

00623
19



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y
ADMINISTRACIÓN

**ANÁLISIS Y DISEÑO DEL ETIQUETADO DE UN
CORPUS LINGÜÍSTICO EN INGENIERÍA**

TESIS PROFESIONAL QUE PARA
OBTENER EL TÍTULO DE:

LICENCIADA EN INFORMÁTICA

PRESENTA:

KARLA IVETTE ORTEGA HERNÁNDEZ

ASESOR:

DR. GERARDO E. SIERRA MARTÍNEZ



MEXICO, D.F.

2003

A



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



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.

DEDICATORIAS

Con amor para mamá y papá por el apoyo incondicional que me han brindado en todas las facetas de mi vida, y por darme la oportunidad de concluir mi carrera profesional, este trabajo también es de ustedes.

Para mi hermanito Fernando y mi abuelita Iris, gracias por todo su apoyo.

A todos ustedes gracias y los quiero mucho.

Karla

B

CONTENIDO

CAPITULO 1

INTRODUCCIÓN	1
1.1 Grupo de Ingeniería Lingüística	2
1.2 La ingeniería lingüística y corpus lingüísticos	3
1.3 Objetivos de la tesis	5
1.4 Estructura de la tesis	7

CAPITULO 2

GENERALIDADES DE CORPUS LINGÜÍSTICOS	9
2.1 ¿Qué es un corpus lingüístico?	10
2.2 Características que debe cumplir un corpus	11
2.3 Tipos de corpus y clasificación	12
2.4 Ejemplos de corpus	14
2.5 ¿Por qué usar un corpus?	22
2.6 Aplicaciones diversas	24

CAPITULO 3

EL CORPUS DE INGENIERÍA	32
3.1 Objetivo general	33
3.2 Importancia del corpus	33
3.2.1 Descripción de contextos definitorios	35
3.2.2 Requerimientos de etiquetado	36
3.3 Clasificación del corpus de ingeniería	38
3.4 Aspectos técnicos del corpus de ingeniería	39
3.5 Definición de descriptores y códigos	40
3.6 Criterios de selección de descriptores y códigos en la selección de información y la captura	42
3.7 Diseño de búsqueda	42
3.8 Diseño de salida	43

3.9 Diseños y tipos de reportes	43
3.10 Definición de procedimientos para la formulación de búsquedas	44
3.11 Diseño de las formas de interacción con el público usuario final	44
3.12 Descripción de la operación de corpus lingüístico	45
3.13 Definición de procedimientos de validación, control de calidad, mantenimiento y actualización del corpus lingüístico	46
3.14 Criterios generales para seleccionar información para el corpus	48

CAPITULO 4

INTRODUCCIÓN AL ETIQUETADO CON XML	50
4.1 Historia de XML	51
4.2 Diferencias entre HTML y XML	53
4.3 Objetivos de XML	55
4.4 Usos de XML	58
4.5 Fundamentos de la sintaxis de XML	60
4.5.1 Etiquetas	61
4.5.2 Referencias de entidades	62
4.5.3 Comentarios	63
4.5.4 Instrucciones de procesamiento	63
4.5.5 Declaraciones de tipos de documento	64
4.5.6 Secciones marcadas	65
4.6 Normalización de etiquetas para corpus lingüísticos	66

CAPITULO 5

ETIQUETADO DE DOCUMENTOS	68
5.1 Etiquetado de libros	69
5.2 Etiquetado de informes	71
5.3 Etiquetado de memorias	73
5.4 Etiquetado de revistas	75

5.5 Estructura del etiquetado de los documentos	77
---	----

CAPITULO 6

ETIQUETADO DEL TEXTO	80
6.1 Etiquetas para el texto	81
6.2 Estructura (elementos del documento)	82
6.2.1 Salto de párrafo	82
6.2.2 Encabezamiento	83
6.2.3 Resumen	84
6.2.4 Bibliografía	84
6.2.5 Texto especial	86
6.2.6 Notas a pie de página	86
6.2.7 Notas a fin de texto	87
6.2.8 Figura	88
6.2.9 Tabla	89
6.2.10 Mapa	90
6.2.11 Gráfica	90
6.2.12 Título de figura	91
6.2.13 Título de tabla	92
6.2.14 Título de mapa	92
6.2.15 Título de fórmula	93
6.2.16 Título de gráfica	93
6.3 Formato	94
6.3.1 Cambio de tipo de letra	94
6.3.2 Cambio de espaciado de letras	95
6.3.3 Letra más grande	96
6.3.4 Letra más pequeña	96
6.3.5 Itálicas	97
6.3.6 Negritas	97
6.3.7 Subrayado	98
6.3.8 Mayúsculas	99
6.3.9 Versales	99

6.3.10 Cambio de margen	100
6.3.11 Viñetas	101
6.4 Referencias	102
6.4.1 Referencias internas	102
6.4.2 Referencias bibliográficas	102
6.4.3 Llamado a pie de página	103
6.4.4 Llamado a fin de texto	103
6.5 Notaciones	104
6.5.1 Fórmulas	104
6.5.2 Subíndice	105
6.5.3 Superíndice	105
6.5.4 Numeración de fórmula	106
6.5.5 Notación de unidad	106
6.5.6 Notación de fórmula	107
6.5.7 Siglas	107
6.5.8 Abreviaturas	108

CAPITULO 7

ADMINISTRACIÓN INTERNA DE LOS DOCUMENTOS

DEL CORPUS	108
7.1 Desarrollo de la base de datos del GIL	109
7.2 Contenido de la base de datos	109
7.3 Registro de documentos	110
7.3.1 Llenado de registro de informes	111
7.3.2 Llenado de registro de libros	112
7.3.4 Llenado de registro de memorias	113
7.3.5 Llenado de registro de revistas	115
7.4 Registro de escaneos	116
7.5 Reportes	120
7.6 Opción de variables de entorno	123
7.6.1 Alta de usuarios	124
7.6.2 Alta de estado del documento	125

7.6.3 Alta de formatos de documentos	126
--	-----

CAPITULO 8

CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS	127
8.1 Conclusiones de la tesis	128
8.2 Líneas de trabajo futuras	131

BIBLIOGRAFIA

Referencias	133
Fuentes usadas en los ejemplos	134
Páginas de internet consultadas	135

ANEXOS

ANEXO A	136
ANEXO B	142
ANEXO C	144
ANEXO D	146

CAPITULO 1

INTRODUCCIÓN

Para dar pie a este trabajo de tesis se dará un panorama general de la ingeniería lingüística, en donde se describen las necesidades del corpus para la ingeniería lingüística, así como los objetivos, metodología y descripción de la tesis.

