dirigido el corpus, primordialmente por tratarse de un corpus dirigido a un público amplio.

Este tipo de conflictos se presentaron a través del proceso de elaboración del corpus, de tal forma que debió aplicarse un método para la solución de éstos, ya sea basado en lo esbozado en los trabajos previos a esta tesis o, en su caso, por medio de evaluar los objetivos que se plantearon para el corpus de contextos definitorios.

1.3. Objetivos

El corpus de contextos definitorios se postula como una herramienta de apoyo para la lexicografía y la terminología. El tema y los fines de esta tesis plantean ya el problema de la elaboración de un corpus lingüístico y de la estructura y tipología de definiciones dentro de contextos definitorios, así como del etiquetado de sus constituyentes a partir de la estructura de cada uno de ellos.

Principalmente, se busca elaborar un corpus informático que sirva de consulta para la investigación terminológica, lexicográfica y para aplicaciones dirigidas a la ingeniería lingüística, tales como la extracción automática de información, la creación de diccionarios electrónicos, la elaboración de bancos terminológicos, entre otras. También se persigue que el corpus le sea útil a especialistas, estudiantes y personas interesadas en diversas áreas científicas. Por tanto, la propuesta de esta tesis es la descripción de la metodología implementada para la elaboración de un corpus de tales características.

1.3.1. Objetivos específicos

De manera más específica, los objetivos que busca la presente tesis son los enumerados a continuación:

a) El análisis de los aspectos más generales de la lingüística de corpus, así como de las metodologías de elaboración de corpus lingüísticos informáticos.
b) De igual forma, se busca estudiar los tipos de anotación de corpus lingüísticos que se han trabajado y las herramientas para elaboración de estos corpus.
c) Planeación y descripción de una interfaz para un corpus lingüístico informatizado que se acople a las necesidades y los objetivos del proyecto.
d) La consideración de los aspectos concernientes a la estructura y tipología de definiciones, así como a la estructura de contextos definitorios para su etiquetado y consulta dentro del corpus.
e) Contribuir en la investigación sobre ingeniería lingüística y también en el desarrollo de la ingeniería lingüística dentro de los proyectos apoyados por el GIL.
f) Colaborar en el desarrollo de una herramienta para lingüistas, es decir, la construcción de un corpus que pueda ser útil para las áreas de la lexicografía y de la terminología, un corpus de contextos definitorios.
g) Se busca elaborar un método para la construcción de corpus de este tipo, con base en herramientas computacionales, así como con la utilización de la lingüística de corpus.

1.4. Estructura de la tesis

La tesis se ha distribuido en seis apartados. Su estructura pretende abordar los temas más básicos al principio para adentrarse, posteriormente, en la metodología, la elaboración y la descripción del corpus de contextos definitorios en los apartados 4 y 5. Por tanto, los primeros temas corresponden a un marco teórico acerca de la lingüística de corpus y a la extracción de definiciones en textos de especialidad, es decir, de los contextos definitorios.

A continuación, a manera de síntesis, enumeramos los capítulos por los que se conforma la tesis además de la introducción:

a) *Lingüística de corpus*. En este capítulo se hace una revisión bibliográfica acerca de la lingüística de corpus, de los métodos que se utilizan en ella, de sus herramientas, enfocándose principalmente en los corpus informáticos; se mencionan sus ventajas y se describe la metodología con la que comúnmente se elaboran.
b) *Contextos definitorios*. Este capítulo aborda los estudios anteriores que se han realizado sobre los contextos definitorios. Se busca describirlos, dar una definición de éstos, mencionar su proceso de obtención y su utilidad en diferentes ámbitos lingüísticos.
c) *Metodología y diseño de un corpus para contextos definitorios*. Aquí se describe los procesos que se llevaron a cabo para la obtención, revisión y etiquetado de los contextos definitorios que conforman el corpus. También se mencionan los problemas en el proceso de elaboración.
d) *El Corpus de contextos definitorios: CORCODE*. Posteriormente, se describe el corpus, su interfaz y cómo se puede acceder a ésta a través de internet, las formas de consulta que se pueden realizar, así como los beneficios que representa contar con un corpus de esta índole.
e) *Conclusiones*. Por último, se presentarán las observaciones finales a las que se llegaron durante la elaboración de esta tesis.

2. Lingüística de corpus

En este capítulo abordaremos el tema de la lingüística de corpus y los corpus lingüísticos, comenzando por definir ambos términos para, después, describir las características de los corpus lingüísticos, su tipología y las herramientas computacionales que se han utilizado para su desarrollo, tales como el etiquetado XML. Finalmente, expondremos las ventajas y las desventajas que el uso de corpus presenta en las investigaciones lingüísticas.

2.1. ¿Qué es la lingüística de corpus?

Para comenzar a hablar de un corpus de contextos definitorios, es importante entender la importancia de la lingüística de corpus, así como aclarar qué se entiende por un contexto definitorio. En este capítulo desarrollaremos el concepto de lingüística de corpus y corpus lingüístico. Para ello comenzaremos por definir estos conceptos y finalizaremos por analizar las ventajas tanto del uso de corpus como de los corpus electrónicos frente a los tradicionales.

2.1.1. Definición de la lingüística de corpus

En su forma más general, podemos entender a la lingüística de corpus como el estudio de la lengua a través de análisis de corpus lingüísticos; de una forma más amplia, podemos decir que esta ciencia se plantea el estudio de fenómenos lingüísticos mediante el uso de muestras que se reúnen dentro de los corpus lingüísticos. Por tanto, podemos afirmar que la lingüística de corpus no es una escuela lingüística en sí misma sino una metodología para el análisis de la lengua y para la realización de estudios lingüísticos específicos. Biber, Conrad, y Reppen (1998) ya afirmaban que el enfoque lingüístico basado en corpus nos provee de la posibilidad de realizar una multitud de nuevas investigaciones del uso del lenguaje. Estos mismos autores presentan cuatro características que consideran esenciales en este tipo de análisis lingüístico:

a) Se trata de un estudio empírico, puesto que proporciona los datos para el análisis objetivo del lenguaje.

b) Utiliza una colección de muestras de lenguaje natural como base para el análisis.
c) Utiliza, asimismo, las nuevas tecnologías, como computadoras, para el estudio del lenguaje, combinándolas con técnicas no automáticas.
d) Depende tanto del análisis cuantitativo como de técnicas de estudio cualitativas.

Encaminado en estos puntos, para McEnery y Wilson (2001), la lingüística de corpus puede ser definida, de forma simple, como el estudio de la lengua a partir de muestras de uso reales. Esta perspectiva percibe a la lingüística de corpus como un camino para el análisis de la lengua a partir de la recolección de muestras de uso real del lenguaje. Más adelante, McEnery y Wilson (20012) afirman:

> Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics and so on. All of these disciplines concentrate on describing/explaining some aspect of language use. Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description.

Esto nos muestra que, como ya hemos mencionado, la lingüística de corpus no es una rama o escuela lingüística sino una metodología para el estudio del lenguaje humano a través de la observación de muestras de éste. Por tanto, no es raro que la mayor parte de los avances logrados en el área de la lingüística de corpus estén enfocados a la creación de modelos probabilísticos de la lengua y a la comprobación de éstos, principalmente debido a la naturaleza cuantitativa de los corpus o, dicho de otra forma, a la pertinencia para realizar análisis cuantitativos a partir de los corpus, por ser éstos un catálogo de muestras del lenguaje. Los modelos probabilísticos, entonces, han permitido que se lleven a cabo avances en análisis gramaticales automáticos de textos (Torruella y Llisterri 1999). Sin embargo, como ya mencionaban Biber, Conrad, y Reppen (1998), es primordial que el estudio de un corpus incluya un análisis cualitativo; de esto hablaremos con mayor profundidad en 2.1.2.

2.1.1.1. Corpus lingüísticos

Un corpus lingüístico puede definirse, en su forma más sencilla, como cualquier repertorio de textos; es decir, cualquier colección de textos puede considerarse ya como un corpus, puesto que reúne muestras de lenguaje humano. Para el caso de la filología, el corpus bien

puede entenderse como la recopilación bien organizada de los textos orales o escritos, así como de los documentos que los contienen (Torruella y Llisterri 1999). McEnery y Wilson (2001) aseveran que para la elaboración de un corpus lingüístico deben considerarse los siguientes puntos:

a) Que esté bien seleccionado y que cuente con representatividad (variedad y equilibrio).
b) Que tenga un tamaño finito de palabras.
c) Que cuente con una estructura capaz de ser interpretada por una computadora.
d) Que contenga una referencia estandarizada.

Por otro lado, Sierra (2008) también considera que las características del corpus son la variedad, representatividad y el tamaño del corpus, lo que coincide con los primeros dos criterios de McEnery y Wilson; agrega, además, la consideración de los derechos de autor.

Para una mejor comprensión de estos puntos, es apropiado explicar más ampliamente cada uno de ellos:

Bien seleccionado y representativo. Se entiende por esta característica que el corpus debe contar con variedad en las muestras que lo componen, además de una buena distribución de éstas, de tal forma que se logre representatividad en el muestreo. Una de las críticas de la lingüística chomskiana hacia los corpus se basaba en afirmar que éstos no podían ser representativos del habla, puesto que no podían reunir todas las variedades que puedan presentarse en el lenguaje humano. Debe tenerse en consideración, empero, que la representatividad de los corpus se basa en el equilibrio de las muestras con que están formados; se debe representar lo mejor posible el grupo de estudio que se pretende analizar. Además, estos criterios se formularon hace poco más de treinta años, ahora la lingüística moderna puede valerse de los medios informáticos, los cuales pueden procesar la información de forma más efectiva y más rápida que un ser humano. No obstante, el análisis de corpus no puede formular una teoría general del lenguaje, pero bien puede centrarse en un ámbito específico de éste (McEnery y Wilson 2001).

Para definir la representatividad de un corpus lingüístico, parece importante tener una organización u estructura jerárquica de la población a estudiar, esta estructura puede basarse en los criterios esbozados por Sierra (2008): geográficos, culturales, dialectales, étnicos, temporales, históricos o aquellos que sean necesarios para la investigación. Después de todo, como hemos mencionado, un corpus difícilmente será capaz de mostrar un panorama general del lenguaje humano o de una lengua en específico; se limitará a mostrar una perspectiva de un fragmento de un lenguaje dado, como puede ser el habla de los estudiantes de derecho de la UNAM o del área especializada de la ingeniería en computación en español.

Sin embargo, existen esfuerzos para crear un corpus que tenga una amplitud mayor del lenguaje; a este tipo de corpus se les conoce como corpus generales. Un caso de este tipo de corpus lo podemos encontrar con el Corpus del Español Mexicano Contemporáneo (CEMC)[1]. Este corpus comenzó a construirse desde 1973 gracias a la iniciativa del Dr. Luis Fernando Lara y a un grupo de personas que se dedicó a recolectar las muestras; finalmente se obtuvieron 2 millones de palabras. Empero, aunque este corpus trate de ser más general que otro, abarca muestras de español únicamente mexicano. Se puede argüir, por otro lado, que incluso la lingüística de índole chomskiana no puede abarcar más que muestras del lenguaje que sean de conocimiento del lingüista. Un estudio realizado por un investigador mexicano no podrá abarcar al español de otras áreas geográficas. Mientras que la lingüística de corpus puede tomar muestras más amplias y representativas, como es el caso de los corpus de la Real Academia Española. Un corpus, pues, buscará sobre todo ser representativo ya sea de una forma en particular de una lengua o desde una perspectiva más general (un idioma, corpus monolingües, o varios, corpus multilingües).

Otra forma de buscar representatividad es mediante el equilibrio, el cual consiste en buscar que la información de cada uno de los rubros o áreas contenidas en el corpus sea proporcional. Aunque finalmente, los criterios de selección y representatividad dependen en gran medida de los objetivos del corpus, por lo que es preciso reflexionar sobre lo que se

[1] Más información sobre este corpus puede encontrarse en el artículo Características del *"Corpus del español mexicano contemporáneo"* (Lara, 1987).

busca al momento de elaborar estas herramientas lingüísticas para determinar los criterios a seguir.

Un tamaño finito. Es importante considerar a un corpus como una muestra de tamaño finito; esto es, que contenga un número finito de palabras. Aunque un corpus puede ir creciendo, la mayoría de los corpus están planeados con un límite en su contenido, de tal forma que el análisis lingüístico no sea exhaustivo. El tamaño del corpus debe considerar los objetivos de éste y la cantidad de personas que laboren en su construcción, así como el tiempo en que se planea elaborarlo y los recursos con los que se cuenta. Debe tomarse en cuenta que un uso excesivo de texto puede dificultar tanto su análisis como su construcción. Un corpus más grande no representa necesariamente uno más rico; la riqueza de los corpus consiste, más que en su tamaño, en la variedad y la representatividad. En otras palabras, los aspectos cualitativos de un corpus enriquecen a éste tanto como los cuantitativos.

Una estructura capaz de ser interpretada por una computadora. En la lingüística actual, los corpus exigen un tratamiento computacional, por lo que es importante que en la construcción de éstos se considere una anotación que permita su interpretación y análisis por medio de una computadora. Este tipo de corpus tiene grandes ventajas en el procesamiento sobre los corpus que no cuentan con las capacidades informáticas; una de éstas es su mejor aprovechamiento en las búsquedas que se pueden realizar, así como una mejor manipulación de sus contenidos (McEnery y Wilson 2001).

Una referencia estandarizada. Para McEnery y Wilson (2001) un corpus hace referencia a una variedad específica de lengua y representar esta variedad. El corpus, pues, constituye una referencia del lenguaje que representa. Esto permite que el corpus pueda ser utilizado para otras investigaciones y no únicamente para la que fue originalmente concebido. Torruella y Llisterri (1999), ligado a esto, mencionan como una característica importante la neutralidad del corpus, es decir, la capacidad de un corpus de ser actualizable y reutilizable. La concepción de un corpus bajo este supuesto extiende la vida del corpus, al hacer posibles otras investigaciones con base en el mismo.

Derechos de autor. Considera Sierra (2008) que los derechos de autor son un punto de importante atención para la elaboración de un corpus lingüístico, puesto que muchas veces

los corpus buscan ser accesibles para un público amplio. Es importante, por tanto, respetar la propiedad intelectual de los textos que se exponen dentro del corpus. Aunque también se toma en cuenta que hay excepciones a las normas donde no se requiere una autorización del uso de las obras, por ejemplo, cuando los fines del corpus están dirigidos a la investigación o docencia sin fines de lucro; en este caso, se debe indicar la procedencia de los textos y se debe tener acceso sólo a fragmentos de éstos. Bajo estos criterios de derechos de autor se ha desarrollado el corpus presentado en esta investigación (véase 4.2.1).

Ahora bien, dentro de la definición de qué es un corpus lingüístico, Atkins, Clear y Ostler (1992) y Llisterri y Torruela (1999) proponen una clasificación de tipos de conjuntos textuales, la cual mencionamos a continuación:

a) *Archivo o colección informatizados*. Se trata de una colección de textos en soporte informático que no tienen ni buscan ninguna relación entre éstos.
b) *Biblioteca de textos electrónicos*. Se refiere a una colección de textos archivados de forma informatizada y con un formato estandarizado que, aunque siguen ciertos criterios en el aspecto de sus contenidos, no tienen un criterio amplio para la selección de éstos.
c) *Corpus informatizado*. El corpus es un conjunto de textos que contienen muestras de habla que se seleccionan a partir de criterios lingüísticos. Para Atkins, Clear y Ostler (1992), el corpus es una parte de la biblioteca de textos electrónicos. Los textos que lo componen tienen un procesamiento informático que consiste en una codificación estandarizada y homogénea para su procesamiento computacional.

Atkins, Clear y Ostler (1992) también incluyen al *subcorpus* dentro de esta clasificación; sin embargo, nosotros lo consideramos, de acuerdo con Llisterri y Torruela (1999), dentro de otra clasificación, que veremos en 2.1.3. Por el momento, podemos ver que un corpus es una conformación de textos organizados bajo determinados criterios lingüísticos. Es una colección de muestras de un lenguaje determinado seleccionadas y ordenadas bajo criterios lingüísticos estipulados que permite abarcar un panorama de una lengua natural (Sinclair 1996). Son importantes los criterios que los distinguen de otros tipos de colecciones; según lo que acabamos de describir, un corpus no puede ser definido como lo hicimos al principio

de este apartado; si bien es un repertorio de documentos textuales u orales, se construye con criterios lingüísticos bien determinados y se anota bajo estos supuestos para que su manejo sea más sencillo. En palabras de Sierra (2008446):

> es un conjunto de datos reales y aceptables, debidamente ordenado, codificado y organizado, de diferentes textos recopilados, pertenecientes a un código lingüístico determinado, oral o escrito.

2.1.2. Análisis cuantitativo y análisis cualitativo

Un corpus, debido a su constitución y sus características, puede ofrecer en su análisis tanto datos de tipo cuantitativo como de tipo cualitativo, puesto que permite que el estudio de las muestras contenidas en éste se realice, ya sea desde una perspectiva estructural o bien por medios estadísticos. Cabe resaltar que estas dos formas de análisis están estrechamente ligadas y los datos cuantitativos también arrojan información de formas estructurales de la lengua; es decir, obtenemos datos como flexiones verbales, formas de afijos, estructuras sintácticas, etcétera.

Cuando hablamos de un análisis cuantitativo nos referimos al uso de los datos arrojados por el corpus como una base para describir aspectos del uso real de la lengua a partir de datos estadísticos, mientras que el análisis cualitativo se enfoca en la estructura de la lengua, es decir, del análisis de las muestras contenidas en el corpus desde una perspectiva lingüística más descriptiva.

Ambos tipos de análisis revelan información relevante para un estudio lingüístico a partir del uso de corpus; ambos son de gran utilidad para la investigación en un área del lenguaje o un idioma específico, puesto que muestran dos perspectivas de los fenómenos del lenguaje que, como mencionamos, están estrechamente ligadas. Biber, Conrad y Reppen hacen notar la importancia del uso combinado de estas dos perspectivas, alegando que en el análisis basado en corpus “it is essential to include qualitative, functional interpretations of quantitative patterns” (19983). De igual forma, McEnery y Wilson (200177) aseveran que: “Qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalisable results”.

2.1.3. Tipología de corpus

Los corpus lingüísticos pueden ser clasificados de distintas maneras según determinados criterios. Una primera tipología básica de los corpus lingüísticos distingue entre corpus textuales y corpus orales (Jiménez Pozo 1999). Como su nombre lo dice, un corpus textual es aquel conformado por documentos que contengan únicamente muestras de lenguaje escrito, mientras que el corpus oral lo constituyen transcripciones de lengua hablada o grabaciones de ésta. Los corpus textuales pueden considerarse como repertorios de escritos, ya sean físicos o electrónicos, y, por su naturaleza, su manejo es más sencillo para una computadora. Por el otro lado, los corpus orales son repertorios de muestras del lenguaje hablado y su procesamiento digital es más complicado que el del lenguaje escrito.