Cabe hacer mención y agradecer tanto a los programas de apoyo a la investigación de PAPIIT (UNAM) y CONACYT que patrocinaron la realización de este trabajo de investigación, como al Instituto de Ingeniería de la UNAM.

1.1. Grupo de Ingeniería Lingüística (GIL)¹

Esta tesis se desarrolló dentro del Grupo de Ingeniería Lingüística. A continuación se habla del GIL y de los proyectos que ha ido desarrollando desde su existencia en la UNAM.

Con el fin de desarrollar formalmente el área de ingeniería lingüística en la UNAM y a nivel nacional, el Instituto de Ingeniería abre, en septiembre de 1999, el Grupo de Ingeniería Lingüística, dirigido por el Dr. Gerardo Sierra Martínez.

El GIL representa un grupo de investigación en la que dos áreas, al parecer alejadas, retoman el sentido de unidad e interdependencia para formar un solo núcleo. Estas áreas son la lingüística y la ingeniería.

El Grupo de Ingeniería Lingüística, en el seno del Instituto de Ingeniería de la UNAM, se conformó con el fin, primero, de crear una base de conocimiento relativa y concerniente a esta área de trabajo, y segundo, de formar personal especializado y comprometido con el estudio y desarrollo de las diversas áreas que ésta ofrece.

El interés del GIL radica en la realización de proyectos que superen las necesidades y los problemas presentados para el procesamiento del lenguaje natural, incluyendo el desarrollo de aplicaciones específicas que sirvan a las diferentes áreas con las que el GIL interactúa.

Gracias al apoyo del Instituto de Ingeniería y con el patrocinio del Consejo Nacional de Ciencia y Tecnología y de la propia UNAM, el Grupo de Ingeniería Lingüística ha venido realizando proyectos vinculados con el procesamiento de lenguaje natural.

Con fines estratégicos, se está realizando un proyecto central de investigación aplicada, sobre el cual giran las diferentes líneas de investigación,

¹ <http://iling.torcingeneria.unam.mx/>

tales como la lingüística de corpus, recuperación y extracción de información, terminológica, etc.

Este proyecto central, motor del Grupo de Ingeniería Lingüística, persigue crear un sistema de búsqueda onomasiológica, esto es, un diccionario que permita la búsqueda de términos a partir de la descripción del concepto mediante el uso de lenguaje natural.

El Grupo de Ingeniería Lingüística ha empezado a diversificar en otros proyectos, tales como inteligencia tecnológica y traducción automática.

1.2. La ingeniería lingüística y corpus lingüísticos

Esta tesis se centra en el área de la ingeniería lingüística, por ello, es importante definirla, describirla y conocer la importancia para ésta, de la existencia de corpus para el análisis lingüístico.

La ingeniería lingüística es el área interdisciplinaria de investigación aplicada al desarrollo de sistemas computacionales para reconocer, interpretar y generar lenguaje humano. Existe una correspondencia biunívoca, de forma que la lingüística permite la creación de modelos en lenguaje natural que puedan ser utilizados por los sistemas computacionales, mientras que la ingeniería permite el desarrollo de sistemas que puedan resolver las necesidades específicas planteadas por los problemas lingüísticos².

² <http://iling.torreingenieria.unam.mx/>

La ingeniería lingüística o ingeniería del lenguaje natural es el conjunto de las técnicas, fundamentalmente informáticas, que permiten la aplicación de los conocimientos lingüísticos a la industria, las comunicaciones, etc.³

Uno de los terrenos de interés para la terminología es la lingüística de corpus, para el establecimiento de los repertorios terminológicos.

Un corpus no es sólo una colección de archivos de textos acumulada en una computadora: tiene que haber sido objeto de un proceso, el cual es el etiquetado, que permite formalizar tanto las distintas subunidades en que se estructuran dichos textos como las informaciones lingüísticas (categoría gramatical, función sintáctica, etc.) que permitirán, por ejemplo, localizar entre millones de palabras aquellas frases que contienen una definición. Evidentemente no todos los corpus tienen los mismos objetivos ni necesitan la misma profundidad de etiquetado, pero se necesitan herramientas informáticas para automatizarlo. La ingeniería lingüística tiene aportaciones en este campo: analizadores y etiquetadores de varios tipos, etc.⁴

El texto escrito puede introducirse en una computadora tanto desde una fuente impresa —reconocimiento óptico de caracteres (ROC u OCR)— como desde una fuente manuscrita —reconocimiento de caracteres manuscritos o reconocimiento inteligente de caracteres (RIC)—, mientras que, en el caso de la lengua hablada, la entrada de información siempre se lleva a cabo mediante la voz. Sin embargo, en este último caso puede pretenderse la conversión del habla en un texto escrito —reconocimiento de habla—, identificar quién es la persona que habla y verificar su identidad o descubrir automáticamente la lengua que está utilizando un determinado locutor.

³ De Yzaguirre. Ll. (1996) "Ingeniería lingüística y terminología". *Terminómetro. Monográfico: La terminología en España*. págs. 69-71, Unión Latina-IULA, París.

⁴ http://terminotica.upf.es/membres/DE_YZA/PUBLI/INGE.HTM

Además de integrar y generar información lingüística, los sistemas informáticos desarrollados en el marco de la ingeniería lingüística pueden también llevar a cabo el procesamiento de dicha información.

Es por esta razón la importancia para la ingeniería lingüística de la existencia de los denominados recursos lingüísticos, consistentes en corpus textuales, orales o léxicos que proporcionan los datos necesarios para el desarrollo o el funcionamiento de las aplicaciones lingüísticas. La utilización de los corpus está ligada a una serie de procesos y herramientas que facilitan su uso y su explotación.

La codificación consiste en la introducción en el corpus de marcas relacionadas con su estructura y formato, de modo que éste pueda recuperarse para ser utilizado en sistemas informáticos diferentes.

Finalmente, se han creado diversas herramientas para la explotación de los corpus, especialmente en la investigación lingüística y en la lexicografía: entre ellas destacan los programas que realizan listas de palabras ordenándolas en función de su frecuencia de aparición o los que permiten obtener concordancias — en las que las palabras buscadas aparecen ordenadas alfabéticamente, acompañadas de su contexto anterior y posterior—; en esta misma línea, también puede obtenerse información sobre la frecuencia con la que dos o más palabras aparecen seguidas (colocaciones).⁵

1.3. Objetivos de la tesis

El presente trabajo se enmarca en el análisis y desarrollo del etiquetado de un corpus lingüístico en el área de ingeniería. Dentro de este marco, los objetivos planteados en la tesis son los siguientes:

⁵ http://cvc.cervantes.es/obref/anuario/anuario_98/parte2/cap3/llisterri_01.htm

a) Comprensión de las generalidades que rodean al área de la ingeniería lingüística: que son los corpus lingüísticos.

b) Analizar las necesidades del corpus en ingeniería para el GIL y de los objetivos que debe de cubrir el desarrollo de este corpus.

c) Identificar los documentos que pertenecerán al corpus.

d) Diseño y desarrollo de etiquetas que identifiquen cada uno de los datos contenidos en: documentos y los textos que los conforman. Para los segundos es importante identificar su:

- Resumen,
- Bibliografía,
- Notas a pie de página,
- Cambio de tipo de letra,
- Itálicas,
- Negritas,
- Subrayado,
- Apéndices,
- Figuras,
- Tablas,
- Capítulos,
- Subíndice,
- Superíndice, etc.

Mientras que para identificar los elementos de un documento, es necesario distinguir e identificar las características fundamentales que los diferencian, como por ejemplo:

- Nombre del autor,

- Título,
- Subtítulo,
- Editorial, etc.

e) Definir una herramienta necesaria que permita el desarrollo de las etiquetas de documentos y el texto de éstos.

1.4. Estructura de la tesis

La tesis está estructurada en 8 capítulos, seis de los cuales, son parte principal y fundamental de este trabajo. El segundo capítulo (Generalidades de corpus Lingüísticos), constituye una introducción de corpus lingüísticos en general: abarca desde su definición, características que lo rodean, tipos y clasificación, hasta ejemplos de ellos que en la actualidad se usan para investigaciones; todo esto, para dar a conocer el campo de desarrollo de los corpus. Así mismo, este capítulo incluye los diversos usos y aplicaciones de éstos.

Una vez teniendo los conceptos fundamentales de corpus lingüísticos, el tercer capítulo (el corpus de ingeniería), se centra en la importancia y los objetivos fundamentales para el desarrollo del corpus de ingeniería, del Grupo de Ingeniería Lingüística, de los cuales se desprenden los requerimientos del etiquetado para el corpus. En este capítulo, se define la clasificación del corpus, abarcando los aspectos técnicos que van desde su arquitectura general hasta el control de calidad, validación, mantenimiento, actualización, diseño de búsqueda y de salida, tipos de reportes, procedimientos para la formulación de búsquedas, formas de interacción con el usuario final, operación y los criterios generales para la selección de la información.

El cuarto capítulo (Introducción a XML), se explica los fundamentos de este lenguaje para la representación de etiquetado, las diferencias que existen con HTML, los objetivos que persigue XML, los diferentes usos en que se puede utilizar este lenguaje. Se da brevemente la sintaxis de éste y la definición de los elementos que lo conforman.

El quinto capítulo (Etiquetado de documentos), se definen las etiquetas para identificar los datos de un documento, los cuales son: autor, título, subtítulo, editor, etc., de libros, memorias, informes y revistas que se requieren para el desarrollo corpus.

El capítulo seis (Etiquetado de textos), muestra las etiquetas que se definieron para identificar los patrones metalingüísticos en el texto (letras itálicas, viñetas, negritas, fines de párrafo, etc.). Las etiquetas se clasificaron en cuatro secciones: estructura, formato, énfasis y notaciones. Para ello fue necesario identificar los requerimientos de aplicación para el GIL.

El capítulo siete (Administración interna de los documentos), se centra en el desarrollo de una base de datos que permite el control interno de los documentos en el GIL, lo cual permitirá identificar que documentos ya fueron recabados, escaneados, el usuario que fue el encargado de la digitalización, y toda aquella información que es de interés particular para el grupo. Cabe mencionar que sólo se explica la interfaz del usuario ya que por no ser tema principal de este trabajo, sólo se tiene como objetivo dar a conocer que se elaboró esta herramienta en el tiempo de investigación para el desarrollo de este tesis.

CAPITULO 2

GENERALIDADES DE CORPUS LINGÜÍSTICOS

En este capítulo se hablará de la definición de un corpus lingüístico, así como de las características que debe cumplir un corpus, sus aplicaciones y sus diversos usos, y se darán algunos ejemplos de corpus que actualmente existen para delimitar el corpus de ingeniería, que es el tema central de esta tesis.

2.1. ¿Qué es un corpus lingüístico?

Es importante hablar primeramente de cuál es el significado de un corpus lingüístico para llegar así a comprender y analizar el tema central de este trabajo. Tony McEnery y Andrew Wilson en su libro "Corpus Linguistics" definen un corpus como: "In principle, any collection of more than one text can be called a corpus, (corpus being Latin for "body", hence a corpus is any body of text). But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition."¹

Un corpus lingüístico consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados sobre una misma área para realizar ciertos análisis lingüísticos; estos textos deben ser representativos y se recogen según criterios lingüísticos para poder ser utilizados en el análisis.

De hecho, se usan a menudo textos individuales para muchos tipos de análisis literario y lingüístico, como - el análisis estilístico de un poema, o un análisis de la conversación de una muestra de charla de tv.

Un corpus es una colección de textos en soporte informático, que llega a ser muy extensa, de varios millones de palabras. Los corpus, manejados con programas informáticos apropiados, nos proporcionan un excelente material para el trabajo de investigación.

Cada vez es más extendido el uso de corpus hoy en día. Se han desarrollado diversos corpus para distintas áreas en particular, y actualmente cada vez es mayor el uso de las computadoras por la gente. Hoy contamos con algunos corpus muy potentes, en Internet, de donde los podemos consultar gratuitamente por nuestra computadora.

¹ McEnery Tony, Wilson Andrew, "Corpus Linguistics" Edinburg University Press Koinonia, Manchester 2001. Publicación electrónica: <http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>

2.2. Características que debe cumplir un corpus

Entendiendo ahora el término de corpus lingüístico, se mencionarán las características que debe un cumplir un corpus.

José Manuel Blecua, et al, en su libro "Filología e informática", contemplan que el corpus debe tener las siguientes características²:

- Los corpus deberán estar compuestos de datos reales.
- El corpus tiene que mostrar a pequeña escala como funciona una lengua natural.
- Debe de ser selectivo, ya que no es posible recopilar todo lo escrito y/o hablado de una lengua. Debe de ser representativo.