Otra clasificación es la dada por Llisterri y Torruela (1999), que se asemeja a la de Atkins, Clear y Ostler (1992), y está basada en el nivel de distribución dentro del corpus; estos autores distinguen entre *corpus, subcorpus* y *componente*, los cuales explicamos a continuación:

a) *Corpus*. Un corpus es definido por los autores como "un conjunto homogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.), los cuales se toman como modelo de un estado o niveles de lengua predeterminado" (Torruella y Llisterri 1999, 8). Este conjunto de enunciados contenido en un corpus, para los autores, debe permitir un análisis que dé pie al mejoramiento en el conocimiento de las estructuras al interior del sistema lingüístico que representan.
b) *Subcorpus*. En su forma más simple, se puede definir como un subgrupo de un corpus (Atkins, Clear y Ostler 1992). Para Llisterri y Torruela un subcorpus puede ser de dos tipos: el primero está representado por una selección estática de textos derivados de un corpus de mayor tamaño y complejidad, que divide, a su vez, en muestras textuales más específicas; el segundo tipo es definido como una selección dinámica de textos pertenecientes a un corpus en crecimiento, en otras palabras se trata de textos cuyo fin es integrarse al apartado de un corpus general, de mayor tamaño.
c) *Componente*. Es una colección de muestras de un corpus o subcorpus, las cuales responden a un criterio lingüístico específico muy concreto. Los componentes

reflejan un tipo determinado de lengua. Podemos decir que tanto los corpus como los subcorpus son muy heterogéneos, mientras que los componentes son muy homogéneos.

Para desarrollar una división más amplia de los tipos de corpus, tomamos en consideración principalmente el trabajo sobre tipología de corpus de Sierra y Rosas (2009)[2], quienes clasifican los corpus a partir de los siguientes criterios:

2.1.3.1. El origen de los elementos.

Según el origen de sus elementos se considera a un corpus como oral y escrito; esto se refiere a que —como sus nombres lo dicen— responden a las muestras que conforman el corpus. Tales muestras pueden ser de tipo oral o material escrito, las primeras responden a grabaciones o transcripciones fonéticas o fonológicas del lenguaje hablado, mientras que las segundas son propiamente muestras de lenguaje escrito.

2.1.3.2. La codificación y anotación

Conforme a la anotación, la distinción es la misma sugerida por McEnery y Wilson (2001) y Torruella y Llisterri (1999); es decir, se distingue entre un corpus simple y un corpus anotado o codificado. De igual forma, Sierra y Rosas (2009) proponen un esquema más amplio de tipos de anotación, que veremos más adelante (2.2.1).

2.1.3.3. La especificidad de los elementos

Conforme a la especificidad de sus elementos, se distingue entre dos tipos de corpus: los corpus generales y los corpus especializados o específicos. Los primeros aportan información de tipo general, esto es que recogen todo tipo de géneros y tipologías textuales. Los corpus especializados, por su parte, recogen información de una o varias áreas en particular; éstos pueden, a su vez, ser informativos y contener textos periodísticos,

[2] Otros autores como Atkins, Clear y Ostler (1992) o Llisterri y Torruela (1999) también presenta una tipología de corpus basada en criterios similares a los presentados por Sierra y Rosas. Estos últimos autores se basan, empero, en el trabajo de Llisteri y Torruela, además de otros autores, varios de los cuales se incluyen en este trabajo; no mencionan, sin embargo, a Atkins, Clear y Ostler, aunque, como hemos dicho, los criterios y la clasificación que se hace de los corpus son similares para todos los autores consultados.

científicos o similares, mientras que, por otro lado, están los literarios, que se enfocan a textos del área de la literatura.

2.1.3.4. La temporalidad

Conforme al criterio de temporalidad, una primera distinción se hace entre los corpus diacrónicos y los corpus sincrónicos. Un corpus diacrónico puede definirse como aquel que responde a diferentes períodos de tiempo, mientras uno sincrónico sólo responde a un período temporal. Los corpus diacrónicos, a su vez, pueden subdividirse en cronológicos y periódicos; los cronológicos contienen textos de años en orden consecutivo; los periódicos se encargan de estudiar la lengua en diversos periodos históricos. Los corpus sincrónicos también presentan una subdivisión en contemporáneos e históricos; los contemporáneos se componen de textos actuales, mientras los históricos de textos de una época pasada, sin llegar a abarcar más de un periodo temporal.

2.1.3.5. El propósito

Según el propósito, encontramos también dos tipos de corpus. El primero de propósito específico y el segundo multipropósito. Los de propósito específico son corpus construidos para un estudio lingüístico concreto, a diferencia de los multipropósito que tratan de abarcar un análisis lingüístico más amplio y por tanto pueden ser reutilizables para diferentes investigaciones.

2.1.3.6. El lenguaje

A partir del criterio de la lengua, encontramos los corpus monolingües, es decir que cuentan con textos en un solo idioma; los corpus comparables son una especie de corpus monolingües que cuentan con traducciones de textos a una misma lengua; por último, los corpus multilingües, que consisten en textos en varios idiomas[3].

[3] La distinción que hace Atkins, Clear y Ostler (1992) sobre el idioma de los corpus considera también tres tipos. El primero de ellos es el monolingüe y otros dos que son el bilingüe y el plurilingüe, los cuales entrarían dentro de lo que aquí hemos llamado multilingüe.

2.1.3.7. La cantidad de texto

De acuerdo con la cantidad de texto, tenemos los siguientes tipos: corpus grande, que pueden contener una cantidad considerable de texto (desde diez millones de palabras); corpus pequeño que son aquellos que contienen una cantidad menor de textos y corpus monitor, que contiene un volumen fijo que se actualiza constantemente.

2.1.3.8. La distribución de los textos

Por la distribución de los textos, clasificamos los corpus en desequilibrados y equilibrados. El corpus desequilibrado contiene textos en cantidades no proporcionales entre sí; por otro lado, el equilibrado procura distribuir sus textos de manera proporcional entre sí, dentro de éste se encuentran los corpus piramidales, cuyos textos están distribuidos en diferentes niveles ascendentemente; el primer nivel de un corpus de este tipo contiene poca variedad temática en una cantidad grande de textos, el segundo contiene más variedad temática en menos textos, el tercer nivel tiene más variedad temática en pocos textos y así sucesivamente, según se determinen los niveles de la pirámide en el corpus.

2.1.3.9. La documentación

Según su documentación, se tienen los corpus documentados y no documentados; los primeros contienen registros de la documentación de los textos que permiten hacer búsquedas específicas y conocer la proveniencia de los textos, mientras que los segundos carecen de esto.

2.1.3.10. La autoría

Según la autoría se tienen dos tipos de corpus: los canónicos y los genéricos. Los primeros responden a textos de un único autor, mientras que los segundos responden a documentos de un solo género literario. Por otra parte, cuando la información contenida en el corpus no responde a ninguno de estos dos criterios, se puede decir que tenemos un corpus de autoría variada.

2.2. Corpus electrónicos

Como hemos mencionado más arriba (2.1.1.1), Torruella y Llisterri (1999) proponen una clasificación de conjuntos textuales informatizados que distingue entre tres tipos de éstos: los archivos o colecciones informatizados, las bibliotecas de textos electrónicos y los corpus informatizados. De estos últimos ya hemos esbozado una descripción; podemos decir, además, que cuando los corpus se encuentran en un formato computacional presentan muchas ventajas sobre un formato no informatizado.

Hoy en día, las herramientas informáticas juegan un papel importante en la conformación de los corpus; debido a las grandes cantidades de texto que deben ser procesadas, los medios computacionales se han vuelto indispensables gracias a la gran utilidad que aportan para procesar de manera fácil y rápida grandes cantidades de información. Se hace innegable, entonces, que "Cada vez parece más evidente la conveniencia de utilizar recursos informáticos en las investigaciones humanísticas" (Torruella y Llisterri 1999, 1).

Una de las principales características de los corpus informáticos es el uso del etiquetado, el cual permite anotar información metatextual dentro de éstos, lo que permite que las búsquedas en el interior de los corpus sean más exactas y eficientes, combinado con el soporte electrónico, más rápidas, pues ya no se realizan de forma manual a través de grandes cantidades de texto, sino que se pueden crear interfaces informáticas que permitan a los usuarios revisar cualquier información necesaria dentro del corpus a partir de las etiquetas de éste.

Se puede entender al corpus informatizado como textos elegidos y anotados bajo normas y criterios de análisis lingüístico, se sirve de las herramientas computacionales que permiten obtener resultados más precisos que un corpus no informatizado (Sierra, 2008; Torruella y Llisterri, 1999). Incluso este mismo autor entiende que cualquier corpus es una colección de textos que han pasado por un proceso informático (etiquetado) y ya no cualquier colección de textos, lo que nos deja ver la necesidad de la utilización de medios computacionales en los estudios lingüísticos. Por tanto, la lingüística de corpus se dedica al estudio de grandes muestras textuales para la obtención de información lingüística a partir

de medios meramente computacionales, puesto que esta gran cantidad de datos sería inabordable de otro modo.

2.2.1. Etiquetado

Conforme a la tipología de corpus, hemos visto que Sierra y Rosas (2009) consideraban el corpus simple y uno codificado o anotado; de igual forma, McEnery y Wilson (2001) considera dos tipos de corpus: los no-anotados (unannotated) y los anotados (annotated). Los primeros hacen referencia a texto plano, es decir, a un texto que si bien computarizado, no tiene anotaciones metatextuales; por su parte, un corpus anotado cuenta con etiquetas que referencian a información lingüística no explícita en el texto en sí, esto es, por ejemplo, una anotación de los componentes sintácticos de la oración, de rasgos fonéticos, etcétera.

Torruella y Llisterri (1999) también hacen una clasificación similar según la codificación y anotación de los corpus; ellos distinguen entre un corpus simple y uno codificado o anotado. El primero corresponde al mismo que el no-anotado, es decir, a aquel sin codificación alguna; por otro lado el segundo lo definen como:

> Corpus formado por textos a los cuales se les ha añadido, ya sea manual o automáticamente, etiquetas declarativas de algunos elementos estructurales de los documentos […] —codificación— o etiquetas analíticas de algunos aspectos lingüísticos […] —anotación (Torruella y Llisterri 1999, 12).

Conforme al etiquetado de los corpus, McEnery y Wilson (2001), referenciando a Leech (1993), enumeran siete máximas para la aplicación de la anotación a la información, las cuales son las siguientes:

a) Debe ser posible retirar la anotación para que se tenga un corpus en texto plano si es necesario.
b) Debe ser posible extraer la anotación y almacenarse.
c) El esquema de anotación debe estar basado en directrices que estén al alcance del usuario final.
d) Debe quedar claro cómo y por quién se llevó a cabo la anotación.

e) El usuario final debe percatarse de que la anotación del corpus no es infalible, sino simplemente una herramienta potencialmente útil.
f) Los esquemas de anotación deben estar basados, en lo posible, en principios estándar y en una teoría neutral.
g) Ningún esquema de anotación debe ser considerado *a priori* un estándar. Los estándares surgen a través de consensos prácticos.

Un corpus anotado trae ventajas sobre uno que no lo esté. Las etiquetas utilizadas permiten que se reconozcan patrones meta-informativos que facilitan las tareas de exploración en el corpus, así como la manipulación de éste por los usuarios que lo consulten; mientras que un corpus sin anotaciones resultará en búsquedas más tediosas y de menos facilidad para quien lo consulte.

Un etiquetado de corpus consiste, básicamente, en el marcaje de determinadas partes de texto que indiquen datos lingüísticos que puedan utilizarse para determinado estudio. Sierra y Rosas (2009) proponen una clasificación del tipo de etiquetado según el nivel de lengua al que se haga referencia; el siguiente cuadro presenta esa propuesta:

<table>
<tr><td rowspan="8">Corpus Codificado o Anotado</td><td rowspan="3">Textual</td><td>Estructura Textual</td></tr>
<tr><td>Tipología Textual</td></tr>
<tr><td>Ortográfica</td></tr>
<tr><td rowspan="2">Morfológica</td><td>Lematización</td></tr>
<tr><td rowspan="3">POS Tagging</td></tr>
<tr><td rowspan="3">Sintáctica</td></tr>
<tr><td>Chunking</td></tr>
<tr><td>Parsing</td></tr>
<tr><td rowspan="2">Semántica</td><td>Características Semánticas</td></tr>
<tr><td>Semánticas Ontológica</td></tr>
</table>

		Relaciones
	Fónica	Fonética
		Fonológica
		Prosódica
	Discursiva	Anafórica y Referencial
	Pragmática	

Tabla 2.1. Etiquetado por niveles lingüísticos

Una anotación por estructura textual es la que busca marcar los componentes de un corpus a partir de diferentes niveles dentro del documento, como pueden ser párrafos, oraciones, secciones, etc., como ejemplo, un corpus formado de novelas en las que éstas se anotan por capítulos. Por otro lado, el marcado de la tipología textual implica indicar el tipo de texto que se está utilizando, como puede ser novela, cuento, revista, artículo especializado, etc. La anotación ortográfica, como su nombre lo dice, consiste en marcar los elementos ortográficos de un corpus.

El etiquetado morfológico puede realizarse por lematización, que consiste en el marcado de los lemas de las palabras. También se puede realizar por truncamiento, proceso que radica en eliminar las partes flexivas de las palabras, los afijos, para llegar a las raíces de éstas. Existen también otro tipo de anotaciones morfológicas que no mencionamos. Entre la anotación morfológica y sintáctica, tenemos el llamado POST o *Part of Speech Tagging*, el cual consiste en etiquetar morfosintácticamente las oraciones que componen el corpus.

A nivel sintáctico, el *chunking* hace referencia al etiquetado de patrones sintácticos básicos, como pueden ser sintagmas nominales, sintagmas verbales, sintagmas preposicionales, etc. Por otro lado, el *parsing* es el proceso de etiquetado de los componentes sintácticos de la oración y su relación con los otros componentes, que al final dará como resultado un árbol sintáctico o una representación en paréntesis categorizados, por ejemplo:

> [O [SN El_Art hombre_Sust SN] [SV vio_Vb [SP a_Prep [SN la_Art nena_Sust SN] SP] [SP en_Prep [SN el_Art parque_Sust SN] SP] [SP con_Prep [SN el_Art telescopio_Sust SN] SP] SV] O]

Conforme a la anotación semántica tenemos tres tipos: el primero de ellos, de características semánticas, que se dedica a anotar los semas o rasgos semánticos de una palabra (McEnery y Wilson 2001); por relaciones semánticas, que se refiere al etiquetado de las relaciones que se presentan entre las palabras de una oración, tales como homonimia, hiperonimia, sinonimia, antonimia, etc.; por último, la anotación ontológica consta de etiquetas que sirven para relacionar sentidos entre las palabras, utilizando tanto las características semánticas como las relaciones entre éstas. Con este último tipo de anotación se pretende crear ontologías que, entre otras cosas, permitan un acceso inteligente a los recursos de navegación y búsqueda de información en la web, lo que se ha dado en llamar web 3.0 o web semántica.

Cabe señalar que no existen estándares de anotación semántica bien establecidos, sino que, muchas veces, este tipo de marcaje se define por el usuario dependiendo de sus intereses: verbigracia, en este proyecto se utilizó un marcaje determinado según los elementos que constituían cada uno de los contextos definitorios.

La anotación fónica puede ser de tipo fonético, es decir, enfocada a la lengua hablada; fonológica, señala los fonemas con los que se constituye una palabra; y prosódico, que comprende el etiquetado de elementos suprasegmentales del lenguaje hablado, tales como entonación, intensidad y otros.

También se pueden hacer marcados discursivos y pragmáticos, que consisten en la anotación de rasgos de la lengua pertenecientes al nivel del discurso y de la praxis, es decir, de la construcción y estructuración del sentido de un texto o acto elocutivo. Tales rasgos pueden ser, por ejemplo, la relación entre un párrafo y otro, los emisores de determinados fragmentos textuales, o bien de sentido pragmático, como la anotación del sentido cómico, sarcástico u de otro tipo de algún texto.

2.2.1.1. Lenguaje XML

Por otra parte, debe tomarse en cuenta que para la anotación de corpus se debe utilizar un lenguaje de etiquetado. El que consideramos para esta tesis fue el XML (eXtensible Markup Language), el cual se basa en etiquetas de apertura y cierre que pueden ser

determinadas por el usuario que se ocupe de la elaboración del corpus, lo cual permite elaborar marcas definidas para objetivos específicos.

Fue creado por el World Wide Web Consortium (W3C)[4] con el propósito de estructurar la información contenida en internet. Este consorcio define a su lenguaje de etiquetado como un formato sencillo basado en texto para representar información estructurada, como pueden ser documentos, libros, transacciones, etcétera.