Tony McEnery y Andrew Wilson en su libro "Corpus Linguistics"³ proponen cuatro características que debe de reunir un corpus:

- **Representativo.**- básicamente esto se refiere a la representatividad del corpus, delimitar la información contenida en el corpus, los textos que conformarán el corpus deben ser representativos del tema de estudio que se llevará a cabo.
- **Tamaño finito.**- el corpus debe ser finito, por ejemplo, puede ser de 1,000,000 de palabras. La ventaja del tamaño finito es que no son estáticos, sino que pueden irse agregando nuevos textos al corpus, así mismo, mantienen una muestra mayor del tema que se trata.

² Blecua José Manuel, Clavería Gloria, et al "Filología e Informática" ed. Nuevas tecnologías en los estudios filológicos. Seminario de filología e informática. Barcelona 1996, pags. 45,46.

³ McEnery Tony, Wilson Andrew. "Corpus Linguistics" Edinburg University Press Koinonia, Manchester 2001 pags 29-32.

- **Manejable por la computadora.**- actualmente contamos con corpus electrónicamente diseñados, ya que anteriormente los corpus estaban desarrollados en forma impresa, como por ejemplo "A Corpus of English Conversation" (Svartvik and Quirk 1980). Las ventajas que se tienen de tener corpus en formato electrónico sobre en forma impresa son: que se pueden manipular más fácilmente y a mayor velocidad; pueden enriquecerse con información extra.
- **Referencia estándar.**- el desarrollo de nuevos corpus debe llevarse a cabo con estándares, con una sola metodología para obtener la elaboración de cualquier corpus.

Para complementar las características de un corpus, conviene observar las siguientes:

- El corpus debe tener un soporte electrónico.
- Debe facilitar al usuario la consulta por medios electrónicos, que en la mayoría de los casos es una conexión directa al corpus a través de Internet.
- Debe cumplir con reglas establecidas para el acceso a la información, derechos de autor para evitar el mal uso del corpus.
- Debe ser de interés para alguna área.

2.3. Tipos de corpus y clasificación

Una vez que se ha explicado el término de corpus lingüístico y las características que debe cumplir, conviene conocer los diferentes tipos y clasificaciones de los corpus que proponen los autores consultados, para poder llegar así a definir el corpus de ingeniería.

Se puede hablar de dos diferentes tipos de corpus según el origen, en corpus textuales y corpus orales. Los corpus textuales consisten, como su nombre lo indica, todo lo relacionado a lo que está escrito, mientras que los orales es todo lo relativo a lo que esté en transcripciones ortográficas de la lengua hablada; como ejemplo de estos últimos podemos mencionar alguna grabación hecha con su respectiva transcripción.

En la figura 2.1 se muestra un cuadro sinóptico para facilitar la distinción de los tipos y clasificación de los corpus.

Existen también clasificaciones de los corpus según la especificidad de los textos. Así tendremos corpus generales y corpus especializados o también llamados específicos. Los primeros se encargan de recoger todo tipo de géneros y son útiles para describir la lengua común de una colectividad. Los corpus especializados, al contrario de los generales, recogen material que puedan aportar datos para la descripción de un área o tema en particular.

En lo referente a la clasificación según el lenguaje, existen dos tipos de corpus, el corpus monolingüe y el corpus multilingüe. El primero se refiere que utiliza un solo idioma, como por ejemplo, puede ser un corpus de lengua inglesa, española, etc. Mientras que el segundo a diferencia del primero hará referencia a más de una lengua.

Según la cantidad de texto que se recoge de cada documento tenemos corpus textual y corpus de referencia. El corpus textual es el que recoge íntegramente los documentos que lo componen. El corpus de referencia es aquel que solo toma fragmentos de los documentos, en este tipo de corpus es muy importante los aspectos de equilibrio y representatividad cuando se hace la selección de los fragmentos. Un corpus de referencia es aquel que está diseñado para proporcionar información exhaustiva acerca de una lengua en un momento

determinado de su historia y, por tanto, ha de ser lo suficientemente extenso para representar todas las variedades relevantes de la lengua en cuestión⁴.

También existe una clasificación de corpus según la codificación y anotación donde encontramos el corpus simple y el corpus codificado o anotado. El corpus simple es el que ha sido guardado en un formato ASCII y que no tiene una codificación para ninguno de sus aspectos. Mientras que el corpus codificado o anotado, es aquel corpus que está formado por textos a los cuales se ha añadido electrónica o manualmente, etiquetas para reconocer algunos de sus elementos en los documentos.

2.4. Ejemplos de corpus

Una vez conociendo los tipos de corpus que existen hoy en día, se mencionarán algunos ejemplos de corpus, así como en la clasificación en que se encuentran cada uno de ellos.

a) Corpus Diacrónico del Español (CORDE)⁵

El Corpus diacrónico del español (CORDE) es un corpus textual de todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1975, en que limita con el *Corpus de referencia del español actual*. El CORDE está diseñado para extraer información con la que estudiar las palabras, sus significados, la gramática y su uso a través del tiempo.

⁴ <http://www.rae.es/>

⁵ <http://www.rae.es/>

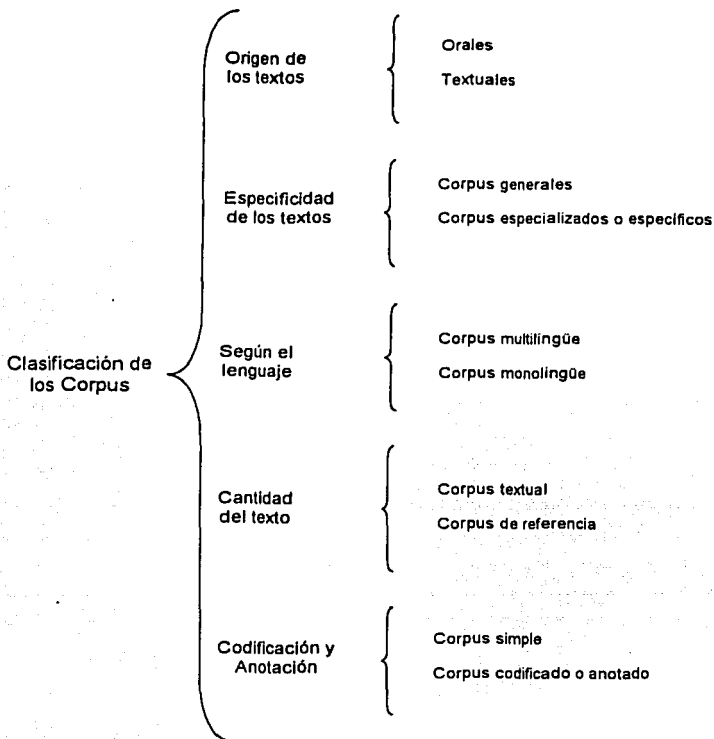


Figura 2.1. Cuadro sinóptico de la clasificación de los corpus que existen

Su andadura comenzó en 1994, cuando la Real Academia Española se planteó la posibilidad de aplicar las nuevas técnicas informáticas para construir un banco de datos que mejorara la calidad de sus materiales de trabajo y el acceso a estos. Hasta octubre de 2001 cuenta en la actualidad con más de 136 millones de

registros. Este volumen de información es el mayor conjunto de registros de la historia de la lengua española.

El corpus recoge textos escritos de muy diferente género. Se distribuyen estos en prosa y verso y, dentro de cada modalidad, en textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc. Se pretende recoger todas las variedades geográficas, históricas y genéricas para que el conjunto sea suficientemente representativo.

Hoy es fuente obligada para cualquier estudio diacrónico relacionado con la lengua española. La Academia utiliza sistemáticamente el *CORDE* para documentar palabras, para calificarlas de anticuadas o en desuso, para saber el origen de algunos términos, su tradición en la lengua, primeras apariciones de palabras. Pero uno de los cometidos fundamentales del Corpus diacrónico será servir de material básico para la confección del Diccionario histórico.

El *CORDE* es un corpus textual ya que solo recogerá textos, también es un corpus general ya que abarcará varios temas como los históricos, jurídicos, etc., es un corpus monolingüe ya que solo será para el español.

b) Corpus de Referencia del Español Actual (CREA)⁶

El Corpus de referencia del español actual (CREA), constituido por la Real Academia Española, es un banco de datos del español contemporáneo, es decir, un conjunto de textos de diversa procedencia, almacenados en soporte informático, del que es posible extraer información para estudiar las palabras, sus significados y contextos.

El *CREA* cuenta hasta octubre de 2001 con 130 millones de registros, que está previsto vayan aumentando hasta conseguir al menos 160 millones, a finales

⁶ <http://www.rae.es/>

de 2004. Se compone de una amplia variedad de textos escritos y orales, producidos en todos los países de habla hispana desde 1975 hasta la actualidad. Los textos escritos, procedentes tanto de libros como de periódicos y revistas, abarcan más de cien materias distintas. La lengua hablada está representada por transcripciones de documentos sonoros, procedentes, en su mayor parte, de la radio y la televisión.

El *CREA* es, hoy por hoy, la única herramienta lingüística de gran magnitud existente para lengua española. Debe ser el punto de partida forzoso para investigaciones de diverso tipo, principalmente aquellas estrictamente lingüísticas, pero también pertenecientes a campos tan dispares como el de la publicidad, la terminología o la sociología, así como para la elaboración de una enorme cantidad de productos derivados: gramáticas, diccionarios, tesauros, correctores ortográficos, métodos de didáctica del español, desarrollos informáticos de traducción automática, etc.

El *CREA* es un corpus general, un corpus de referencia, es un corpus especializado ya que se quiere extraer información para estudiar las palabras, sus significados y contextos, es un corpus monolingüe por que solo será del español.

c) Archivo de textos hispánicos (ARTHUS)⁷

El *Archivo de textos hispánicos de la Universidad de Santiago de Compostela* contiene en la actualidad textos pertenecientes a diferentes etapas de la historia del español. Todos ellos han sido introducidos en ordenador mediante escáner y programas de reconocimiento óptico de caracteres, están en formato ASCII y tienen una codificación mínima en formato COCOA que permite, con los programas de recuperación adecuados, conocer texto, página y línea en que se encuentran los ejemplos buscados.

⁷ <http://www.sintx.usc.es/>

La parte contemporánea comprende en la actualidad treinta y cuatro textos narrativos, teatrales, ensayísticos, periodísticos y orales procedentes de España e Hispanoamérica con un total aproximado de 1,450,000 formas.

El corpus incluye textos de diferentes períodos de la historia de la lengua española y correspondiente a distintos géneros literarios y periodísticos, así como también transcripciones de textos orales.

ARTHUS es un corpus textual, así mismo es un corpus específico ya que solo contiene textos históricos, es un corpus monolingüe por que solo abarca al español, es un corpus codificado.

d) CRATER⁸

El proyecto europeo CRATER (Corpus Resources and Terminology Extraction) es un corpus de textos técnicos consistente en textos de la ITU (International Telecommunications Union), etiquetado morfológicamente e incluye el alineamiento de frases con sus equivalentes francés e inglés; este corpus está disponible en la Universidad Autónoma de Madrid. Para este proyecto fue creado un etiquetador part-of-speech en español.

Generado este recurso el proyecto rectificó errores en las versiones inglesas y francesas una vez existente el corpus, así mismo los errores alejados del corpus en español. El corpus contiene un millón de palabras que ha demostrado ser de beneficio para proyectos en el área de traducción automática, la lingüística computacional y corpus en general.

Una aportación final del proyecto fue el de crear un juego de herramientas para la recuperación del corpus, y examinar las alineaciones de términos o palabras entre los distintos idiomas que lo conforman.

⁸ <http://www.htcentral.org/projects/CRATER>

El corpus CRATER es un corpus textual, así como, un corpus específico ya que sólo fue creado para la ITU, es un corpus multilingüe ya que abarca tres lenguajes: español, inglés y francés; es un corpus codificado ya que fue etiquetado morfológicamente.

e) Proyecto Corpus: Corpus textual especializado plurilingüe⁹

El proyecto Corpus es el proyecto de investigación prioritario del IULA (Instituto Universitario de Lingüística Aplicada) de la Universidad Pompeu Fabra, Barcelona. Recopila textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de especialidad de la economía, el derecho, el medio ambiente, la medicina y la informática. A través del establecimiento del corpus, se intentan inferir las leyes que rigen el comportamiento de cada lengua en cada área. Este corpus es el soporte principal de las actividades de investigación y docencia del IULA.

Las investigaciones previstas sobre el corpus son las siguientes: detección de neologismos y términos, estudios sobre variación lingüística, análisis sintáctico parcial, alineación de textos, extracción de datos para la enseñanza de segundas lenguas, extracción de datos para la construcción de diccionarios electrónicos, elaboración de tesauros, etc.

Los textos son seleccionados por especialistas de cada área y agrupados sobre la base de una clasificación temática y de uso propuesta por los propios especialistas (derecho, economía, medio ambiente, medicina e informática). Posteriormente los textos son marcados de acuerdo con el estándar SGML y siguiendo las directrices marcadas por el "Corpus Encoding Standard" (CES) de la iniciativa EAGLES.