Se caracteriza por el uso de picoparéntesis (< >) que encierran un metadato predefinido por el usuario. Se utiliza en conjunto con un DTD (Definición de Tipo de Documento) o un Esquema XML para especificar su estructura. Generalmente, un documento XML comienza por un encabezado que contiene información sobre el documento y un cuerpo que contiene lo que sería, en nuestro caso, el corpus. Un esquema sencillo de esta organización puede ser el siguiente:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<CUERPO_DEL_TEXTO>
Aquí se contiene el corpus y pueden usarse otras etiquetas para el marcado de los rasgos concernientes.
</ CUERPO_DEL_TEXTO >
```

El corpus se ubica entre las etiquetas de cuerpo y es ahí donde se marcan los hechos lingüísticos relevantes para la investigación, en nuestro caso, las partes constituyentes de los contextos definitorios, proceso que se explicará en 4.2.3.2.

Por su parte, algunas de las ventajas que el W3C encuentra en este lenguaje de marcado frente a otros estándares son los siguientes:

a) Redundancia, es decir que trata de ser incluir en el marcado toda la información posible para evitar errores en su ejecución.

[4] Para mayor información acerca de XML puede consultarse la página de la W3C en la dirección: www.w3.org.

b) Autodescriptible o, en otras palabras, resulta fácil entender este lenguaje debido, principalmente, a que las etiquetas son asignadas por el usuario y se pueden agregar atributos que especifiquen el contenido de cada una de las etiquetas.
c) Su aceptación en la red por ser legible para cualquier analizador de XML, es decir, puede ser procesado por diferentes herramientas y, por tanto, utilizado para diferentes propósitos. Finalmente el W3C afirma que XML es el lenguaje más utilizado a nivel mundial, lo que le añade un valor de universalidad.

Un ejemplo de cómo se etiqueta XML se presenta a continuación, en donde se marcan los elementos que constituyen la cita de un libro:

```
<LIBRO>
        <TITULO> Cien años de soledad</TITULO>
        <AUTOR>Gabriel García Márquez</AUTOR
        <EDITORIAL>Cátedra</EDITORIAL>
        <AÑO>1987</AÑO>
        <ISBN>84-376-0494-X</ISBN>
</LIBRO>
```

Como observamos, se marcan cada una de las partes relevantes del libro, lo que hará más sencillo su proceso por medio de un soporte computacional.

2.2.2. Lingüística de corpus en México

En México se han desarrollado grandes e importantes corpus lingüísticos. Uno de éstos es Corpus del Español Mexicano Contemporáneo (CEMC) coordinado por el Dr. Luis Fernando Lara del Colegio de México (COLMEX) para el *Diccionario del español de México* (DEM), del cual se han desprendido otros trabajos como el *Diccionario fundamental del español de México* o el *Diccionario del español usual en México*.

De igual forma, el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), de la UNAM, ha desarrollado el corpus DIME (Diálogos Inteligentes Multimodales en Español) y posteriormente el DIMEX-100 bajo la responsabilidad del Dr. Luis Pineda y del Mtro. Javier Cuétara; se trata de corpus orales que buscan la creación de sistemas de conversación en español; el primero, DIME, se trata de habla espontánea y el segundo, DIMEX-100, contiene muestras de habla controlada.

Un aporte importante para la lingüística de corpus en México fue la creación del Grupo de Ingeniería Lingüística (GIL) en 1999, también de la UNAM, bajo la dirección del Dr. Gerardo Sierra Martínez. El GIL ha desarrollado investigaciones enfocadas directamente a la creación de corpus informáticos. El mismo Sierra (2008) ha descrito a la lingüística de corpus como:

> se ha llamado *lingüística de corpus* a aquella parte de la lingüística en la que se estudian con medios informáticos de diferentes tipos grandes masas de datos, inabordables de otro modo, para obtener de ese análisis, por ejemplo, las características lingüísticas de una lengua en un cierto momento de su historia, de cierto tipo de textos, de un conjunto de autores o un autor determinado, etc.

Podríamos decir que el GIL ha sido pionero de la elaboración de corpus informáticos en México; de igual forma, han desarrollado herramientas que apoyan la creación de corpus lingüístico y, a su vez, se ha encargado de producir distintos corpus enfocadas en diversas áreas del conocimiento. Enumeramos a continuación algunos de los corpus que han surgido dentro del GIL:

a) Corpus Lingüístico en Ingeniería (CLI). Desarrollado en 2004 con el propósito de reunir información de las áreas de las ingenierías. Este corpus se encuentra disponible en internet en la liga http://www.iling.unam.mx/cli/.
b) Corpus Histórico del Español en México (CHEM)[5]. El CHEM reúne documentos en español de los siglos XVI a XXI y también cuenta con una interfaz disponible en internet en http://www.iling.unam.mx/chem/.
c) Corpus de textos Científicos en Español de México (COCIEM). Este corpus surge como parte de la colaboración entre el GIL y el COLMEX para el *Diccionario Científico de México*.
d) Corpus de las Sexualidades en México (CSMX). Este corpus busca reunir información acerca de sexualidad a partir de textos especializados, así como de aquellos extraídos de foros de discusión. Actualmente también cuenta con una interfaz en línea y se pretende utilizarlo para el desarrollo de un Diccionario de la

[5] Para más información sobre este corpus, resulta muy enriquecedor la consulta del artículo "Arquitectura del Corpus Histórico del Español en México (CHEM)" (Medina y Méndez 2006)

Sexualidad en México y se encuentra disponible en la siguiente liga de internet: http://www.iling.unam.mx/csmx/.

Por último, podemos mencionar el corpus que este trabajo aborda, el cual también se ha desarrollado dentro del GIL con las herramientas que este grupo de investigación ha aportado a la lingüística de corpus, lo que explicaremos más adelante.

2.2.3. Utilidad de los corpus lingüísticos

Los corpus lingüísticos traen ventajas al análisis lingüístico empírico y al procesamiento de lenguaje natural, en cuanto a que son una herramienta de gran utilidad para el estudio de distintos factores en la lengua textual o escrita. Sin embargo, la lingüística de pensamiento chomskiano se opone al uso de corpus para los estudios lingüísticos, principalmente arguyendo que éstos tienen una falta de representatividad del lenguaje humano. El punto de vista de esta aseveración se basa en la idea de que un corpus es una muestra pequeña del lenguaje, siendo este último potencialmente infinito y, por tanto, no sería posible abarcar el lenguaje de una población entera de hablantes. Tenemos que tomar en cuenta, en primera instancia, que los avances tecnológicos han contribuido a la conformación de corpus de grandes escalas, que contienen millones de palabras de un idioma o de varios. Sobre todo, es importante considerar la representatividad con que debe formarse un corpus, las muestras que debe contener y el equilibrio entre éstas, pues esto permite un análisis más amplio a partir de muestras de una población determinada. Debemos pensar, también, que es imposible abarcar una muestra totalizadora de un lenguaje humano, por lo que un corpus parece una herramienta más apropiada para un análisis lingüístico. Éste abarca una muestra más amplia que la que se puede obtener a partir de un solo individuo investigador y, por tanto, es una muestra de mayor variedad. Esto implica que las conclusiones a las que se llegue con una análisis basado en corpus son válidas para una mayor población de hablantes (Sierra 2008).

No pretendemos, empero, profundizar aquí en esta discusión, para ello puede consultarse McEnery y Wilson (2001), quienes dedican parte de su primer capítulo a esbozar esta disyuntiva.

Resta decir que a partir de un corpus lingüístico se puede obtener información tanto de orden cuantitativo como cualitativo que es de gran utilidad para los estudios lingüísticos en todos los niveles de lengua. Algunos de los campos de estudio que más se benefician son la terminología y la lexicografía:

> La lexicografía y la terminología son dos de los campos de investigación y de estudio que más se benefician de las informaciones que los corpus textuales y los corpus de lengua oral aportan. Éstos son de gran ayuda para configurar el lemario de los diccionarios […], así como para separar las distintas acepciones de cada lema, para detectar las palabras co-ocurrentes, las combinaciones sintácticas, etc. Los corpus también proporcionan material muy útil para trabajar sobre fraseología, la detección de neologismos y la obtención de ejemplos reales susceptibles de aparecer en los diccionarios (Torruella y Llisterri 1999, 4).

Por otra parte, los beneficios que ofrece el soporte informático también son notables, pues permiten mayor accesibilidad, aumentan la rapidez con que se realizan las búsquedas, permiten con mayor facilidad la reutilización del material que compone el corpus, así como dan pie a anotaciones metatextuales que pueden ser de utilidad no sólo para la computadora, sino para el usuario final del corpus. Los corpus electrónicos traen muchas ventajas al análisis lingüístico, pero también traen ciertas desventajas.

3. Contextos definitorios

Para entender de forma más clara el concepto de corpus de contextos definitorios, ya hemos explicado de qué se trata la lingüística de corpus y los corpus; ahora cabe presentar el concepto de contexto definitorio, comenzando, también, por tratar de esbozar una definición y proseguir en la descripción de su estructura para, finalmente, entender su uso.

3.1. Definición de contexto definitorio

Dentro del GIL se han desempeñado investigaciones enfocadas a la terminología y a la extracción automática de términos. Dentro de este marco, se ha desarrollado lo que se ha dado en llamar contextos definitorios (CDs). En primera instancia, cabe definir lo que se ha entendido por este concepto. Podemos entender por contexto definitorio un fragmento textual extraído de un documento especializado en donde se encuentra información útil que nos ayude a entender un término; en otras palabras, es un fragmento textual en el cual se define un término.

Un contexto definitorio se identifica a partir de patrones definitorios (3.2.2) que, según explican Sierra y Alarcón (2002), los autores utilizan para introducir la definición de un término que se ha expuesto en el texto. De esta forma, se ha desarrollado en el GIL un sistema (ECODE) que extrae automáticamente este tipo de patrones y, por tanto, un término que esté siendo explicado por una definición, es decir, un contexto definitorio.

3.2. Estructura de un contexto definitorio

Un contexto definitorio fundamentalmente debe contar con un término, un patrón definitorio y una definición. Sierra y Alarcón (2002437) describen que: “in a first state we delimitated a definitory context for those structures integrated by a term (T) and its definition (D), where both parts belong to the same paragraph. Then we will attempt to analyse more complex forms”. Otros de los elementos, además del término y la definición, con los que puede contar un contexto definitorio son los patrones pragmáticos (3.2.4). Un

esquema de la estructura con la que puede contar un contexto definitorio es la que propone Aguilar (2009):

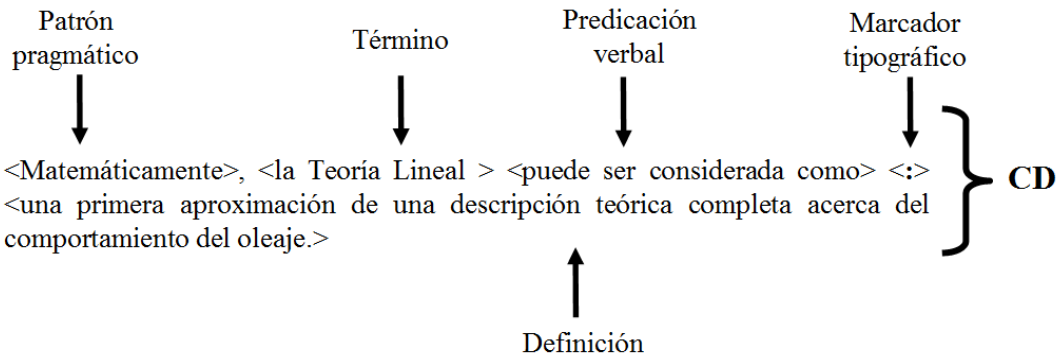

Ilustración 3.1. Estructura de un contexto definitorio (Aguilar 2009)

Los elementos constitutivos de un contexto definitorio se construyen mediante una estructura sintáctica que permite la identificación automática de éstos dentro de un texto más grande. En palabras de Aguilar (2009): "términos, definiciones, predicaciones, patrones pragmáticos y marcadores tipográficos, mantienen una relación estrecha entre sí, lo que permite que los CDs tengan cohesión y coherencia discursivas". La siguiente ilustración muestra la relación que se da entre los distintos elementos de un contexto definitorio:

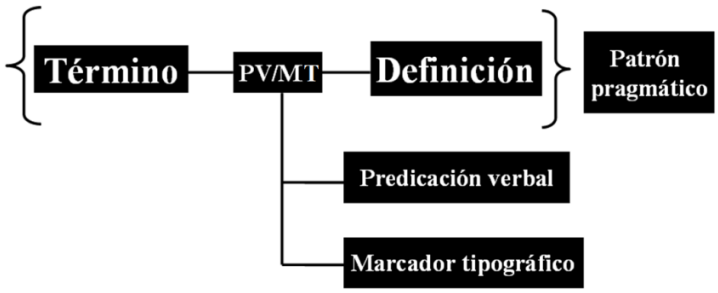

Ilustración 3.2. Jerarquía de los elementos constitutivos de un contexto definitorio (Aguilar 2009)

Como podemos observar, el término y la definición están ligados mediante un elemento que puede ser de dos tipos: el primero de ellos, de índole lingüístico, es la predicación verbal, mientras que el segundo es de índole tipográfica, por lo que se ha dado en llamar un marcador tipográfico, como pueden ser los dos puntos, la coma, punto y coma, etc. Por su parte, el patrón pragmático se encuentra fuera del conjunto término-definición, puesto que un patrón pragmático puede modificar a los elementos que conforman al contexto definitorio, atribuyéndoles valores de autoría, tiempo e instruccionales (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009).

El orden prototípico en español de los elementos de un contexto definitorio es el que se muestra en la ilustración anterior (3.2), esto es, la aparición de un término seguido de un patrón definitorio y una definición:

Término	+	PD	+	Definición
La longitud	+	es	+	el ángulo entre dos planos

Por otro lado, también podemos tener contextos definitorios en los que aparezca un autor de la definición, en este caso, el término pasa a ser objeto directo de un patrón verbal definitorio, y el lugar del sujeto lo ocupa el autor (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009):

Autor	+	PVD	+	Término	+	Definición
Lafourcae (1980)	+	define	+	el perfil profesional	+	como una especificación de…

Empero, pueden presentarse otras estructuras. Por ejemplo se puede presentar un clítico *se* o un verbo auxiliar (*ser*) que modifique la transitividad del verbo o, incluso, puede darse la supresión del autor en una secuencia como la de arriba, fenómeno que es común, sobre todo en las áreas de ingeniería y la informática. Aguilar (2009) encuentra tres tipos de patrones que corresponden con la supresión de la autoría:

Impersonalización. Se refiere a la ya mencionada aparición de un clítico *se* que provoca que el verbo se vuelva impersonal y, por tanto, se elude la presencia del autor. Por ejemplo:

Se conoce como reenganche rápido a la operación de cierre de un interruptor después de una falla[6].

Oraciones en voz pasiva. Se trata, como su nombre indica, de oraciones en voz pasiva, es decir, cuando un argumento interno pasa al lugar de sujeto, presentándose un verbo auxiliar, como puede ser:

A. citrícola es considerado como uno de los posibles vectores de la tristeza.

Oraciones en forma de media voz pasiva. También presenta la aparición de un clítico *se* con una estructura de Sustantivo + *se* + Verbo, al mismo tiempo que el verbo tiene rasgos imperfectivos, por tanto, la media pasiva describe propiedades atemporales de un sujeto (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009). Un ejemplo de esto puede ser el siguiente:

La longitud se define como el ángulo entre dos planos.

Cabe señalar, también, que el tipo de verbo definitorio que funja de núcleo de nuestro contexto definitorio determinará, en gran medida, la estructura de los constituyentes de éste. De tal forma, un verbo como *componer* no permitirá la presencia de un autor, como lo harían verbos como *definir* o *concebir*. Esto mismo se relaciona con el tipo de definición, puesto que una definición extensional será introducida con un verbo como *componer* o *constituir* y no permitirá la presencia de un autor, como veremos en la sección correspondiente a este tipo de definiciones (3.2.2.2).

3.2.1. Término

El término es, en general, aquello sobre lo que se hace una definición; en otras palabras, el término "equivale a la unidad nominal cuya función es designar a la entidad referida por un concepto" (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009). Esta unidad nominal puede estar compuesta por un sustantivo o por un verbo en su forma infinitiva. Un término con un núcleo sustantivo puede encontrarse en el siguiente contexto definitorio:

[6] Todos los ejemplos de contextos definitorios han sido extraídos del CORCODE.

> *Una sustancia infecciosa* es definida como una sustancia que contiene un microorganismo viable, tal como una bacteria, un virus, una rickettsia, un parásito o un hongo, que se sabe o se cree en forma razonable que causa enfermedad en humanos o animales.

Por otro lado, un término con estructura de frase verbal se ejemplifica a continuación:

> *Hacer el diagnóstico* es un paso importante para definir un problema clínico, pero no es un fin en sí mismo.

El término de una definición puede presentar estructuras más complejas, ya sea con expansiones a la derecha o a la izquierda del núcleo; muchas veces, este tipo de ordenación al interior de un sintagma nominal puede dificultar la identificación de un término dentro de éste, lo cual discutiremos más adelante (4.2.3.2).

Por otro lado, también se han considerado términos lingüísticos y no lingüísticos, siendo los primeros palabras o sintagmas mientras los segundos son fórmulas, símbolos, ecuaciones, entre otros.

3.2.2. Definición

La definición se entiende como aquel constituyente que explica a un término; podemos considerarla como la "representación lingüística de un concepto" (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009). Es el elemento nuclear de un contexto definitorio, puesto que en éste se conjunta el conocimiento respectivo a un término.

Sajer y Ndi-Kimbi (1995) caracterizan las definiciones de textos científicos y técnicos a partir de varios patrones que se basan en el esquema de una definición aristotélica de género próximo + diferencia específica. A partir de esto, clasifican dos tipos de patrones generales:

Patrones básicos. Contienen una estructura de “X es un Y + Z”, es decir: término + es un + género próximo + diferencia específica, por lo que, como se observa, su patrón verbal está asociado al verbo *ser*. Como ejemplo tenemos:

> La bomba conocida en la literatura como "Air Lift", o bomba "Mammut", es un sistema utilizado en las estaciones depuradoras para vaciar tanques y extraer aguas o lodos, siempre que la incorporación de aire complementario ofrezca ventajas suplementarias.

Patrones complejos. La estructura de este tipo de patrones es de “X es un Y (que/la cual/el cual/cuyo) + verbo + Z, es decir: término + es un + género próximo + (que/la cual/el cual/cuyo) + verbo + diferencia específica. Un ejemplo de este tipo de patrones puede ser el siguiente:

> El ratón es un dispositivo de entrada que sirve para introducir información gráfica o seleccionar coordenadas (x,y) de una pantalla

En la lexicografía computacional, Vossen y Copestake (1993) conceptualiza a la definición como una descripción lingüística de un término y plantea una estructura general para una definición, que se basa en el modelo aristotélico, que consiste en dos unidades: un género próximo y una diferencia específica.

El género próximo se puede describir como una palabra que engloba al término o, en otras palabras, se trata de un hiperónimo del término. Por otro lado, la diferencia específica es un conjunto de rasgos que distinguen al término de su género próximo; esta diferencia puede darse por medio de la enumeración de las partes específicas del término o de su función en determinado ámbito. En el siguiente ejemplo marcamos los dos constituyentes de una definición:

> Así, un molino de viento es *un artefacto útil* *para captar y aprovechar parte de esta energía*.

En este caso, el género próximo correspondería al sintagma *un artefacto útil*, que podría considerarse hiperónimo de “molino de viento”, mientras que la diferencia específica es

"para captar y aprovechar parte de esta energía". Aquí, la diferencia específica está explicando la función de un molino de viento dentro del conjunto mayor de *artefactos*.

A partir de estos dos elementos, podemos esbozar una tipología de definiciones según se explica en la siguiente ilustración:

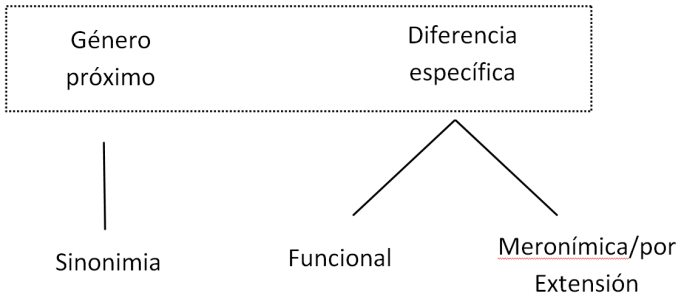

Ilustración 3.3. Tipología de definiciones basada en el modelo aristotélico (Aguilar 2009)

Entendemos que una definición, si bien consta básicamente de un género próximo y una diferencia específica, puede prescindir de alguno de estos elementos. Como se muestra en la ilustración anterior, si sólo se cuenta con un género próximo obtenemos una definición sinonímica. Cuando un contexto definitorio sólo cuenta con la diferencia específica, entonces podemos obtener definiciones de tipo funcionales y extensionales o meronímicas.

A continuación, describimos cada uno de los tipos de definiciones para que sea más clara la concepción de éstas.

3.2.2.1. Definición analítica

Una definición analítica es aquella que cuenta con un género próximo y una diferencia específica. Como se ha mencionado, se trata de una concepción aristotélica que describe un elemento de nivel superior al término y las características que diferencian al término de su género próximo. Para Aguilar (2009), "se da una definición de este tipo cuando el género próximo y la diferencia específica aparecen de manera explícita dentro de una definición". También se puede considerar que este tipo de definiciones denotan el uso, la utilidad o la entidad que representa un término (Hernández 2009). La diferencia específica puede

enumerar características de tipo extensionales o funcionales. Un primer ejemplo de una definición analítica con características extensionales y funcionales es:

> Una rejilla es un dispositivo con aberturas generalmente de tamaño uniforme, utilizado para retener los sólidos de cierto tamaño que arrastra el agua residual.

Este contexto definitorio describe a "Una rejilla" como "un dispositivo", que funciona como el género próximo del término, mientras que "con aberturas generalmente de tamaño uniforme" funge como la diferencia específica con característica extensionales, mientras que por otro lado se introducen características funcionales a partir del verbo en participio *utilizado* más el patrón *para + infinitivo* (Sánchez 2009). Este último patrón también puede introducir funcionalidad sin necesidad de utilizar una forma verbal:

> Para el Gobierno canadiense, el TLC es una plataforma para revitalizar y reorientar su industria de cara a alcanzar la competitividad internacional a partir del contexto continental (América del Norte).

También aquí se nota el género próximo en "una plataforma" y una diferencia específica con características funcionales introducida por el patrón ya mencionado.

3.2.2.2. Definición sinonímica

Una definición de este tipo se caracteriza por contar, únicamente, con un género próximo que denota un sinónimo del término. Para Aguilar (2009) una definición sinonímica se presenta cuando el género próximo está explícito dentro del contexto definitorio y se presenta una equivalencia conceptual entre éste y el término que se está definiendo. Un ejemplo de este tipo de contextos definitorios es:

> El colesterol LDL se conoce como el "colesterol malo" porque es la mayor fuente de adherencia y bloqueo en las arterias, mientras al colesterol HDL se le llama el "colesterol bueno" porque ayuda a eliminar el colesterol extra de su cuerpo.

Verbigracia, en este contexto definitorio existe una similitud conceptual entre el término "colesterol LDL" y el género próximo "colesterol malo".

3.2.2.3. Definición funcional

Una definición funcional es aquella que cuenta con una diferencia específica que explica al término a partir de la descripción del uso o la aplicación de éste. Una definición funcional, en otras palabras, define al término a través de sus funciones en un área o situación determinada. Como ejemplo de este tipo de contextos definitorios tenemos:

> La técnica de velocimetría de imágenes de partícula permite medir la velocidad de un campo de flujo bi o tri dimensional.

En este contexto definitorio, el verbo "permitir" introduce la funcionalidad del término "técnica de velocimetría de particula". Es común en este tipo de definiciones que el elemento que sigue al patrón verbal definitorio sea un verbo en infinitivo. La estructura *para + infinitivo* ha sido estudiada por Sánchez (2009).

3.2.2.4. Definición extensional