El procesamiento de los textos del corpus sigue los siguientes pasos:

⁹ <http://www.iula.upf.es/corpus/corpuscs.htm>

- marcaje estructural
- preproceso (detección de fechas, números, locuciones, nombres propios...)
- análisis y marcaje morfológicos de acuerdo con los etiquetarios morfosintácticos diseñados en el IULA.
- desambiguación lingüística y/o estadística
- almacenamiento en una base de datos textual

El proyecto corpus es un corpus textual, también es un corpus específico debido a que abarca sólo áreas de especialidad como la economía, el derecho, el medio ambiente, la medicina y la informática. Es un corpus multilingüe porque abarcará catalán, castellano, inglés, francés y alemán. Es un corpus codificado porque es con una anotación estructural.

f) Archivo Gramatical de la Lengua Española¹⁰

El Archivo Gramatical de la Lengua Española (AGLE), orientado hacia los estudios gramaticales, está constituido por más de 100,000 citas recogidas por el gramático español Salvador Fernández Ramírez (1896-1983). Actualmente se está editando y anotando en el Instituto Cervantes.

El *Archivo* en su primera entrega consta de unos 75 ficheros, cada uno de los cuales contiene alrededor de 1,500 fichas a las que el autor se refiere siempre como *cédulas*. No todos los ficheros poseen el mismo grado de ordenación interna ni todos poseen una articulación similar. Los ficheros seguían aproximadamente el orden que el autor tenía previsto para su *Gramática*, pero aun así eran muy numerosas las fichas que se agrupaban en apartados como VARIOS o SIN CLASIFICAR. La intención de este archivo es de respetar las clasificaciones establecidas, ordenar las partes menos articuladas, clasificar las fichas que el autor no llegó a ordenar, y completar, sin añadir ni una sola papeleta, los bloques

¹⁰ <http://cvc.cervantes.es/obref/agle/prologo/>

temáticos existentes tomando siempre como guía el criterio que hubiera sido el de su gramático.

El archivo gramatical de la lengua española dentro de la clasificación de un corpus se concluye que es un corpus textual y un corpus especializado.

g) Base de datos ETDEWEB¹¹

La base de datos ETDEWEB contiene la colección más grande del mundo de la literatura sobre energía. Con más de 3.8 millón archivos abstraídos. El banco de datos contiene referencias bibliográficas y artículos de periódico, informes, conferencias, libros, y otros tipos de documentos. El banco de datos cubre varios aspectos medioambientales del uso y producción de energía y políticas de energía y planeación de ésta, así como las ciencias básicas que apoyan investigación de energía y desarrollo.

El banco de datos contiene citas publicadas mundialmente considerando áreas como la: nuclear, carbón, y la información de cambio de clima global. Los usuarios del Banco de datos de Energía de ETDEWEB son tan diversos como los temas que se cubren en ésta: científicos, ingenieros, bibliotecarios, líderes de industria, y estudiantes. El Banco de datos de Energía de ETDEWEB está disponible a cualquier país miembro de ETDEWEB (México, Estados Unidos, Japón, entre otros) y para cualquier organización, biblioteca, o institución de algún país miembro. Esto beneficia llevando oportunidades inestimables para aquéllas áreas comerciales y académicas, así como para organizaciones gubernamentales.

ETDEWEB está públicamente disponible vía Internet como ETDEWEB, y en varios formatos a través de los organizadores del online comerciales y en productos de CD-ROM.

¹¹ <http://www.ctde.org/ctdeweb/>

El banco de datos ETDEWEB es un corpus textual, lo referente a la especificidad de los textos es un corpus especializado o específico, ya que solo trata temas relacionados con la energía.

2.5. ¿Por que usar un corpus?

Actualmente es más evidente la utilización de recursos informáticos para llevar a cabo las investigaciones humanísticas. Para poder llevar a cabo la utilización de este tipo de recursos es necesario contar con el material donde aplicarlos, este material son los textos orales o escritos y los documentos que los contienen, como ya se ha mencionado anteriormente, debidamente recopilados estos documentos y textos forman los corpus.

Hoy en día, con la ayuda de la informática es posible el análisis de corpus, ya que permite llevar a cabo cálculos complejos en cuestión de segundos, y sin error alguno. Así por ejemplo, se puede dar una instrucción a la computadora para que recupere todas las ocurrencias (una ocurrencia se refiere al número de veces que una palabra aparece en un texto, por ejemplo, "de", "para", "el", "la", etc.) de una palabra junto con las palabras que se encuentran a su derecha o izquierda. Así como se le puede solicitar a la computadora que presente los resultados con el contexto en donde aparece la palabra de interés, además se puede especificar un cierto orden (normalmente en forma alfabética). Esta presentación de las palabras se conoce con el término de concordancias, y su utilidad ha sido reconocida como medio de investigación lingüística para estudios en las áreas de lexicología, gramática, semántica, pragmática, estilística, y dialectología, entre otros. Otro posible uso de un corpus de textos es encontrar cada una de las apariciones en las que una palabra desempeña una función determinada o pertenece a una categoría gramatical concreta, lo que facilita la investigación de la variación sintáctica existente en la lengua.

Es imprescindible, dentro de un análisis lingüístico, identificar y detallar aquello que es útil para el fin de la investigación; por ello, y con la ayuda de la información que proporciona un corpus, al investigador le es posible la realización de tablas de identificación de palabras significativas, eliminando elementos que nada aportan al análisis, por ejemplo, efectuar análisis lingüístico, relación entre palabras, entre otros procesos.

Hoy en día, muchas áreas de la lingüística pretenden trabajar con datos reales y lo más completos posibles que permitan reproducir las características del objeto de estudio. El auge que actualmente ha tenido la aplicación informática en cualquier campo de la investigación ha facilitado las tareas de recopilación y organización en formato electrónico de los textos, lo que ha permitido que el investigador pueda encontrar grandes cantidades de documentos y la organización de los datos de éstos.

Los corpus informatizados han demostrado ser una herramienta excelente para muchas investigaciones, principalmente en el campo de la lingüística, ya que proporcionan bases mucho más reales para el estudio de las lenguas que los métodos intuitivos tradicionales.

Los corpus informatizados, así mismo, han influido y cambiado bastante los métodos de investigación e, incluso, han propiciado el nacimiento de nuevas tendencias lingüísticas. Muchos trabajos que antes tenían que desarrollarse a mano, ocupando mucho tiempo y esfuerzo estudiando, leyendo y repasando los textos para encontrar datos concretos que sirvieran para demostrar hipótesis, hoy en día, con la ayuda de la informática, se pueden hacer todo esto en menos tiempo de igual forma en forma más ordenada y exhaustivamente, es decir, con mayor eficacia y eficiencia y una delimitación del campo de estudio, además de la potencialidad que representa el uso y explotación de los recursos que el corpus es capaz de realizar.