Se da una definición del tipo extensional cuando la diferencia específica está explícita en el contexto definitorio y ésta enumera los elementos que componen al objeto del término definido. Es decir, las definiciones de este tipo especifican las partes constituyentes de un término. La enumeración de estos constituyentes puede estar basada en una relación del todo hacia las partes o de las partes hacia el todo (Aguilar, Metodología de Análisis Lingüístico de definiciones en contextos definitorios 2009). Un ejemplo de este tipo de contextos definitorios es el siguiente:

> Para iniciar el estudio de un cortocircuito es necesario primero la preparación del diagrama unifilar de la instalación que muestre la conexión de todas las fuentes de las corrientes de cortocircuito, que ya sabemos son generadores, motores y condensadores síncronos, motores de inducción, conexiones de la red pública, convertidores rotativos y todos los elementos del circuito que se puedan incluir, tales como transformadores, cables, etc.

Es común que las definiciones de este tipo se introduzcan por verbos como "componer", "consistir", "comprender", "constar", y similares; empero, como observamos en este

ejemplo, también pueden introducirse por el verbo “ser”, que generalmente acompaña a definiciones analíticas.

3.2.3. Patrones definitorios

Los patrones o predicaciones definitorias son elementos que permiten relacionar al término con la definición, como se puede apreciar en la ilustración 3.2 (pág. 32). Los más generales son los patrones verbales, pero también se puede presentar la definición de un término por medio de patrones tipográficos o marcas reformulativas. A continuación explicamos de qué se trata cada uno de estos elementos.

3.2.3.1. Patrones verbales definitorios (PVDs)

Un patrón verbal definitorio puede ser concebido como un verbo que funciona como introductor de una definición, ya sea por su valor semántico intrínseco o bien bajo determinados contextos (Hernández 2009). Podemos, a su vez, hablar de patrones verbales simples y compuestos. Un patrón verbal definitorio simple es aquel que sólo cuenta con un verbo definitorio. Por otro lado, uno compuesto se estructura a partir de otras partículas gramaticales que se encuentran dentro del sintagma verbal, tales como clíticos o verbos auxiliares.

Estructuralmente, la principal función de éstos dentro de un contexto definitorio es unir al término con la definición. A partir de estos patrones se pueden extraer automáticamente los contextos definitorios, puesto que los patrones verbales determinan en gran medida el tipo de definición. El siguiente cuadro elaborado por Sierra y Aguilar (2009) muestra la relación existente entre la predicación verbal de un contexto definitorio y el tipo de definición que introducen:

Definición	**Verbos**	**Partículas asociadas**
Analítica (Predicación primaria)	*referir* *representar* *ser* *significar*	*a* (preposición)

Analítica (predicación secundaria)	*caracterizar* *comprender* *concebir* *conocer* *considerar* *definir* *describir* *entender* *identificar* *visualizar*	*como* (adverbio) *por* (preposición)
Sinonímica	*equivaler* *llamar* *nombrar* *ser_igual* *ser_similar*	*también* (adverbio) *a* (preposición) *igual a* (frase adverbial) *similar a* (frase adverbial)
Funcional (Predicación primaria)	*emplearse* *encargar* *funcionar* *ocupar* *permitir* *servir* *usar* *utilizar*	*de* (preposición) *para* (preposición)
Extensional (Predicación primaria)	*componer* *comprender* *consistir* *constar* *contar* *constituir* *contener*	*de* (preposición) *por* (preposición) *con* (preposición)

	incluir *integrar*	

Tabla 3.1. Relación de la tipología de definiciones con los verbos que integran los CDs

La predicación primaria refiere a las estructuras conformadas por la secuencia *término + patrón verbal + definición*, mientras que las predicciones secundarias refieren a la estructura de *autor + patrón verbal + término + definición.* Asimismo, como se observa en el cuadro, el tipo de verbo también puede regir el nexo que introduce la definición. El nexo es una partícula que forma parte del patrón verbal definitorio y se encarga de unir al verbo con la definición. En el caso de los contextos definitorios se habla, principalmente, de dos tipos de nexos: las preposiciones y los adverbios (en la tabla anterior se puede observar qué categoría corresponde a cada uno de los nexos).

3.2.3.2. Marcadores tipográficos y reformulativos

Los marcadores tipográficos pueden tratarse de formas tipográficas como dos puntos, coma, punto y coma, etc. También pueden funcionar como indicadores de un término o de una definición. En este caso hablamos de formas como cursivas, negritas, subrayados, etc. Estas formas pueden funcionar con o sin un patrón verbal definitorio. Por ejemplo, un contexto definitorio con un patrón verbal y un marcador tipográfico es el siguiente:

> Planificación Ambiental puede ser concebida como: El instrumento dirigido a planear y programar el uso del territorio, las actividades productivas, la organización de los asentamientos humanos y el desarrollo de la sociedad, en congruencia con el potencial natural de la tierra, el aprovechamiento sustentable de los recursos naturales y humanos y la protección y calidad del medio ambiente.

Por otro lado, un contexto definitorio que elude el patrón verbal por el uso del marcador tipográfico bien puede ser:

> El efecto cóctel: la capacidad para atender selectivamente a una sola voz entre muchas.

Por otra parte, tenemos los marcadores reformulativos, que son formas que permiten la dar una definición dentro de un contexto definitorio a partir de la introducción de la explicación de un elemento mencionado con anterioridad (Hernández 2009, 16). En el siguiente contexto definitorio podemos encontrar un marcador reformulativo:

> Sin duda, lo más revolucionario en el aspecto médico es la terapia génica, es decir el uso del ADN para intentar corregir algún gen defectuoso.

En este caso el marcador reformulativo es "es decir", el cual introduce una definición al explicar el concepto de "terapia genética".

3.2.4. Patrones pragmáticos

Dentro de un contexto definitorio pueden encontrarse uno o más patrones pragmáticos, que Aguilar (2009) define como aquellos elementos "cuya función es la de ubicar tanto al término y la definición en un contexto dado […], o añadir información complementaria, como la mención al autor o autores de dicho término y su definición, o algún indicio temporal (p. e., en una referencia bibliográfica)". Es decir, un patrón pragmático no cumple una función primordial en la definición de un término, sino que agrega información acerca de la definición; este tipo de información complementaria puede ser bien de tipo autoral, temporal o de otro tipo, como complementos circunstanciales, fechas de definición, etc. Por tanto, los patrones pragmáticos han sido clasificados en tres tipos: temporales, autorales e instruccionales.

Temporales. Los patrones pragmáticos de este tipo, como su nombre lo indica, agregan información temporal a la definición del contexto definitorio. Pueden presentarse como fechas o como complementos que indican temporalidad (como *actualmente*, *en la actualidad*, *antiguamente*, etc.). Un ejemplo de patrón pragmático dentro de un contexto definitorio:

> Ya en los 80s la Estrategia Mundial de la Conservación de la UICN planteó que el desarrollo en relación con la naturaleza se debía concebir como la modificación de la Biosfera y la aplicación de los recursos humanos y financieros, tanto bióticos y

abióticos, a la satisfacción de las necesidades humanas y al mejoramiento de la calidad de vida.

Autorales. Un patrón pragmático del tipo autoral es aquel que indica la pertenencia autoral de una definición. Una de las estructuras de los contextos definitorios que plantea Aguilar (2009) es *autor + PVD + término + definición*, donde la aparición del autor se considera como un patrón pragmático que indica quién es el autor de la definición que se encuentra en el contexto definitorio. En este caso, un ejemplo es el siguiente:

> Carlos Godino en su libro "Teoría del Buque y sus aplicaciones", define la Arquitectura Naval como la ciencia que trata de los conocimientos necesarios para la construcción de los buques , dividiendo estos conocimientos en dos grandes grupos: La "Construcción Naval" que comprende el estudio de las condiciones que deben satisfacer los barcos, desde el punto de vista constructivo, y determina la forma y espesores que deben tener sus diferentes partes, según los esfuerzos a que han de estar sometidos y las condiciones a que deben satisfacer en cada caso.

Instruccionales. Un patrón pragmático de este tipo introduce información pragmática adicional al tiempo y el autor a un contexto definitorio; este tipo de información puede indicar diferentes características semánticas atribuidas a la definición; por ejemplo, puede indicar pertenencia a una materia u otras características. Un patrón pragmático de este tipo se presenta en el siguiente contexto definitorio:

> Entendemos aquí el término visión en el sentido schumpetenano como acto cognoscitivo preanalítico que nos hace ver las cosas bajo una luz cuya fuente no se encuentra en los hechos.

3.3. Usos de los contextos definitorios

Los contextos definitorios, en su naturaleza de fragmentos textuales que contienen un término definido, pueden ser de gran utilidad para la terminología y la lexicografía. Como ya comentaban Sierra *et al* (2006), un corpus que contenga Contextos Definitorios funge como una valiosa herramienta para estas dos áreas académicas. A través de este tipo de

herramientas la lexicografía y la terminología se enriquece de definiciones que, por sus características, presentan diferentes tipos de perspectivas a un mismo término.

Los contextos definitorios proveen un amplio campo de definiciones para diferentes ámbitos de especialidad que, de otra forma, serían de difícil acceso incluso para el especialista en terminología, debido a su proceso de extracción que consiste en la búsqueda de las definiciones en distintos textos especializados en diferentes áreas de conocimiento. Así, actualmente se cuenta —como ya se ha mencionado— con contextos definitorios del área de medicina, computación, matemática, política, física, medicina, entre otras. A partir de los contextos definitorios, por tanto, se puede obtener información terminológica sobre términos de estas áreas de especialidad.

Con los contextos definitorios, la terminología y lexicografía pueden obtener la información que necesitan a partir de la extracción automática. De igual forma, otras áreas pueden verse beneficiadas por herramientas de este tipo. Los contextos definitorios pueden servir para la construcción de bancos terminológicos, la elaboración de diccionarios, para el análisis de relaciones léxicas y para la construcción de ontologías, además de que pueden ser de utilidad para cualquier persona que esté interesada en alguna de estas áreas del conocimiento.

Por tanto, el corpus de contextos definitorios, al agrupar una buena cantidad de éstos, facilita la consulta de términos y definiciones especializadas en las áreas ya mencionadas y permite que su búsqueda sea más sencilla y accesible a todo tipo de público interesado.

4. Metodología y diseño de un corpus para contextos definitorios

Una vez analizados los conceptos de corpus y de contextos definitorios, en este capítulo explicaremos el proceso que se siguió para la elaboración de este corpus. Explicaremos, en principio, los objetivos que se persiguieron, así como las características que se buscaban para el corpus; finalizado esto, puntualizaremos cómo se obtuvieron los contextos definitorios y su proceso de selección y etiquetado para su incorporación a la interfaz web.

4.1. Diseño del corpus

En primera instancia, planteamos los criterios bajo los cuales se diseñó el corpus de contextos definitorios. Tomamos en cuenta los objetivos que se perseguían dentro de la investigación y las características que se decidieron para que estos objetivos fueran alcanzados.

4.1.1. Objetivos

Los objetivos con los que se planeó el corpus fueron los que a continuación se enumeran:

a) Desarrollar un corpus que contenga los contextos definitorios que han surgido y surjan dentro del marco de investigaciones realizadas en el GIL.
b) Que este corpus se complemente con el ECODE para que, a partir del corpus, se haga más investigación sobre la extracción de contextos definitorios y que a su vez los contextos definitorios extraídos pasen a formar parte del corpus.
c) Que el corpus sea accesible para todo público y fácil de consultar a través de internet.
d) Que sea una herramienta útil para especialistas y que en éste se reúna información especializada en diversas áreas del conocimiento.

4.1.2. Características

Cada corpus cuenta con características específicas exigidas por los propósitos y objetivos por los que fueron concebidos; en nuestro caso —una vez expuestos los objetivos— nos basamos, principalmente, en los criterios ofrecidos por Sierra y Rosas (2009) para caracterizar el corpus de contextos definitorios. Dichos criterios se basan en el formato en que se encuentra el corpus, en el tipo de textos que recoge, en su estructura y en los criterios con que se concibe éste.

Uno de los primeros criterios que podemos tomar en cuenta, propuesto por Torruella y Llisterri (1999), es s*egún su formato*; a este respecto, se trata de un corpus informatizado, puesto que los elementos integrantes se presentan en un formato electrónico, codificados con un estándar XML, y seleccionados bajo criterios determinados. Con estos factores, el corpus tiene un tratamiento computacional que permite realizar búsquedas y determinar factores lingüísticos con mayor facilidad que un corpus que no utilice métodos informáticos, lo que se resume en una gran ventaja sobre todo si se trata de corpus con gran cantidad de texto.

Según el origen de sus elementos. Por tratarse únicamente de documentos escritos, nuestro corpus se considera dentro de la categoría de corpus textual.

Según su codificación o anotación. Se trata de un corpus anotado. Se utilizaron etiquetas en el estándar XML. La anotación que utilizamos para este corpus tiene características de lo que en la tabla 2.1 (p. 29) hemos llamado etiquetado semántico, puesto que éstas marcan los constituyentes como el término, el patrón definitorio y la definición, por lo que, más específicamente, podríamos llamarlo un etiquetado de relaciones semánticas. Como mencionamos, el etiquetado semántico no tiene un estándar bien definido, no se hizo caso ni a rasgos morfológicos, relaciones sintácticas, ortográficas, ni otras, sino que nos basamos en la teoría sobre contextos definitorios desarrollada dentro del mismo GIL para determinar la anotación que se uso en el primer corpus y, posteriormente, para proponer una nueva.

Según la especificidad de sus elementos. Estamos hablando de un corpus especializado en diferentes ramas; los textos que componen el corpus forman parte de documentos de áreas

de especialidad y, por tanto, son fragmentos de lenguaje especializado; se extrajeron de distintas fuentes (4.2.1) que permiten la diversidad de temas de especialidad, que abarcan áreas como la de medicina, biología, derecho, informática, matemática, etcétera. Específicamente, se trata de un corpus especializado conformado por *textos informativos,* excluyendo, dentro de éstos, los periodísticos; se conforma, entonces, de documentos científicos, académicos y técnicos.

Según su temporalidad. Se trata de un corpus sincrónico, compuesto de textos actuales, por lo que podemos decir, también, que es contemporáneo.

Según su propósito. Los propósitos del corpus de contextos definitorios no son otros que ser una herramienta de consulta y de trabajo tanto para el público más especializado (lexicógrafos, terminólogos y, en general, lingüistas) como para el usuario común. Si bien surge de trabajos de investigación que perseguían un propósito específico, el corpus de contextos definitorios es un corpus multipropósito, que no surge con el objetivo de arrojar datos lingüísticos precisos y que puede consultar cualquier tipo de persona; por tal motivo, podemos clasificarlo como un corpus de referencia.

Según el lenguaje de sus muestras. Es un corpus monolingüe, puesto que únicamente contiene contextos definitorios en español. Aunque, para un futuro, se planea incluir textos en inglés.

Según la cantidad de texto que se recoge. Afirman Sierra y Rosas (2009) que un corpus grande debe contar con un aproximado de diez millones de palabras; en nuestro caso, el corpus cuenta con menos de cien mil palabras, por tanto podemos considerarlo un corpus pequeño.

Según la distribución de sus elementos. Para la construcción del corpus no se tomaron en cuenta la proporción de los textos que se recogieron, por lo que en principio no podemos hablar de un corpus equilibrado ni piramidal sino, más bien, de un corpus desequilibrado, puesto que las proporciones de los contextos definitorios que se incluyeron no contienen ningún criterio con respecto a la proporción que ocuparían en éste.

Según su documentación. Podemos decir que se trata de un corpus no documentado, puesto que por medio de la interfaz no se puede consultar una referencia específica que indique de dónde se extrajo cada contexto definitorio; aunque se sabe de dónde se extrajo cada uno (véase 4.2.1), no se cuenta con los documentos completos y la interfaz no cuenta con una opción que envíe a éstos.

Según la accesibilidad. Como ya lo hemos mencionado en otro punto, el corpus de contextos definitorios está disponible para todo tipo de usuarios. Se trata, entonces, de un corpus público, que permite su consulta en línea a través de una interfaz en internet.

Según su autoría. Por tratarse de un corpus que recoge muestras sin tomar en cuenta más que se trate de contextos definitorios, hablamos de un corpus de autoría variada.

4.2. Metodología de elaboración del corpus

Establecidos los objetivos y las características del corpus, el siguiente paso para la elaboración de éste fue la recolección de los contextos definitorios a partir de diversas fuentes. Una vez recolectados los contextos definitorios se prosiguió con su etiquetado por medio del estándar XML. Estos procesos serán explicados a continuación.

4.2.1. Fuentes de extracción

Con respecto a los derechos de autor del corpus, cabe señalar que los contextos definitorios que lo integran actualmente fueron obtenidos a partir de otras investigaciones que se han realizado en el Grupo de Ingeniería Lingüística y no de las fuentes originales directamente, por lo que no se tienen documentos de autorización de uso de obras, sino que se ha estipulado presentar sólo los fragmentos de éstas donde se encuentra cada uno de los contextos definitorios y mostrar las fuentes de extracción de donde se han obtenido. Éstos fueron extraídos de diversos corpus y por medio de métodos de extracción diversos. Las fuentes originales de los contextos definitorios, así como las investigaciones dentro del GIL en que fueron utilizados, se enumeran a continuación:

- El Corpus Lingüístico de Ingeniería o CLI (Medina *et al*, 2004). Se trata de un corpus en español orientado hacia el área de ingeniería y desarrollado por el Grupo

de Ingeniería Lingüística (GIL). Está conformado por documentos en texto plano (extensión *.txt*). Se trata de un corpus que reúne textos especializados del área de ingeniería, tales como tesis, artículos, informes, etcétera (utilizado en Sierra *et al*, 2006; Aguilar, 2009).

- El Corpus Técnico del Instituto Universitario de Lingüística Aplicada (IULA) de la Universidad Pompeu Fabra en Barcelona (Cabré y Vivaldi, 1997). Este corpus cuenta con 9 542 000 palabras en su sección dedicada al español, a la cual se puede acceder a través de su herramienta de búsqueda *BwanaNet*, que cuenta con distintas opciones de búsqueda, estas son: básica, estándar y compleja (utilizado en Sierra *et al*, 2006; Hernández, 2009; Alarcón, 2003).
- El Corpus Informático en Español o CIE (L'Homme y Drouin, 2006). Se trata de un corpus técnico desarrollado para las áreas de informática y ciencias de la computación con miras a la creación e implementación de un diccionario electrónico en español. Cuenta con alrededor de 500 000 palabras, divididas en cuatro sub-corpus, estos son: de la revista *PC World Latinoamérica* (PCWLAF), revista *Guía Computación*, *WindowsTI Magazine*, y entradas obtenidas de la Wikipedia en español (utilizado en Aguilar, 2009).
- *Sketch Engine* (Kilgarriff, 2003). Esta herramienta cuenta con 116 900 060 palabras en el idioma español (utilizado en Hernández, 2009; Sánchez, 2009).
- *Google*. Una última fuente de nuestro corpus fue la herramienta de *Google* en español, que es un motor de búsqueda en internet (utilizado en Hernández, 2009).

4.2.2. Proceso de recolección de los contextos definitorios

Los procesos que se siguieron para la extracción utilizaron la metodología del ECODE (Alarcón 2009) y búsquedas manuales a partir de patrones definitorios. A partir de la búsqueda automática sólo se obtuvieron candidatos. No se podía asegurar que todos los candidatos extraídos de esta forma fueran contextos definitorios, y aunque se tratara de ello, muchas veces no contaban con una estructura completa, faltando varios elementos o presentando otros problemas.

Una vez extraídos los contextos definitorios, se prosiguió a reunirlos en un solo archivo de texto plano para que fuera más sencilla su manipulación, ya que éstos se obtuvieron a partir de archivos de *Excel* y *Word*, ya que así se habían manejado en anteriores investigaciones. También se eliminaron algunas etiquetas con las que el ECODE marca el término y la definición, pero que no corresponden a las utilizadas por el corpus.

4.2.2.1. Resultados preliminares

En un principio, los posibles contextos definitorios que se obtuvieron se clasificaron de la siguiente manera según su fuente de extracción:

Fuente	**Número de contextos definitorios**
CLI	238
IULA	1361
CIE	562
Sketch Engine	5
Google	49
Total	2215

Tabla 4.1. Cantidad de contextos definitorios extraídos por fuente

Como se puede observar, en esta etapa se contaba con un total de 2215 candidatos a contextos definitorios, la mayoría procedente del corpus técnico del IULA, mientras que de *Sketch Engine* tan sólo se extrajeron 5, siendo la fuente con menor número de contextos definitorios extraídos.