Las ventajas de trabajar con corpus informatizados, sobre todo con los que están anotados o codificados, es tan grande, que está obligando a los lingüistas tradicionales a trabajar conjuntamente con lingüistas computacionales.

2.6. Aplicaciones diversas

El uso de corpus es de interés en los estudios del lenguaje, ya que proporciona datos empíricos para realizar análisis objetivos sobre un estado de la lengua en particular. Entre los intereses de los corpus para estudios lingüísticos, cabe mencionar:

Lexicografía

La lexicografía hace uso de los corpus para desarrollar diccionarios de lengua. El interés de la lexicografía en los corpus es también el de saber cuantas y cuales palabras se utilizan en el corpus.

Los corpus son de gran ayuda para configurar el vocabulario de los diccionarios, ya sea para incluir nuevas palabras como para quitar las que ya no se usen. Así como para separar las distintas acepciones de cada vocablo para detectar las palabras co-ocurrentes, las combinaciones sintácticas, etc.

El lingüista que tiene acceso a un corpus lingüístico puede llamar a todos los ejemplos de una palabra o de una frase de muchos millones de palabras de texto en algunos segundos. Los diccionarios se pueden producir y revisar mucho más rápidamente que antes, proporcionando así la información actualizada sobre el propio lenguaje. También, las definiciones pueden ser más completas y exactas puesto que un número más grande de ejemplos naturales se examina.

La recopilación (constantemente creciente) permite a lexicógrafos guardar las nuevas palabras que incorporan el lenguaje, o de palabras existentes que cambian sus significados, o el equilibrio de su uso según el género, etc.

En las investigaciones lingüísticas sobre lexicografía, como ejemplo del uso de la lexicografía cabe mencionar el Diccionario del Español de México (DEM) de Luis Fernando Lara¹² el cual es una obra lexicográfica cuyo objetivo fundamental es reflejar el léxico del español utilizado actualmente en el país, en cuanto "lengua nacional" y en cuanto a sus modalidades escritas y orales, cultas y coloquiales, urbanas y rurales.

Este trabajo se refiere también a los problemas de la objetividad en la descripción del léxico mexicano y al de la cantidad de datos necesaria para la labor de lexicografía. Particularmente se ocupa de la aplicación de la estadística lexicológica en la investigación del español de México como el mejor instrumento de documentación y análisis del vocabulario.

Para el desarrollo del DEM se consideró que el método estadístico era el único capaz de dar los registros necesarios y la cantidad de datos suficientes para la tarea lexicográfica de un modo objetivo e imparcial. Del análisis estadístico del corpus de datos para el DEM quisieron obtener:

- a) Un número elevado de vocablos que puedan constituir la mayor parte de las entradas del diccionario.
- b) Una base imparcial de selección de vocablos para la primera edición del DEM.
- c) Un punto de referencia que permita detectar los usos diferentes de los vocablos en la sociedad mexicana.

¹² Lara Luis Fernando, et al. "Investigaciones lingüísticas en lexicografía". El Colegio de México, Jornadas 89. México 1979. pp. 7-83.

Estas tres necesidades los colocan frente a frente, por una parte, con lo que significa un corpus para la lexicografía y la lingüística y, por la otra, con la concepción del corpus para la estadística.

El DEM se define como un diccionario sincrónico, descriptivo y selectivo. Interesa mostrar en él el léxico del español que se utiliza entre las fronteras de México y, a diferencia de los diccionarios de regionalismos, es un diccionario regional de la común lengua española.

Terminología

La terminología utiliza los corpus para desarrollar diccionarios especializados, de igual forma, tiene interés particular en el corpus para hacer comparaciones de cuales palabras se utilizan en un corpus y en otro. La terminología lee todo el corpus y lo toma completo, a diferencia de la lexicografía que solo toma una muestra de éste.

Uno de los puntos que permiten establecer diferencias claras entre el lenguaje común y el especializado, como entre los distintos lenguajes especializados entre sí, es el uso de una terminología específica. La terminología desempeña un papel fundamental para caracterizar el lenguaje especializado, y para establecer y clasificar los distintos lenguajes de especialidad en un corpus lingüístico.

El proceso de trabajo en la terminología sistemática consiste en elaborar la lista de los términos del área que se ha delimitado previamente, e informarlos de acuerdo con las características del trabajo que uno se propone llevar a cabo.

De esta forma, la concisión de la terminología y la frecuencia con que aparecen formantes cultos y el carácter internacional de los términos, obtenibles

de los corpus, favorecen el aspecto conciso de los textos especializados y facilitan enormemente la comunicación internacional, tan importante para los especialistas.

De entre las técnicas de obtención de repertorios terminológicos, cada vez gana más adeptos la de recurrir a la extracción automática de terminología a partir de *corpus*. Para ello dichos *corpus* deben cumplir ciertas condiciones, que nos permiten hablar de "validación".

Para que la extracción automática de terminología de resultados válidos, debe efectuarse sobre un corpus de textos representativo (como ya se ha mencionado arriba) del ámbito en cuestión: por ejemplo, para estudiar los términos de la odontología, habrá que acumular desde artículos científicos y tesis doctorales hasta historias clínicas y facturas del material usado en su práctica profesional y manuales de instrucciones del sofisticado instrumental que tienen. Por el contrario, un repertorio de términos obtenido con metodología "clásica" puede, especialmente en terrenos científicos o profesionales sujetos al trepidante ritmo de las innovaciones tecnológicas, quedar desfasado antes de ver la luz.

Confección de herramientas lingüísticas informatizadas

Otro campo en el que los corpus aportan grandes ventajas es el de la confección de herramientas lingüísticas informatizadas. Dentro de ésta una de la más importante es la de los diccionarios-máquina, de usos tan diversos como son la corrección de textos informatizados o la segmentación de palabras por sílabas. Este tipo de herramientas son de gran importancia para traducción automática del lenguaje y otras tareas que se basan en el tratamiento automático del lenguaje.

Fonética

La fonética se encarga del estudio de cómo se producen los sonidos. En el terreno de la fonética, los corpus que están constituidos por grabaciones de laboratorio, son herramientas imprescindibles para el estudio experimental del habla (lenguaje), mientras que los que contienen registros menos formales son necesarios para caracterizar diversos estilos. En el ámbito de las tecnologías del habla, las bases de datos orales proporcionan datos importantes en la conversión de texto a habla y son esenciales para el entrenamiento y la validación de los sistemas de reconocimientos y de diálogo en entornos de comunicación persona a máquina, cuyas aplicaciones se extienden desde la oferta de servicios telefónicos automatizados hasta las ayudas para personas con alguna discapacidad.