De estos 2215 candidatos se hizo un análisis manual preliminar para delimitar el patrón definitorio que constituía cada uno de ellos. Los lemas verbales y otros patrones que se obtuvieron se muestran en la siguiente tabla:

Lema del verbo	Cantidad de contextos definitorios
Caracterizar	18
Componer	7
Comprender	8
Concebir	106
Conocer	16
Considerar	36
Consistir	24
Constar	10
Constituir	15
Contar	4
Contener	6
Corresponder	2
Definir	273
Denominar	4
Describir	7
Designar	1
Determinar	3
Emplear	12
Encargar	2
Entender	214
Formar	1
Funcionar	1
Identificar	182
Incluir	26
Integrar	2
Interpretar	6
Llamar	26
Permitir	130
Referir	15
Representar	11
Ser	958
Servir	8
Significar	6
Tener	4
Usar	6
Utilizar	18
Ver	1
Visualizar	2
Marcadores referencial definitorio	4
Marcadores tipográficos definitorios	13
Sin determinar	26

Tabla 4.2. Número total de contextos definitorios por patrones definitorios

El lema del verbo fue un factor importante en la selección de los contextos definitorios y en su posterior análisis, ya que éste puede indicar el tipo de definición; además, sirvió para una primera evaluación. Como se puede observar, se cuenta con 26 candidatos que tienen un patrón definitorio indeterminado, es decir que no contaban con uno o no introducía una

definición; a partir de esto, se determinaron como no candidatos a contextos definitorios y fueron descartados del corpus. Aquellos candidatos que contaban con un verbo definitorio o un marcador textual o referencial se mantuvieron; sin embargo, esto no garantizó que se tratara, efectivamente, de contextos definitorios.

Así, de los 2215 candidatos obtenidos en un principio, no todos se pudieron considerar como contextos definitorios, debido principalmente a que se extrajeron mediante un proceso automático que todavía se encuentra en prueba; por tal motivo fue necesario realizar un proceso de revisión manual para poder evaluar los resultados y obtener una nueva lista que contenga mejores candidatos a contextos definitorios.

4.2.2.2. Determinación de candidatos

Para realizar la revisión de los resultados preliminares y descartar aquellos candidatos que no se incluirían en el corpus, se siguieron cuatro procesos: primero, la revisión de las muestras del corpus para eliminar aquellos que se repitieran; segundo, un análisis manual de cada uno de los candidatos extraídos para descartar aquellos que no fueran contextos definitorios; tercero, identificación de los elementos constitutivos de cada uno de ellos; por último, una clasificación tanto por el tipo de definición como por el lema verbal.

Candidatos que se repetían en el corpus. Se tuvo que descartar, en primera instancia, aquellas muestras que se repitieran. Para esto se implementó un proceso semiautomático que consistió en dos pasos:

a) La ejecución en el corpus de un programa informático que consiste en encontrar coincidencias en varias cadenas de caracteres; de esta forma, aquellos contextos definitorios con coincidencias fueron agrupados para ser revisados manualmente y descartados si, en efecto, se trataba de contextos definitorios repetidos.
b) Como no todos los contextos definitorios que se repetían fueron encontrados por el procesamiento computacional, se procedió a hacer una revisión manual del corpus para eliminar los que no fueron reconocidos por el método informático. Algunos de los factores que produjeron que el programa computacional no encontrara candidatos repetidos fueron desde pequeños cambios en la secuencia de caracteres

como un espacio en blanco de más (lo que impedía la coincidencia exacta de los caracteres) hasta que el contexto definitorio haya sido extraído con mayor cantidad de texto que otro, aunque se tratara del mismo.

Análisis manual del corpus para descartar aquellos candidatos que no fueran contextos definitorios. Una vez eliminados aquellos candidatos repetidos, se realizó una revisión manual de cada uno de los contextos definitorios que conforman el corpus, para, en primer lugar, evaluar su calidad como contextos definitorios; posteriormente, eliminar aquellos candidatos que no fueran contextos definitorios; y por último determinar el tipo de verbo y de definición;

Aguilar *et al* (2006) proponen que para delimitar qué clase de predicaciones verbales son las que tienen una mayor posibilidad de estar ligadas a una definición debe considerarse:

- Que estén conjugadas en tercera persona del singular y plural (define, caracteriza, es/son, etcétera).
- Que sean participios y gerundios, los cuales estarían asociados a pretéritos perfectos (Se ha definido como) o a formas en pasiva (es identificado como, está siendo interpretado como).
- Que se trate de infinitivos ligados a perífrasis verbales con el auxiliar poder (se puede considerar, se puede definir, etc.).

Con base en estos puntos, se comenzó el proceso de clasificación, que consistió en determinar qué tan posible era que se tratara de contextos definitorios. Se descartó aquellos qué no cumplieran con tales consideraciones y aquellos que tuvieran un verbo que no fuera definitorio o en una conjugación que no permitiera una definición. También se eliminaron aquellos que no contaran con un término o que se tratara de términos anafóricos (como pronombres). De igual forma, se rechazaron los que no contaran con una definición. Hubo casos en que se presentaban elementos a la izquierda y derecha del verbo; sin embargo, no correspondían a contextos definitorios estrictamente. Por ejemplo:

> Hay otra manipulación, más sutil, que es mostrar *lugares donde hay hambre, no las zonas de la pobreza, porque los primeros son pequeñas colonias cuyo problema se*

> *puede solucionar enviando aviones con alimentos y ya nos quedamos satisfechos*; en cambio, la pobreza es más universal, no sabemos cómo arreglarlo, plantea interrogantes centrales sobre el sistema y angustias irresueltas.

En este caso, estructuralmente se puede decir que se presenta un patrón verbal “es” y una definición que va desde “mostrar” hasta “satisfechos”, pero el sentido que describe este fragmento no es lo descarta como candidato a contexto definitorio.

Identificación de los elementos constitutivos de los contextos definitorios. Se buscó que cada contexto definitorio del corpus contará con los elementos de término, definición y patrón definitorio (3.2). Si no se contaba con uno de estos elementos, se descartaba al candidato.

Clasificación tanto por el tipo de definición como por el lema verbal. Para el proceso de clasificación de los contextos definitorios obtenidos, se hizo una revisión que primeramente fue manual y posteriormente automática; estas revisiones consistieron, primero, en marcar cada uno de los contextos definitorios con su lema verbal y su tipo de definición; posteriormente se aplicó de un programa informático que identificaba (gracias al marcaje) y ordenaba cada uno de los contextos por su lema verbal y su tipo de definición. De esta forma, se obtuvieron dos listas que agrupaban los contextos definitorios según su definición y según el lema de su verbo definitorio, aunque finalmente se decidió mostrarlos en la interfaz ordenados por lemas y no por definiciones.

4.2.3. Proceso de etiquetado

Una vez descartados los candidatos que no se consideraron como contextos definitorios, así como las muestras que se repetían, se prosiguió con el etiquetado.

Como hemos mencionado, el etiquetado permite que las búsquedas al interior del corpus y su análisis sean más sencillos de realizar, pues ayuda a que la computadora interprete el contenido textual de los documentos. Con respecto al etiquetado, Arrarte (199922) lo entiende como:

marcas textuales con las cuales señalamos determinados elementos del documento e indicamos al sistema las funciones de procesamiento que debe llevar a cabo con cada uno de ellos. Podemos definir estas marcas textuales como texto añadido al contenido de un documento con información sobre el mismo.

El proceso de etiquetado llevado a cabo en nuestro corpus se realizó, primero, llevando a cabo una selección de las etiquetas que se iban a utilizar y, segundo, mediante la implementación de estas etiquetas en el corpus.

4.2.3.1. Determinación de etiquetas utilizadas

Para el corpus, se utilizó el estándar de anotación XML (2.2.1), puesto que éste permite definir las etiquetas. Las etiquetas XML que utilizamos para el corpus delimitan a cada contexto definitorio de forma integral, así como los elementos que los constituyen.

En primera instancia, el corpus tiene que mostrar un encabezado que ayude a identificar ciertas características de éste. Sierra *et al* (2006) proponen:

Etiquetas	**Función**
Cabeza	Contiene la información del documento (nombre, tipo de verbo, fuente, fecha, recopilador, etcétera).
Fuente	Indica la fuente original del documento (CLI, IULA, CIE, *Google*, Sketch Engine).
Fecha	Indica la fecha del recopilado y del etiquetado del documento.
Nombre	Contiene el nombre de la recopilación hallada en el documento, como puede ser “verbo definir”.
Verbo	Muestra el nombre del verbo definitorio cuando el documento sólo analiza uno.
Tipo	Se indica si el criterio de clasificación del documento es la *definición*. Estas pueden ser: analítica, funcional, sinonímica, instruccional.
Recopilador	Muestra el nombre de la persona que recopiló el documento.

Tabla 4.3. Encabezado del documento XML (Sierra, Alarcón y Aguilar, y otros 2006)

El encabezado ayuda a identificar tanto los factores metatextuales del corpus, como su filiación y las características de cada archivo XML. Los contextos definitorios que integran el corpus, por otra parte, están contenidos en el cuerpo del documento. Para cada elemento de un contexto definitorio se determinó una etiqueta que puede contener determinados atributos que especifiquen al constituyente. Por ejemplo, se especifica qué tipo de nexo es el que se utiliza o qué tipo de término, así como la estructura de éste. En la tabla 4.4 presentamos las etiquetas básicas que se utilizaron en esta tesis para el etiquetado.

Etiqueta	**Significado**	**Función**
CD	Contexto definitorio	Marca a todo el contexto definitorio, dentro de esta etiqueta se incluyen todas las que se enumeran a continuación.
TERM	Término	Abarca el término que se va a definir.
DEF	Definición	Indica la definición. En ella se debe omitir cualquier texto complementario que de manera estricta no forme parte de dicha definición. Existen 5 tipos de definiciones (véase el apartado dedicado a tipología de definiciones).
PVD	Predicación Verbal Definitoria	Contiene todos los componentes de una PVD: clítico *se*, verbo auxiliar, verbo definitorio y nexo.
VD	Verbo Definitorio	Marca el verbo que une al término con su respectiva definición.
SEmarc	Clítico *se*	Se marca el clítico *se* cuando acompaña a un verbo definitorio.
VAUX	Verbo Auxiliar	Contiene cualquier verbo auxiliar dentro del PVD (por ejemplo, *se* ***puede*** *considerar como, se* ***ha*** *definido, se* ***debe*** *concebir como,* etc.)
NX	Nexo	Señala la función que cumple un adverbio o preposición entre el verbo y la definición.
MRD	Marcadores	Abarcan estructuras sintácticas con la función de explicar el

	Reformulativos Definitorios	propio lenguaje. Son frases que retoman algún elemento discursivo para reintroducirlo al discurso de otra forma; algunos ejemplos pueden ser las siguientes estructuras: *es decir, por ejemplo, esto es*, etc.
MTD	Marcadores Tipográficos Definitorios	Señala cualquier marca tipográfica que sirva como introducción o indicador de una definición. Se distingue en dos tipos: a) Marcadores definitorios (mdef): unen a un término con su definición con alguna marca de puntuación o un cambio de línea, sustituyendo o complementando la función del PVD. b) Marcadores tipográficos (mt): indicación de negritas, cursivas, subrayado y otras marcas que dan prominencia al término definido o a la definición.
PP	Patrones Pragmáticos	Dan información sobre el uso de los términos.

Tabla 4.4. Etiquetas utilizadas en el corpus

Por otra parte, cada uno de los constituyentes debe ser determinado con características que especifiquen sus valores, estructuras, etc. Por ejemplo, se debe determinar el tipo de término y su estructura, también se debe especificar el lema verbal y otros valores en cada uno de los elementos. Determinar los constituyentes de los contextos definitorios permite que en el corpus final se puedan realizar búsquedas más específicas para los usuarios. Por tanto, se deben agregar atributos a las etiquetas utilizadas. Los atributos que pueden contener cada una de las etiquetas, así como los elementos que los determinan, se presentan a continuación:

Etiqueta	Atributo	Descripción del atributo
CD	num	Indica el número del CD etiquetado.
TERM	tipterm	Indica el tipo de término y tiene los valores: L = Un término lingüístico, construido por un lexema. NL = Término no lingüístico, como números, fórmulas matemáticas o químicas, etc.
	stterm	Indica la estructura sintáctica del término. Tiene los valores: fn = Frase nominal fn-fp = Frase nominal seguida de una frase preposicional fv-fn = Una frase nominal (artículo más verbo en infinitivo) seguida de una frase nominal.
DEF	tipdef	Indica el tipo de definición según la tipología de definiciones (3.2.2). Tienes los valores: GD = Definición analítica (género próximo y diferencia específica) EXT = Definición extensional FUN = Definición funcional Sin = Definición sinonímica
VD	lema	Indica el lema del verbo definitorio, es decir, su forma en infinitivo.
NX	tipnx	Indica el tipo de nexo que introduce la definición. Tiene los valores: adv = Un nexo adverbial (En este caso, *como* se considera adverbio) prep = Una preposición
MTD	mdef	Se usa cuando se trata de un marcador definitorio (marcadores discursivos que introducen una definición). Pueden ser: dp = dos puntos

		viñ = viñetas par = paréntesis gui = guiones dl = comillas
	mt	Refiere a un marcador textual (marcas tipográficas en el texto). Tiene los valores: curs = Cursivas neg = Negritas subr = Subrayado otr = Cualquier otro tipo de marcador tipográfico definitorio.
PP	tipp	Indica el tipo de patrón pragmático. Tiene los valores: Aut = Indica el autor de una definición. Temp = Indica fechas en que se definió, puede incluir valores temporales como *actualmente, anteriormente, ahora*, etc. Inst = Indica una instrucción que se da en la definición, es decir, alguna frase que modifique el sentido o valor de la definición, como *también, en el campo de XX, en tal ciencia*, etc.

Tabla 4.5. Atributos de las etiquetas en el corpus

Las etiquetas y los atributos que le corresponden a cada una de éstas determinan los elementos constitutivos de un contexto definitorio, así como las características de éstos, y por tanto, describen su estructura. Cabe resaltar que los elementos de un contexto definitorio pueden encontrarse dentro de otro, como una especie de subelementos. Las etiquetas también pueden anidarse. Es común, por ejemplo, que el patrón verbal definitorio contenga otros constituyentes subordinados dentro de sí mismo; prototípicamente, éste se compone por el verbo definitorio, verbo auxiliar, clítico *se* y nexo; por tanto, la estructura del documento XML, y del corpus en general, así como la de un contexto definitorio, se puede representar de la siguiente forma:

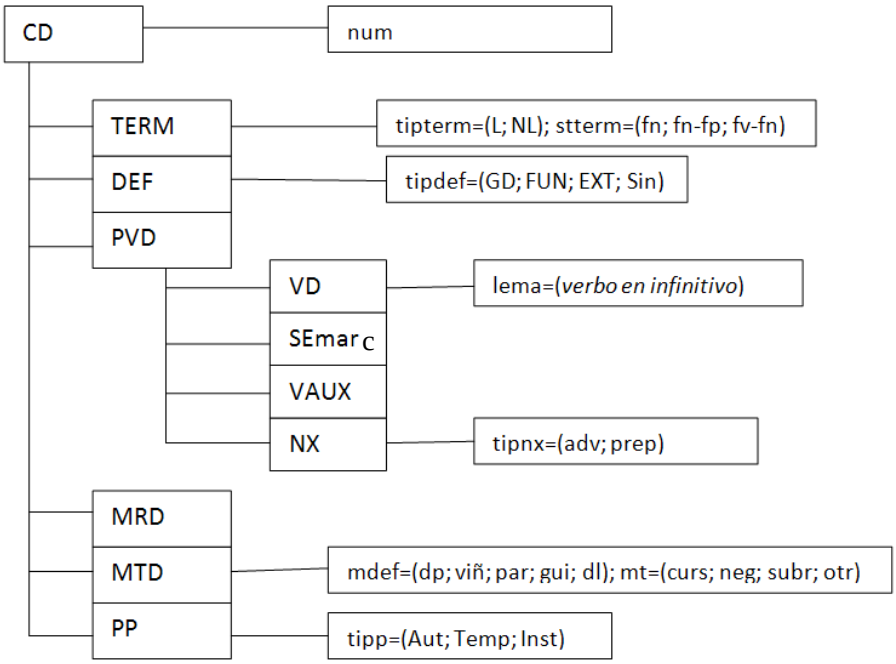

Ilustración 4.1. Estructura de un contexto definitorio dentro del etiquetado del corpus

El elemento más alto dentro de esta jerarquía está representado por la etiqueta CD y representa al contexto definitorio; por tanto, dentro de ésta se encuentran todos los constituyentes posibles que pueden presentarse dentro de un contexto definitorio. Las etiquetas para el término (TERM), la definición (DEF), el marcador referencial definitorio (MRD), el marcador tipográfico definitorio (MTD) y el patrón pragmático (PP) se encuentran al mismo nivel y únicamente contienen sus atributos específicos. El patrón verbal definitorio (PVD), por su parte, incluye a otros elementos que pueden aparecer dentro de esta etiqueta, aunque PVD se encuentra al mismo nivel que los constituyentes antes mencionados.

4.2.3.2. Implementación del etiquetado

Para llevar a cabo el proceso de etiquetado se utilizó el editor *XML Writer*. Se trata de un programa informático que permite una visualización más clara de las partes que integran el documento, es decir, muestra con claridad la diferencia entre el texto y las etiquetas, así como de los atributos de éstas. También permite que se puedan validar las etiquetas del

documento para que éstas sean correctas; si existe algún error en una etiqueta el editor nos avisará. De igual forma, facilita el proceso de etiquetado, al hacer el marcaje de una forma semiautomática, gracias a la implementación de un Esquema XML, el cual contiene las definiciones de las etiquetas y los atributos; de esta forma, el proceso se hace más rápido que si se realizara sin un Esquema XML.

Así, se etiquetaron los 1427 contextos definitorios que conforman actualmente el corpus. A continuación, se presenta la metodología llevada a cabo para el etiquetado de cada uno de los constituyentes de los contextos definitorios que integran el corpus.

Etiquetado del contexto definitorio. Para englobar todo el contexto definitorio se utilizó la etiqueta de CD, como ya se ha indicado. Todos los contextos definitorios resultantes de los procesos anteriores fueron etiquetados de esta forma; por tanto, no se presentaron mayores complicaciones. Una muestra de este etiquetado se presenta a continuación:

> <CD num="1">En la escuela aprendimos que “la energía es la capacidad que tiene un sistema para producir trabajo”</CD>

El atributo de número es determinado por el lugar que ocupa el contexto definitorio dentro del corpus, y permite llevar un control de la cantidad de éstos.

Etiquetado del término. La etiqueta de TERM agrupa a todo aquel elemento que se considere el término de una definición, según los parámetros expuestos en 3.2.1. Con respecto a su estructura, puede corresponder a una o más palabras, ya sea en forma de un sintagma nominal con un núcleo sustantivo (seguido o no de un sintagma preposicional) o de un verbo en infinitivo. Por otro lado, se considera también si se trata de un valor numérico o simbólico; este último factor determina si se trata de un término lingüístico o de uno no lingüístico, el primero corresponde a un sustantivo o verbo, mientras que el no lingüístico responde a números o símbolos; por ejemplo, el sintagma “la energía” es un término lingüístico, por lo que se etiqueta:

> <TERM tipterm=”L”>la energía</TERM>

Sin embargo, si el término de una definición es, verbigracia, una fórmula matemática, se trata de un tipo no lingüístico, por lo tanto el valor de su atributo cambia:

<TERM tipterm=”NL”>d=√(x-y)^2</TERM>.

Acerca de la estructura del término, ésta depende de si el término es lingüístico o no. Un término no lingüístico no contiene el atributo de estructura. En contraste, un término lingüístico puede presentar distintas estructuras sintácticas que se marcan en el etiquetado. Se puede presentar un sintagma nominal como el del ejemplo anterior, cuyo etiquetado correspondería al siguiente:

<TERM tipterm=“L” stterm=”fn”>la energía</TERM>

Empero, al núcleo nominal se le pueden agregar complementos tanto a la izquierda como a la derecha. En nuestro caso, son de mayor relevancia los que se encuentran del lado derecho, puesto que pueden tratarse de complementos introducidos por preposiciones. Cuando se trata de uno o más adjetivos, el etiquetado se conserva tal como el ejemplo anterior (stterm=”fn”), pero al tratarse de sintagmas preposicionales el atributo cambia (fn-fp); por lo tanto, se encontrará una etiqueta similar a la siguiente:

<TERM tippterm=”L” stterm=”fn-fp”> La jerarquía de dependencias</TERM>

Esto indica que pueden presentarse uno o más sintagmas preposicionales a la derecha del núcleo nominal.

Otra variante que presenta la estructura del término es la aparición de un verbo en infinitivo acompañado de un sintagma nominal (con o sin complementos), por ejemplo:

<TERM tipterm="L" stterm="fv-fn">Desactivar la casilla «Sólo lectura»</TERM>

Un problema común en la determinación del término es encontrar los límites precisos de éste. Como hemos visto, se puede presentar la estructura de un sintagma nominal seguido de uno o más sintagmas preposicionales. Se tiene que determinar, entonces, si el término corresponde al sintagma completo (SN + SP) o sólo al núcleo del sintagma preposicional.

En estos casos debe analizarse la definición y determinar cuál es el término según la relación que se establezca entre estos dos elementos. Un ejemplo claro de una frase preposicional cuyo término se encuentra después de la preposición se puede apreciar en el siguiente contexto definitorio:

> La presente Ley respeta el punto de vista clásico sobre la misión del <TERM tipterm="L" stterm="fn">Registro Civil</TERM>, concebido como instrumento para la constancia oficial de la existencia, estado civil y condición de las personas.

Como se observa, el sintagma de "el punto de vista clásico sobre la misión del Registro Civil" no guarda relación con respecto a la definición, pero sí "Registro Civil", puesto que éste es el que se define como "instrumento para la constancia oficial de la existencia, estado civil y condición de las personas".

Etiquetado del patrón verbal definitorio. La etiqueta de PVD agrupa a uno o más elementos, cuyo núcleo es una forma verbal, conjugada, infinitiva, en participio o en gerundio, que une al término con la definición. Este verbo comúnmente va acompañado de un nexo (sea éste preposición o adverbio) y puede tener uno o varios verbos auxiliares, tales como *ser, haber, poder*, etc. Asimismo, puede ir acompañado por un clítico *se*, cuando la estructura del contexto definitorio no es "X define Y como Z", sino: "*se* define Y como Z". Esto es común para varios verbos definitorios que introducen definiciones analíticas o sinonímicas (se define, se conoce, se describe, etc.), extensionales (se compone) y funcionales (se usa, se utiliza).

Para verbos con valores funcionales no es común que se presente un sujeto antes del término; sin embargo, es posible que se dé la estructura: "X utiliza Y para Z", aunque no se presentó ningún caso en el corpus. Ahora bien, para verbos que introducen valores extensionales, no puede presentarse una estructura como "X + verbo definitorio + Y + nexo + Z", pues no se puede decir que "X compone a Y de Z", sino que necesariamente tendrá que ser "Y se compone de Z". Como puede observarse, la estructura del contexto definitorio está determinada en gran medida por su patrón verbal. Verbos como *ser* o *servir* no permiten la inclusión de un clítico *se* y el sujeto de su sintagma oracional siempre será el término. El verbo *ser* tampoco permite la inclusión de nexos ni de verbos auxiliares (sin

ningún caso en el corpus), por lo que su estructura siempre será “X es Y”, introduciendo definiciones analíticas o sinonímicas.

Los nexos también están determinados en gran medida por el verbo definitorio (Tabla 3.1). Para verbos que introducen definiciones analíticas suele aparecer *como*, para extensionales *por* o *de*, y para funcionales *para*.

Los verbos auxiliares introducen valores semánticos determinados al verbo definitorio, por ejemplo, pueden presentarse estructuras como “se ha definido” o “se puede definir”, así como con el verbo *ser* o *estar* (es definido, está caracterizado), en donde cambia la temporalidad o la potencialidad, indicando otras connotaciones a las que se pueden presentar con “se define”.

El etiquetado de un patrón verbal definitorio, con todos sus elementos, se ejemplifica a continuación:

> <PVD><SEmarc>se</SEmarc> <VAUX>han</VAUX> <VD lema="utilizar"> utilizado</VD> <NX tipnx="prep">para</NX></PVD>

Como ya hemos mencionado, los patrones verbales dan mucha información sobre el tipo de definición de un contexto definitorio; en el caso del ejemplo arriba citado es claro que se trata de una definición funcional.

Cabe señalar que dentro del etiquetado del PVD se marcan como atributos el lema y el tipo de nexo, tal información ayuda a que la interfaz permita búsquedas a través del lema verbal y de las características del nexo, como veremos más adelante.

Etiquetado de la definición. Una vez etiquetados tanto el término como el patrón verbal definitorio, determinar el tipo de definición se vuelve mucho más sencillo. Generalmente, la definición queda determinada por el tipo de verbo definitorio y, con ayuda del nexo que la introduce, cada verbo definitorio, por sus características semánticas, suele tener asociado un tipo de definición (Sierra y Aguilar 2009). Aunque muchas veces la determinación del tipo de definición no tiene una relación directa con el verbo definitorio, según lo esbozado en la ilustración 3.1 (p. 37). Obsérvese el siguiente ejemplo del corpus:

> las fuentes de las corrientes de cortocircuito […] son: <DEF tipdef=”EXT”>generadores, motores y condensadores síncronos, motores de inducción, conexiones de la red pública, convertidores rotativos y todos los elementos del circuito que se puedan incluir, tales como transformadores, cables, etc</DEF>.

El verbo *ser* generalmente introduce definiciones analíticas o sinonímicas —esto es consistente en la gran mayoría de las muestras de corpus—; pero en el ejemplo anterior se trata de una definición extensional introducida por el verbo *ser*. No obstante, en este caso también deben tomarse en cuenta los dos puntos después del patrón verbal, hecho que ya reporta Aguilar (2009). Se puede decir, entonces, que el patrón verbal definitorio no es completamente determinante para el tipo de definición de un contexto definitorio, si bien muchas veces se puede confiar en que la definición tiene una relación precisa con el verbo que la introduce.

Otro punto que se tiene que tomar en cuenta en la definición es la extensión de ésta, puesto que no siempre corresponde al final de la oración (como es común que pase). Obsérvese el siguiente ejemplo:

> Las tabletas de Fansidar (25 mg de pirimetamina y 500 mg de <TERM tipterm="L" stterm="fn">sulfadoxina</TERM>, que *es una sulfonamida de acción prolongada*), son eficaces para prevenir el paludismo por P. falciparum resistente a cloroquina, aunque también se ha observado resistencia al Fansidar en el sudeste de Asia y en Brasil.

Aquí observamos que existen dos contextos definitorios, uno dentro del otro, el primero es más amplio y su término es “Las tabletas de Fansidar”, mientras su definición abarca “eficaces para prevenir el paludismo por P. falciparum resistente a cloroquina, aunque también se ha observado resistencia al Fansidar en el sudeste de Asia y en Brasil”. El otro contexto definitorio se encuentra dentro de la oración entre paréntesis cuya definición termina en el cierre de éste. De igual forma, ciertos patrones léxicos pueden ayudar a determinar los límites de una definición, tales como *mientras*, *y*, *por otro lado*, entre otros (Hernández, 2009). Un ejemplo se presenta a continuación:

> Un sistema cerrado es <DEF tipdef="GD">aquel que ofrece un proveedor, y en el que se puede instalar el hardware y el software y obtenerse sólo los servicios de dicho proveedor</DEF>, *en tanto que* un sistema abierto es <DEF tipdef=”GD”>aquel en el que se puede trabajar simultáneamente con varios proveedores diferentes</DEF>.

Aquí se encuentran dos definiciones, así como dos términos y dos patrones verbales coordinados por “en tanto que”. La etiqueta de definición concluye antes de este nexo, mientras que la otra definición se etiqueta como un nuevo elemento, es decir se tendrían aquí dos contextos definitorios, dos entradas en el corpus, una con el término “sistema cerrado” y otra con “sistema abierto”.

Etiquetado del patrón pragmático. El patrón pragmático no aparece en todos los contextos definitorios por tratarse de un elemento que, a diferencia del término, la definición y el patrón definitorio, no es necesario para obtener una definición, sino que modifica o agrega información al contexto definitorio. De hecho, actualmente, la interfaz despliega sólo 34 casos: 10 de autoría, 4 de temporales y 22 instruccionales.

El etiquetado del patrón pragmático es, quizás, uno de los más difíciles en determinar. Se tomó en cuenta todos aquellos elementos del contexto definitorio que agregaban información adicional a éste. En el caso de los patrones pragmáticos de autoría, la mayoría correspondían a una estructura de Autor + Patrón definitorio + Término + Definición. Este tipo de patrones siempre se etiquetó con la característica de autoría. Por otro lado, los temporales fueron todos aquellos que indicaban el tiempo en que se dio la definición. Por último, se consideraron los patrones pragmáticos instruccionales, una categoría más complicada de definir que las anteriores, puesto que en ella entran todos aquellos patrones que modifiquen al contexto definitorio y que no sean ni autorales ni temporales. Muchos de éstos correspondían a adverbios modales (“normalmente”, “matemáticamente”, etc.) o con características de espacialidad (“en las matemáticas”, “en el marco legal”, etc.), entre otras construcciones más complejas, tales como frases más elaboradas (“si nos colocamos en el punto de vista de Marx”, “siguiendo un procedimiento común en estudios de alcance de olas”, etc.). Un ejemplo del etiquetado del patrón pragmático fue:

> <PP tipp="Inst">Cuando el ducto está expuesto a la acción del oleaje</PP>, el número de Keulegan–Carpenter o parámetro del período (KC) se utiliza para describir el comportamiento de los coeficientes hidrodinámicos (<PP tipp="Aut"> Aguilar</PP>, <PP tipp="Temp">2000</PP>).

Etiquetado de marcadores definitorios, tipográficos y reformulativos. A este respecto, el etiquetado de estos dos marcadores consistió en identificar aquellas marcas tipográficas que introdujeran una definición o marcadores definitorios como dos puntos, coma, guión, etc.; el siguiente contexto definitorio sirve como ejemplo a este etiquetado:

> A. Richling y T. Bartkowski, entre otros; quienes consideran a la Ecología del Paisaje como <MTD mdef="dp">:</MTD> <MTD mdef="dl">"</MTD>ciencia transdisciplinar que tiene como objetivo principal la resolución del problema de la gestión y desarrollo de los territorios a escala regional y local a lo que le llaman algunos, Ecosistema Humano Total.

Por otro lado, los marcadores reformulativos presentan una perspectiva similar, también introducen una definición de forma similar a los patrones verbales, pero se trata de frases que reformulan una idea. Por ejemplo:

> Sin duda, lo más revolucionario en el aspecto médico es la terapia génica, <MRD>es decir</MRD> el uso del ADN para intentar corregir algún gen defectuoso.

En este caso, la frase "es decir" funciona como introductor de la definición del término "la terapia genética". Sobre este tipo de marcadores se puede encontrar en Hernández (2009), quien considera tanto "es decir", como "o sea".

Así, se etiquetó cada uno de los elementos de los contextos definitorios con sus respectivos atributos, de tal forma que la interfaz pudiera leer las etiquetas XML e interpretarlas para que el usuario final pudiera consultar, de manera fácil y rápida, los contextos definitorios que conforman el corpus.

5. El Corpus de contextos definitorios: CORCODE

Ya se ha descrito la forma en que se realizó la selección y el etiquetado de los contextos definitorios para el corpus, lo que es un paso importante para el manejo rápido y eficaz de las grandes cantidades de texto que utilizamos, puesto que —como se ha mencionado ya varias veces— esto permite un mejor procesamiento computacional del corpus. Sin embargo, estos últimos pasos sólo nos dieron un archivo de difícil interpretación; por tanto, debió realizarse una interfaz que permitiera una consulta sencilla del corpus de contextos definitorios o CORCODE. A continuación nos enfocamos en explicar la realización de esta interfaz sin ahondar en aspectos técnicos de ésta; de igual forma, explicamos cómo funciona y la manera de usarla.

5.1. Conformación de la interfaz del corpus

Posterior al etiquetado XML, el archivo final con los contextos definitorios ya etiquetados por sus constituyentes tuvo que ser procesado por medio de una interfaz que permitiera búsquedas a partir de distintos criterios que se consideraron importantes para un corpus de esta índole, estos son: búsquedas a través de cadenas de caracteres y búsquedas de cada uno de los constituyentes de los contextos definitorios y sus respectivos atributos; además, se tomó en cuenta que el corpus buscaba ser multipropósito y reutilizable, por lo que se decidió que pudiera consultarse a través de internet. Con estos criterios, un equipo de programadores del GIL prosiguió con la conformación de la interfaz web del corpus de contextos definitorios.

5.1.1. Procesos de elaboración de la interfaz

La interfaz consiste en un programa informático que permite acceder al archivo creado en XML a partir de una forma amigable y sencilla para los usuarios que deseen consultarla. Como se ha dicho, el público al que va dirigido el CORCODE es diverso, pues abarca desde lexicógrafos y terminólogos hasta usuarios interesados en el tema o simplemente aquellos que quieran consultar algún término en las áreas de especialidad con las que cuenta el corpus.

Los contextos definitorios etiquetados tuvieron que ser validados de acuerdo a un esquema XML, el cual permite identificar automáticamente etiquetas que tengan errores o que sean incorrectas. El esquema contiene información sobre las etiquetas y es el que determina cuáles van a ser éstas y cuáles serán los atributos que lleve cada una de ellas, por tanto, es de gran importancia para la elaboración del corpus. Se realiza a la par del proceso de etiquetado y, por tanto, sirvió para validarlo.

Después de esta validación se corrigieron aquellas etiquetas que contaban con errores, de tal forma que el sistema informático pudiera identificarlas correctamente. Posteriormente, el archivo que contenía los contextos definitorios con el etiquetado fue implementando en la interfaz web gracias al apoyo del Mtro. Carlos Francisco Méndez Cruz y de uno de sus becarios, Daniel Alberto Medrano Domínguez.

Cabe destacar que ya existía un corpus de contextos definitorios que se podía consultar en internet. Tal corpus se describe en Sierra *et al* (2006). Esta interfaz contaba con consultas sólo a partir de los elementos constituyentes de los contextos definitorios y contenía menos de 200 de éstos. Las etiquetas que describen los autores también varían conforme a las que se utilizaron en esta tesis, las variaciones hechas a las etiquetas XML entre el primer corpus y el actual son las siguientes:

Etiquetas y atributos actuales	**Etiquetas y atributos del antiguo corpus**
CD → num	Se mantuvo sin ningún cambio
TERM → [tipterm= (L,NL) | stterm= (fn, fn-fp, fv-fn)]	Utilizaba, para la estructura del término, *fnYfprep* y *fvYfn* que se sustituyó por el uso del guión para no mezclar mayúscula y minúscula.
DEF → [tipdef= (GD, FUN, EXT, Sin)]	Se dejó de utilizar el valor Ges, que refería a definiciones de género específico.

PVD	Se mantuvo sin ningún cambio
VD → [lema= "verbo en infinitivo"]	Se eliminaron los atributos de argumentos *[args=(2,3)]* y de modo *[mdo=(inf,ger,part,fin)]*
SEmarc	Se quitó el atributo *tipsemarc=(enc,prec)*, que refería a si era un *se* enclítico o proclítico
VAUX	Antes se mezclaban mayúsculas con minúsculas y se usaba *Vaux*.
NX → [tipnx= (adv, prep)]	Se mantuvo sin ningún cambio
MRD	Se mantuvo sin ningún cambio
MTD → [MDEF = (dp, vin, par, gui, dl), MT = (curs, neg, subr, otr)]	Se cambiaron las minúsculas (*mdef* y *mt*) por mayúsculas
PP → [tipp = (Aut, Temp, Inst)	Se modificó *tippp* a *tipp*, ya que presentaba confusiones.

Tabla 5.1. Variaciones entre las etiquetas y los atributos del CORCODE actual y el de Sierra *et al* (2006)

Las etiquetas de CD, NX y MRD se conservaron con sus atributos respectivos.

Se cambiaron los campos de *stterm* de la etiqueta TERM, pasando a minúsculas las letras y cambiando *Y* por un guión, ya que de esta forma el marcado se hacía más sencillo al no estar intercalando entre mayúsculas y minúsculas.

De igual forma, se eliminó el campo del atributo *lema* GS de la etiqueta DEF que hacía referencia a un tipo de definición ya no considerado, genus específico; los contextos definitorios que ya contenían está indicación fueron analizados y cambiados, en la mayoría de los casos, por las marcas de definiciones sinonímicas o analíticas.

También, se eliminaron los atributos de VD *mdo* y *args*, que hacían referencia al modo verbal y los argumentos de éste, porque resultaban innecesarios para los objetivos planteados del corpus y dificultaban en gran medida el etiquetado, haciendo que el tiempo de éste se extendiera al tratar de determinar el modo y el argumento del verbo definitorio.

Por otro lado, se quitaron todos los atributos de SEmarc, pues tampoco nos eran útiles para los fines perseguidos y decidió sólo marcar cuando aparecía.

Se hicieron cambios en VAUX, sustituyendo las minúsculas por mayúsculas para que no se tuviera que intercalar entre diferentes tipos de letra y el etiquetado fuera más rápido.

Por último, el atributo tipp de PP fue cambiado para que fuera más fácil al reducir el número de caracteres similares que se presentaban.

Podemos decir, entonces, que los cambios realizados en las etiquetas que plantearon Sierra *et al* (2006) permitieron que fuera más sencillo el proceso de marcado, así como la programación del sistema al uniformar, en cierta medida, la tipografía de éstas y también combinar, en la menor medida posible, caracteres en minúsculas con caracteres en mayúsculas. De esta forma, se hizo más sencilla la búsqueda de las etiquetas y, por tanto, de los constituyentes de los contextos definitorios a través de la interfaz.

Finalmente, se desarrolló una nueva interfaz completamente distinta a la que ya existía en internet, además de que se agregaron los nuevos contextos definitorios que surgieron a lo largo del proyecto desarrollado en esta tesis. La antigua interfaz sólo contaba con una sola sección donde se hacían búsquedas a través de los constituyentes de cada contexto definitorio; a la nueva interfaz se le agregaron otras secciones: al abrir la página se despliega una sección de bienvenida, también se tiene información acerca del CORCODE, su desarrollo, las fuentes de éste, bibliografía acerca del proyecto y las personas e instituciones participantes; las búsquedas, como ya hemos mencionado, ahora se pueden hacer también por cadenas textuales.

Para el desarrollo de la interfaz se utilizaron varios lenguajes de marcado y programación. Se manejó, por ejemplo, el HTML, los *scripts* desarrollados en Java, hojas de estilo que

permiten la visualización de los archivos XML tal como aparecen ya en la interfaz, entre otras herramientas computacionales.

5.1.2. Modificaciones requeridas para la nueva interfaz

Empero, en el proceso de elaboración de la interfaz se necesitaron distintos tipos de modificaciones que correspondían a los cambios hechos. Principalmente se presentaron modificaciones debido a la variación de las etiquetas, a la falta de etiquetas en algunos constituyentes o de elementos de éstas, a la aparición de dos o más términos en un solo contexto definitorio y a que se eliminaron o agregaron elementos a los contextos definitorios.

Modificaciones por la variación de las etiquetas. La variación de etiquetas entre las propuestas en Sierra *et al* (2006) y las que se presentan en esta tesis causó una gran cantidad de dificultades en el momento de integrar los nuevos contextos definitorios a la interfaz. La validación de las etiquetas se hace, como ya hemos mencionado arriba, a partir de un esquema XML que busca etiquetas incorrectas; para validar las nuevas etiquetas se tuvo que cambiar el esquema. También hubo conflictos en el momento de realizar las búsquedas de los constituyentes de los contextos definitorios por medio de la interfaz, ya que ésta buscaba las etiquetas antiguas, por lo que se tuvo que reprogramar esta opción de búsqueda. Otra modificación que se consideró fue integrar los contextos definitorios que existían en la interfaz antigua con los nuevos; se tuvo, entonces, que revisar el marcado de esos contextos definitorios y cambiar las etiquetas viejas por las nuevas para, posteriormente, validarlos con el nuevo esquema. Se eliminó, de igual forma, la etiqueta FUENTE, que no era interpretada por el sistema.

Problemas por falta de etiquetas o elementos de éstas. Algunos de los contextos definitorios no estaban completamente etiquetados o faltaba por rellenar uno o más de sus atributos. Esto implicaba que el sistema no detectara tales constituyentes o atributos y, por consecuencia, no los devolviera en la búsqueda. Al implementar el sistema se dio cuenta de cuáles eran los contextos definitorios a los que les hacía falta uno o más elementos de marcado. Ante esto, se prosiguió a revisar qué atributos o etiquetas faltaban y a agregárselas. También hizo falta cambiar el atributo *tipdef* cuando aparecía *Ges* (Género

específico), puesto que este tipo de definición no era reconocido por el sistema, ya que ha dejado de ser considerado.

Problemas por la aparición de uno o más términos en un contexto definitorio. Al aparecer más de un término también se presentan patrones verbales definitorios y definiciones. Ante esto, se decidió etiquetar dos veces el mismo contexto definitorio; sin embargo, al realizar este proceso, también se tuvo que cambiar la forma en que se presentaban los contextos definitorios en la interfaz puesto que parecería que éstos se repiten al aparecer más de una vez en el corpus. Se decidió, entonces, mostrar en colores los constituyentes principales de cada uno de ellos (término, patrón verbal y definición) para poder distinguir a qué término hace referencia cada uno de los contextos definitorios. Esto también ayuda a distinguir más claramente en el corpus cada una de las partes de estas unidades definitorias, como podemos ver en la imagen siguiente:

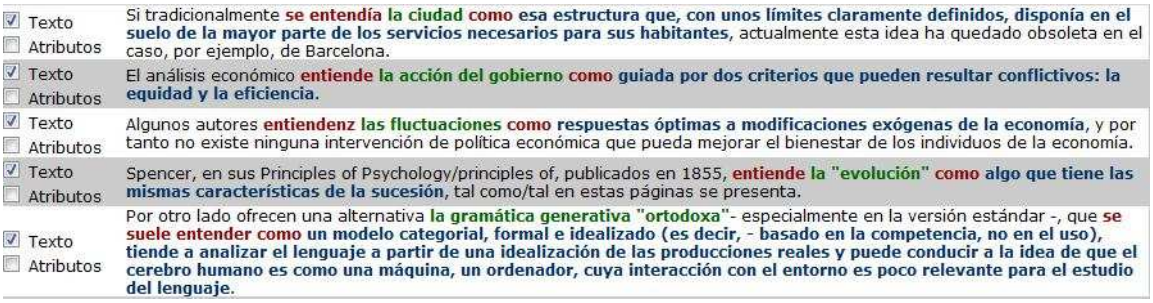

Ilustración 5.1. Muestra de los colores utilizados en la interfaz para ilustrar los constituyentes de los contextos definitorios

Problemas al eliminar o agregar elementos. Para la nueva interfaz del corpus se eliminaron diversos elementos, además de que se obtuvieron más verbos definitorios de los que se tenían considerados en la antigua interfaz. Para esto, simplemente se agregaron o eliminaron a la herramienta de búsqueda por criterios los elementos correspondientes. Se quitó la búsqueda por género específico y se agregaron los lemas de aquellos verbos definitorios que faltaban.

5.2. Interfaz web para el CORCODE

Solucionados los problemas, se procedió ya con la implementación del sistema en la web. Éste puede consultarse a través de la liga http://www.iling.unam.mx/corcode/ y es de libre

acceso para todo aquel público que tenga interés en él. A continuación describimos la interfaz tal como se encuentra actualmente en internet.

5.2.1. Presentación del CORCODE

Como ya hemos mencionado, la nueva interfaz muestra una pantalla de bienvenida que explica brevemente de qué se trata el CORCODE; el menú se presenta en la parte superior de la pantalla y muestra las opciones de *Inicio*, *Acerca de*, *Herramientas*, *Proyecto* y *Ayuda*. A continuación mostramos la pantalla de bienvenida de la interfaz:

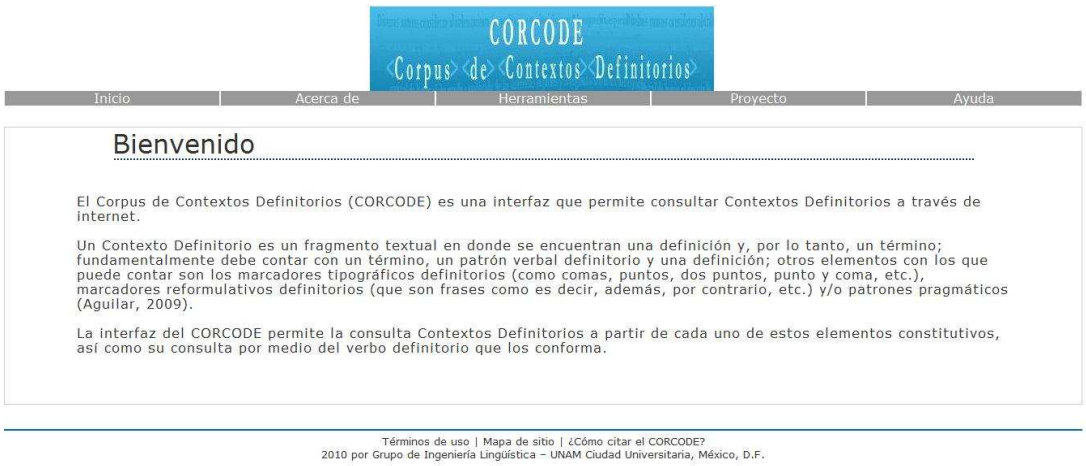

Ilustración 5.2. Pantalla del menú de Inicio de la interfaz del corpus

Inicio. Este apartado muestra la pantalla de bienvenida de la página, lo que permite que se pueda entrar a esta sección cuando ya se ha accedido a las otras opciones de la parte superior de la página.

Acerca de. En este apartado de la página web se describe el proyecto al que pertenece el CORCODE, así como la motivación de éste y los criterios bajo los que se ha desarrollado. También se explica brevemente las fuentes a partir de las cuales se obtuvieron los contextos definitorios del corpus.

Herramientas. Este apartado se refiere a las herramientas de búsqueda que permite el CORCODE. Como hemos dicho, se tienen dos formas de búsqueda dentro del corpus, a

partir de cadenas textuales y por criterios. Estos dos tipos de búsqueda se explicarán detalladamente en 5.2.2.

Proyecto. Aquí se puede acceder a tres apartados distintos: *Participantes*, *Bibliografía* e *Instituciones*. El primero de estos, *Participantes*, muestra las personas que colaboraron para el desarrollo del corpus, quienes son el Dr. Gerardo Eugenio Sierra Martínez, jefe del Grupo de Ingeniería Lingüística (GIL) de quien está a cargo el proyecto; el Mtro. Carlos Francisco Méndez Cruz, quien se ha hecho cargo de la parte informática y la implementación en internet del CORCODE con apoyo de uno de sus becarios, Daniel Alberto Medrano Domínguez. El apartado de *Bibliografía* enumera los artículos y tesis que han surgido dentro del proyecto del CORCODE y aquellos que han servido para su elaboración. Y, por último, el apartado de *Instituciones* muestra a aquellas gracias a las cuales se ha llevado el proyecto: el GIL, de dónde ha surgido el CORCODE, y por tanto el Instituto de Ingeniería y la UNAM; también se desarrolló gracias al apoyo de la Dirección General de Asuntos del Personal Académico (DGAPA) sin cuyo apoyo no hubiera sido posible llevar a cabo el corpus.

Ayuda. Muestra ayuda para la utilización del CORCODE; actualmente, muestra la forma en que debe citarse el corpus para los usuarios que lo consulten. Aún falta desarrollar más esta sección y presentar opciones de ayuda que permitan que los usuarios puedan acceder al corpus de forma más sencilla.

5.2.2. Formas de consulta

Como ya hemos mencionado, existen dos formas de consulta por medio de las cuales se puede consultar el CORCODE, por cadena y por criterios. En este apartado nos enfocamos en la descripción de la utilización de ambas formas de consulta del corpus, describiendo todas las posibilidades que ambas búsquedas permiten.

a) *Por cadena*. Las búsquedas a partir de cadenas textuales se dan a partir de la introducción de una palabra, serie de palabras o simplemente de cadenas de caracteres que se buscan dentro del corpus y se muestran al usuario marcadas en amarillo dentro de todo el contexto definitorio. Por ejemplo, se puede buscar una palabra como *computadora*, un término compuesto como *computadoras*

programables o bien sólo una cadena textual como *compt*. Aparecerán los contextos definitorios que contengan alguna de estas cadenas y éstas se marcarán en amarillo cada vez que aparezcan. Las cadenas de búsqueda se introducen dentro de un campo textual y se puede elegir en qué elemento del contexto definitorio se desea buscar la cadena; se puede buscar dentro de:

a. Todos. Busca las cadenas dentro de todo el contexto definitorio.
b. Término. Busca la cadena textual introducida solamente dentro del término del contexto definitorio. De esta forma se pueden buscar las definiciones de términos puntuales, lo que hace del CORCODE una herramienta de consulta no sólo para especialistas, sino también para aquellos interesados en algún término especializado. Cabe señalar que el CORCODE no es un diccionario, pero, a partir de esta herramienta, se pueden solucionar dudas sobre ciertas definiciones de ámbitos especializados, o bien se puede utilizar para la terminología y el desarrollo posterior de diccionarios especializados.
c. Definición. Con esta opción se obtiene la cadena textual ingresada en la búsqueda únicamente dentro de toda la definición. Tampoco se trata de búsquedas como las que se podrían hacer en un diccionario onomasiológico, empero permite buscar palabras y cadenas dentro de las definiciones de cada uno de los contextos definitorios que pueden ayudar a encontrar un término desconocido.
d. Verbo definitorio. Busca cadenas textuales dentro del verbo definitorio, lo cual complementa la búsqueda de lemas que se puede realizar en la búsqueda por criterios, puesto que esta última devuelve todos los verbos pertenecientes a los lemas definitorios considerados en el CORCODE, mientras que la búsqueda por cadenas dentro del verbo definitorio permite también buscar los verbos en una forma conjugada específica, como por ejemplo puede buscarse una cadena como *considerado*, lo que devolvería sólo los contextos definitorios con un verbo definitorio conjugado en esta forma.

La forma en que se muestra en la interfaz la búsqueda por cadenas es la siguiente:

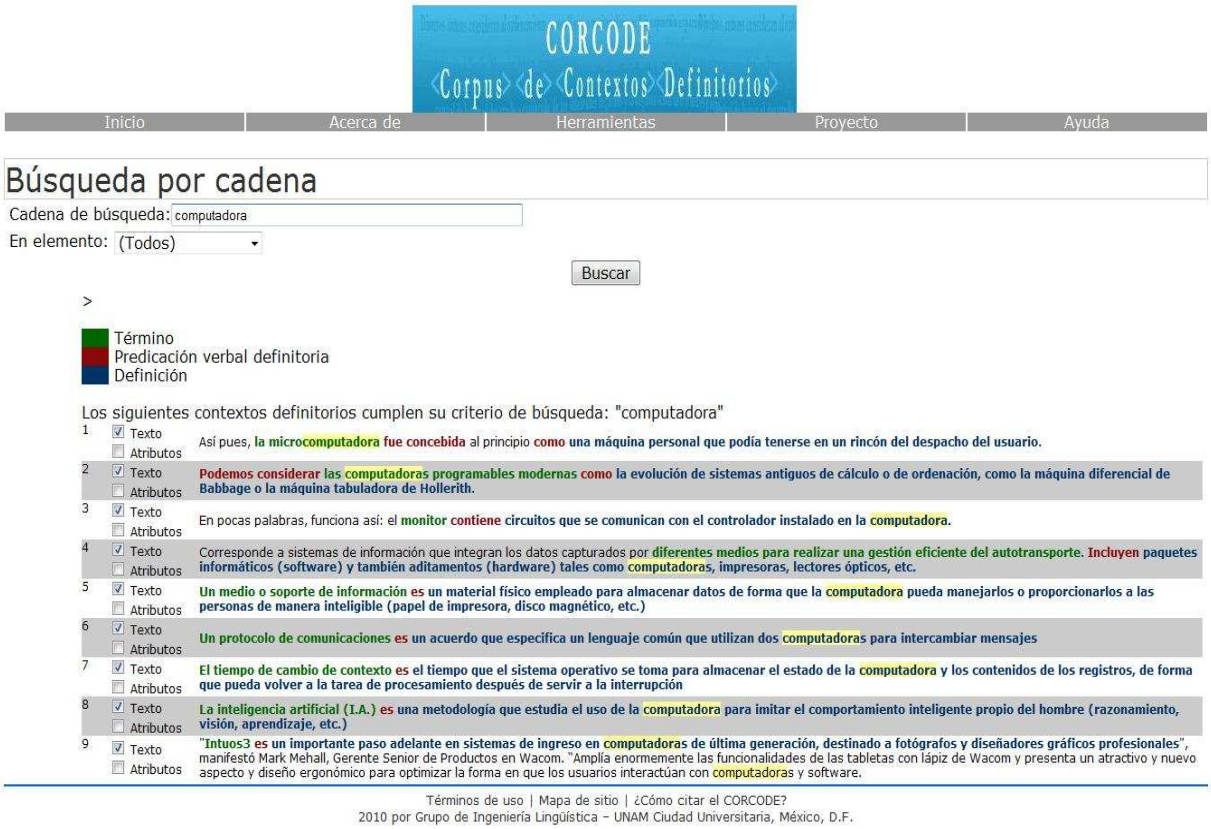

Ilustración 5.1. Ejemplo de búsqueda por cadena dentro del CORCODE. Las cadenas buscadas aparecen remarcadas en color amarillo

b) *Por criterios*. La búsqueda por criterios permite buscar cada uno de los constituyentes de los contextos definitorios, esto es, partir de *Término*, *Definición*, *Verbo definitorio*, *Marcadores* y *Patrón pragmático*. Cada uno de estos elementos permite consultas a partir de sus atributos gracias al etiquetado XML que se realizó. A continuación describimos cómo se realizan tales consultas según cada uno de los elementos de los contextos definitorios.

Término. La herramienta de búsqueda por criterios dentro del término despliega un menú que permite buscar términos lingüísticos y no lingüísticos en contextos definitorios, además permite seleccionar la estructura del término que se desea obtener, ya sea una frase nominal, una frase nominal y una prepositiva, o una frase

verbal más una nominal; se mostrará, entonces, los contextos definitorios que cumplan con los criterios seleccionados por el usuario.

Definición. En el menú de búsqueda por definición se pueden seleccionar el tipo de definición que se desee obtener, esto se basa en la tipología ya definida más arriba (3.2.2), es decir, por genus y diferencia o analítica, funcional, extensional y sinonímica. Este tipo de herramienta permite obtener contextos definitorios según la tipología de las definiciones, lo que bien puede servir para las investigaciones que se realizan dentro del GIL o en otros grupos de estudios acerca de definiciones. En el GIL, verbigracia, Sánchez (2009) estudió las definiciones funcionales y actualmente se realizan investigaciones sobre las de tipo analítica.

Verbo definitorio. Por criterios del verbo definitorio se pueden buscar elementos que complementan al verbo. El primero de éstos es el *nexo*, donde se puede seleccionar si se desean nexos adverbiales o preposicionales, contando actualmente como adverbial únicamente el nexo *como*. También se pueden obtener aquellos verbos que sean acompañados por un *clítico se* o un *verbo auxiliar*. Podemos seleccionar, de igual forma, si contienen o no *marcadores reformulativos*, aunque la interfaz actual no permite seleccionar qué tipo de marcadores reformulativos se buscan, ya que no existe un estudio de clasificación detallado de éstos. Por último, se puede seleccionar el *lema* del verbo definitorio para obtener sólo los contextos definitorios compuestos por este verbo conjugado en todas sus formas posibles. Actualmente se cuenta con los siguientes lemas:

Lemas	
Caracterizar	Ver
Concebir	Componer
Consistir	Constar
Considerar	Constituir
Conocer	Contener
Definir	Corresponder
Comprender	Determinar
Describir	Emplear

Constar	Encargar
Entender	Identificar
Permitir	Incluir
Servir	Integrar
Usar	Interpretar
Utilizar	Referir
Denominar	Representar
Llamar	Significar
Ser	Visualizar

Tabla 5.2. Lemas de los verbos que actualmente se pueden consultar en la interfaz del CORCODE

Marcadores. En marcadores se puede seleccionar la búsqueda a partir de dos tipos de éstos: *definitorios* y *tipográficos.* Los primeros pueden ser: *dos puntos, viñetas, paréntesis, guiones, comillas* u *otros*, que enmarcan a aquellos marcadores que pueden insertar una definición. Conforme a los marcadores tipográficos se puede buscar por: *cursivas, negritas, subrayado* u *otra marca*.

Patrón pragmático. En los patrones pragmáticos se puede hacer una consulta según la tipología de estos patrones esbozada ya en 3.2.4. Por tanto, se pueden obtener contextos definitorios con patrones pragmáticos *autorales, temporales* e *instruccionales.*

La imagen de esta interfaz se muestra a continuación:

Ilustración 5.3. Imagen de la búsqueda por criterios en donde se puede apreciar, en la barra de la izquierda, algunos de los criterios de este tipo de búsquedas

Por otro lado, cabe aclarar cómo funcionan las etiquetas dentro de la interfaz para que ésta pueda realizar las búsquedas. Para esto, presentamos el siguiente cuadro que relaciona cada una de las etiquetas con las opciones de búsqueda de la interfaz:

Opción de búsqueda		**Etiqueta XML**	**Atributos**
Término	Tipo	<TERM>	tipterm
	Estructura: Fr. nominal Fr. nominal y fr. prepositiva Fr. nominal y fr. verbal		stterm “fn” “fn-fp” “fv-fn”
Definición	Tipo: Genus y diferencia Funcional Extensional Sinonímica	<DEF>	tipdef “GD” “FUN” “EXT” “Sin”
Verbo definitorio	Nexo: Preposición Adverbio	<PVD>	<NX> “prep” “adv”
	Clítico SE:		<SEmarc>

	Presente No presente		Con marca Sin marca
	Verbo auxiliar: Presente No presente		<VAUX> Con marca Sin marca
	Marcadores reformulativos: Presente No presente		<MRD> Con marca Sin marca
	Lema		<VD lema=“”>
Marcadores	Definitorios: Dos puntos : Viñetas Paréntesis () Guiones -X- Comillas “” Otros	<MTD>	mdef “dp” “viñ” “par” “gui” “dl” Cualquier valor
	Tipográficos: Cursiva Negritas Subrayado Otra marca		mt “curs” “neg” “subr” “otr”
Patrón pragmático	Tipo: Autoría Temporales Instruccionales	<PP>	tipp “Aut” “Temp” “Inst”

Tabla 5.1. Relación de las etiquetas XML con las opciones de búsqueda que ofrece la interfaz web del corpus

Como observamos en el cuadro, las opciones de búsqueda que permite la interfaz web dependen completamente del etiquetado previo en XML. Este marcado, como hemos mencionado, es el que hace posible que el sistema identifique las búsquedas de los constituyentes de cada contexto definitorio. Cómo observamos en la tabla, la búsqueda de términos corresponde a la etiqueta <TERM> y los atributos de esta etiqueta determinan el tipo de búsquedas que permite la interfaz. Por ejemplo, si el atributo “stterm”, que corresponde a la estructura del término, está marcado con “fn-fp”, la interfaz responderá a

búsquedas de “Fr. nominal y fr. prepositiva”. Igual pasa con las otras etiquetas y atributos; si buscamos definiciones funcionales, la interfaz devolverá todos aquellos contextos definitorios cuyo atributo “tipdef” sea “FUN”; si son extensionales corresponderá a “EXT”, etc. En los caso del clítico SE, Verbo Auxiliar y los Marcadores reformulativos, la interfaz sólo busca si existe la etiqueta o si está ausente para determinar la búsqueda, mostrando si tales elementos están presentes o no presentes.

Notamos, entonces, que la interfaz permite dos tipos de búsqueda, la primera por medio de cadenas textuales, que consiste en encontrar dentro del corpus las cadenas de caracteres que sean introducidas por el usuario; la segunda se ayuda de las etiquetas XML y consiste en encontrar a partir de criterios específicos los elementos constituyentes de cada contexto definitorio y sus características.

5.3. Beneficios del CORCODE

Las opciones de búsqueda permiten la consulta del corpus, de tal forma que cualquier usuario pueda tener acceso al CORCODE de forma rápida y sencilla. Como hemos visto, se pueden buscar palabras dentro de los términos, en las definiciones y en los verbos definitorios, de tal forma que el corpus puede ser consultado por aquellos que no tengan grandes conocimientos dentro del área de la terminología y que sólo busquen solucionar una duda acerca de un término especializado o buscar un término con palabras clave de la definición. Pero también, como planteábamos en la introducción de esta tesis, un corpus de esta índole es de gran beneficio para lexicógrafos y terminólogos, puesto que permite la consulta de contextos definitorios a partir de criterios bien delimitados; puede, por tanto, permitir la consulta de puntos precisos que resultarían de gran utilidad para investigaciones precisas. Dentro del GIL es una herramienta de gran utilidad, ya que permite y permitirá que se sigan realizando investigaciones en el área, de tal forma que el conocimiento que se tienen sobre estas unidades definitorias se enriquezca, lo que permitirá que la extracción automática de contextos definitorios (actualmente realizada a través del ECODE) sea más precisa y que otras herramientas, como el Describe® o el mismo CORCODE, se desarrollen y se mejoren.

6. Conclusiones

6.1. Resumen de la tesis

Antes de las conclusiones finales y de mencionar el trabajo a futuro, haremos un resumen breve de cada uno de los capítulos de la tesis.

Lingüística de corpus. Vimos cómo se entiende a la lingüística de corpus y a los corpus lingüísticos. Se definió un corpus como "un conjunto de datos reales y aceptables, debidamente ordenado, codificado y organizado, de diferentes textos recopilados, pertenecientes a un código lingüístico determinado, oral o escrito" (Sierra 2008), y por tanto, la lingüística de corpus como el estudio de la lengua mediante la utilización de corpus lingüísticos. También hemos descrito algunos procesos informáticos que permiten las consulta de un corpus, como el etiquetado y el estándar XML (eXtensible Markup Language), que realiza el marcado de las partes de un documento por medio de etiquetas editables.

Contextos definitorios. Por otro lado, también se definió qué es un contexto definitorio, es decir, un fragmento textual extraído de un documento y que contiene la definición de un término y, por tanto, un término que se define. Describimos su estructura basada en componentes como término, definición, patrones definitorios y patrones pragmáticos; a su vez se describieron estos constituyentes y una tipología de definiciones que concibe cuatro de éstas: analítica, sinonímica, extensional y funcional.

Metodología y diseño de un corpus para contextos definitorios. Una vez estudiados los puntos anteriores hemos pasado a describir la metodología seguida para la conformación de un corpus de contextos definitorios; desde la recolección de los candidatos, su selección hasta su etiquetado. De igual forma, se han descrito las etiquetas utilizadas en el marcado de los elementos y el proceso cómo se etiquetó, que consistió principalmente en el uso de XML Writer, una aplicación que permite un marcado semiautomático con la implementación de un esquema XML.

El corpus de contextos definitorios: CORCODE. Por último, nos hemos dedicado a describir la interfaz, cómo se elaboró y cómo utilizarla. En esto entra la descripción de los elementos integrados a la página web y a los tipos de búsqueda, mismos que se realizan de dos formas: por criterios y por cadenas textuales. Para terminar, se ha analizado la utilidad de un corpus de este tipo para la lingüística.

6.2. Conclusiones

Podemos concluir que la metodología aquí expuesta para la elaboración del CORCODE es de gran utilidad para el desarrollo de corpus informáticos. Esta metodología permite que los corpus puedan consultarse a través de internet y que las búsquedas dentro de éstos sean eficientes y rápidas gracias a la implementación del marcaje XML. La metodología aquí esbozada ha demostrado su funcionalidad para la elaboración de corpus informáticos. En este caso, la prueba de esto es el corpus desarrollado al margen de esta tesis: el CORCODE.

Dentro de las ventajas que presenta un corpus como el CORCODE, podemos decir que es de utilidad para lexicógrafos y terminólogos interesados en las áreas de especialidad que se contienen en este corpus. Como ya hemos comentado, el CORCODE no sólo está pensado para usuarios interesados en estas áreas o en otras de la lingüística, sino que también busca ser una herramienta de utilidad para cualquier usuario, estudiante o usuario interesado en alguna materia contenida en el corpus.

De igual forma, el corpus se encuentra en una página web, lo que permite que cualquier persona con acceso a una computadora con internet pueda consultarlo. Esto hace del CORCODE un corpus accesible. Además se trata de un corpus gratuito, puesto que no necesita de pago para acceder a él, ni tampoco de un registro, de tal forma que cualquier usuario interesado puede consultarlo de manera fácil y sin costo.

Resumiendo, podemos decir que el CORCODE presenta las siguientes ventajas:

a) El CORCODE es un corpus útil para lexicógrafos y terminólogos.
b) Contiene información en diversas áreas del conocimiento.
c) Su interfaz es accesible en internet para todo público interesado.

6.3. Trabajo a futuro

El COROCDE, empero, aún tiene cosas que pueden ser ampliadas y mejoradas. Como se ha mencionado, el CORCODE se retroalimenta con el ECODE, por lo que el corpus seguirá creciendo con los contextos definitorios que se extraigan a partir de este sistema; además, a partir del CORCODE se pretende desarrollar otros proyectos, como el de implementar un corpus multilingüe que contenga contextos definitorios en otras lenguas.

Por último, es importante mencionar que el CORCODE puede ser utilizado en futuras investigaciones que se realicen dentro del GIL, puesto que los contextos definitorios son una herramienta útil para pulir técnicas de extracción de información, además para la realización de diccionarios electrónicos y otras aplicaciones del procesamiento de lenguaje natural. También, los contextos definitorios que ya han sido reunidos en el corpus pueden servir para la investigación en el desarrollo de herramientas como el ECODE o el Describe®.

7. Bibliografía

Aguilar, César. *Metodología de Análisis Lingüístico de definiciones en contextos definitorios.* México: Facultad de Filosofía y Letras, UNAM, 2009.

Aguilar, César, Rodrigo Alarcón, Carlos Rodrúguez, y Gerardo Sierra. «Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados.» En *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad: Actas del IX Simposio Iberoamericano de Terminología RITERM04*, de Teresa Cabré, Rosa Estopà y Carles Tebé, 259-268. Barcelona: IULA-UPF, 2006.

Alarcón, Rodrigo. *Análisis lingüístico de contextos definitorios en textos de especialidad.* México: Facultad de Filosofía y Letras, UNAM, 2003.

—. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios.* Barcelona: Ciencias del Lenguaje y Lingüística Aplicada, Universidad Pomepu Fabra, 2009.

Arrarte, Gerardo. «Normas y estándares para la codificación de textos y para la ingeniería lingüística.» En *Filología e informática. Nuevas tecnologías en los estudios filológicos*, de Juan Manuel Blecua, G. Clavería, C. Sánchez y Joaquim Torruella, 17-44. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, UAB, 1999.

Atkins, Sue, Jeremy Clear, y Nicholas Ostler. «Corpus design criteria.» *Literary and linguistic computing.* 1992. http://llc.oxfordjournals.org/cgi/content/abstract/7/1/1 (último acceso: 25 de febrero de 2010).

Biber, Douglas, Susan Conrad, y Randi Reppen. «Introduction: Goals and methods of the corpus-based approach.» En *Corpus linguistics: Investigatin language structure and use*, de Douglas Biber, Susan Conrad y Randi Reppen, 1-10. Cambridge: Cambridge University Press, 1998.

Hernández, Ariadna. *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática.* México: Facultad de Filosofía y Letras, UNAM, 2009.

Jiménez Pozo, A. «Corpus lingüísticos.» En *Adaptación y mejora de un sistema de preprocesamiento y categorización gramatical*, 31-49. Madrid: Universidad Politécnica de Madrid, 1999.

Lara, Luis Fernando. «Caracterísitcas del "Corpus del español mexicano contemporáneo".» En *Actas del I Congreso Internacional sobre el español de América*, de María Vaquero de Ramírez y Humberto López Morales, 579-586. Academia Puertorriqueña de la Lengua Española, 1987.

McEnery, Tony, y Andrew Wilson. *Corpus Linguistics. An introduction.* Edinburgh: Edinburgh University Press, 2001.

Medina, Alfonso, y Carlos Méndez. «Arquitectura del Corpus Histórico del Español en México (CHEM).» En *Avances en la ciencia de la computación*, de A. Hernández y J. Sechinelli, 248-253. México: Sociedad Mexicana de Ciencia de la Computación, 2006.

Procházková, Petra. «Fundamentos de la lingüística de corpus: Concepción de los corpus y métodos de investigación con corpus.» 3 de Noviembre de 2006. http://www.prochazkova.de/fundamentos_de_la_lingüística_de_corpus.pdf (último acceso: 24 de Marzo de 2010).

Sajer, Juan, y August Ndi-Kimbi. «The conceptual structure of terminological definitions and their linguistic realisations.» *Terminology*, 1995: 87-106.

Sánchez, Octavio. *Análisis de relaciones léxicas en definiciones analíticas, extensionales y funcionales.* México: Facultad de Filosofía y Letras, UNAM, 2009.

Sierra, Gerardo. «Diseño de corpus textuales para fines lingüísticos.» En *IX encuentro Internacional de Lingüística en el Noroeste, Tomo 2*, 455-462. Sonora, 2008.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, Alberto Barrón, Valeria Benítez, y I. Baca. «Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología.» *X Simposio Iberoamericano de Terminología.* Montevideo, 2006.

Sierra, Gerardo, y Alejandro Rosas. «Una clasificación de cospus lingúísticos informatizados.» *Versión en artículo extenso, presentada para su publicación en Las memorias del X Encuentro de Lingüística en el Noroeste.* 2009.

Sierra, Gerardo, y César Aguilar. «A formal scope on the relations between definitions and verbal predications.» *1st International Workshop on Definition Extraction.* Borovets, 2009. 7-13.

Sierra, Gerardo, y Rodrigo Alarcón. «Identification of recurrent patterns to extract definitory contexts.» *Lecture Notes in Computer Science*, 2002: 436-438.

Sierrra, Gerardo, Rodrigo Alarcón, Alejandro Molina, y Edwin Aldana. «Web Exploitation for Definition Extraction.» *Proceedings of the 2009 Latin American Web Congress (la-web 2009)*, 2009: 217-223.

Sinclair, John. «Preliminary recommendations on corpus typology.» *EAGLES.* 1996. http://citeseerx.ist.psu.edu/showciting?cid=8351889 (último acceso: 14 de Octubre de 2010).

Torruella, Joan, y Joaquím Llisterri. «Diseño de corpus textuales y orales.» En *Filología e informática. Nuevas tecnologías en los estudios filológicos*, editado por Juan Blecua, G. Clavería, Sánchez C. y Joan Torruella, 45-77. Barcelona: Milenio, 1999.

Vossen, Piek, y Anne Copestake. «Defaults in lexical representation.» En *Inheritance, defaults and lexicon*, de T. Briscoe, V. Paiva y A. Copestake, 246-274. Cambridge: Cambridge University Press, 1993.

World Wide Web Consortium (W3C). http://www.w3.org/ (último acceso: 10 de 12 de 2010).