Enseñanza del lenguaje

Los corpus así mismo también proporcionan elementos muy importantes y útiles en el campo de la enseñanza de lenguas, sobre todo a la hora de preparar materiales o ejercicios de trabajo en clase basados en un uso real de la lengua. Del contenido de los corpus puede desprenderse información tanto de uso (palabras y construcciones más frecuentes en los libros de textos y lecturas recomendadas en relación con los materiales auténticos) como de corrección de barbarismos o malos usos lingüísticos (errores más repetidos, construcciones no normativas, léxico mal usado, grafías incorrectas, etc.). La recopilación de corpus de producciones de estudiantes de lengua extranjera constituye también una fuente de datos sobre la interferencia entre la primera y la segunda lengua en todos los niveles del análisis lingüístico y una base empírica importante para el análisis de errores y de las estrategias comunicativas de los alumnos.

Semántica

La semántica estudia el significado de las palabras. Los corpus lingüísticos han ayudado a la semántica a establecer un acercamiento en su objetividad. Demuestran cómo una recopilación se puede utilizar para proporcionar criterios en los objetivos para asignar significados a los términos lingüísticos. Con frecuencia, en la semántica, los significados de términos son descritos por la referencia a propias intuiciones del lingüista.

Otro papel de los corpus en la semántica ha consistido en establecer más firmemente las nociones de categorías y de los contextos no muy claros. Se desea lograr, con los corpus lingüísticos, que las distinciones semánticas estén asociadas de tal manera que los textos estén relacionados con los contextos observables característicos (sintácticos, morfológicos y prosódicos).

Sociolingüística

La sociolingüística le interesa saber como se habla coloquialmente, como por ejemplo, como hablan los hombres y mujeres. En la sociolingüística también de los corpus se pueden obtener datos de gran utilidad; al contrario que a los que estudian la historia, los sociolingüistas no les interesa tanto el tema del texto o el nombre del autor como la clase social, el sexo o el nivel cultural del receptor. En la sociolingüística los corpus se utilizan como base de estudios dedicados a la diferenciación entre registros o estilos (por ejemplo, entre la lengua escrita y la oral o entre diversos géneros como la correspondencia privada, el discurso jurídico, político, publicitario o religioso, incluyendo incluso trabajos sobre las características de los mensajes de correo electrónico), asociados a variaciones en la situación de comunicación y a dimensiones como el grado de formalidad, el carácter público o privado, etc. Este tipo de trabajos entroncan directamente con

los realizados desde la perspectiva del análisis del discurso, encaminados a establecer tipologías textuales.

Psicolingüística

La psicolingüística se encarga del estudio de los errores del lenguaje, y de cómo se produce desde el habla infantil. En el terreno de la psicolingüística es beneficiada por el uso de los corpus, especialmente en campos como son el análisis de los errores de producción del habla o el desarrollo del lenguaje infantil. En el análisis de patologías del lenguaje y del habla requiere igualmente colecciones sistemáticas de muestras recogidas de personas que presentan trastornos de la comunicación.

Literatura

Los que se dedican al estudio de la literatura pueden tener en los corpus una buena herramienta para realizar sus investigaciones. En el campo de la estilística, por ejemplo, los corpus pueden ayudar a definir los trazos que caracterizan distintos estilos literarios o, en el terreno de la estilometría, los análisis estadísticos del uso de las palabras en los textos pueden dar luz a problemas de adscripción de trabajos de dudosa autoría.

Otros campos de las humanidades

En cuanto a las utilidades de los corpus en otros campos de las humanidades que no sean estrictamente lingüísticos cabe mencionar las posibilidades que ofrecen para los estudios históricos, para los de la teoría de la literatura, etc. Si los textos que componen un corpus están asociados a una

documentación detallada de sus rasgos externos: fecha, tema, región, edad del autor, estatus social, sexo, etc., éstos pueden convertirse en fuente de datos para aquellas personas interesadas en los aspectos de contenido textual los historiadores, por ejemplo, pueden seguir la evaluación de opiniones e ideas mediante el estudio de palabras o frases asociadas a ellas.

CAPITULO 3

EL CORPUS DE INGENIERÍA

Una vez desarrollados los antecedentes y generalidades de los corpus lingüísticos, descritos en el capítulo anterior, se expondrán la definición, objetivos y necesidades que se tienen para realizar el primer corpus en ingeniería. Cabe mencionar que este capítulo está basado en un proyecto de investigación aprobado por el CONACYT, en 2001, y que se está realizando en el Instituto de Ingeniería por el Grupo de Ingeniería Lingüística (GIL).

El proyecto del GIL consiste en elaborar un corpus sobre las diferentes áreas de la ingeniería, y busca abarcar todas sus ramas: eléctrica, electrónica, civil, mecánica, etc. El corpus estará disponible para su consulta en internet, de modo que podrán ser utilizados con fines de investigación para extraer información lingüística en el área de ingeniería.

3.1 Objetivo general

El objetivo general para la creación del corpus de ingeniería es: *elaborar, desarrollar y mantener* un corpus lingüístico multipropósito de textos selectos en el área de ingeniería, debidamente *codificados y organizados*, manejando las *herramientas de programación* adecuadas para poder utilizar el corpus en el desarrollo de diversas investigaciones en las áreas de lingüística e ingeniería lingüística.

Se elaborará y desarrollará el primer corpus de ingeniería, ya que como se ha mencionado, no existe hoy en día un corpus en esta área. El corpus se mantendrá disponible para su consulta a través de Internet y se actualizará periódicamente. Los textos en el área de ingeniería serán codificados y organizados, es decir, se etiquetarán debidamente en XML y se organizarán en base a las necesidades que se describen más adelante.

3.2. Importancia del corpus

Para conocer los objetivos específicos y la importancia del corpus, es necesario, en primera instancia, conocer los objetivos que pretende el GIL con el corpus. Para ello, conviene recordar que el GIL tiene como proyecto central de investigación (como se mencionó en el capítulo 1), el desarrollo de un diccionario de búsqueda onomasiológica¹. El diseño y creación de este diccionario comprende cinco fases bien definidas: adquisición de datos, creación de bases de datos y captura de información, determinación de paradigmas semánticos, diseño del motor de búsqueda y diseño de la interfaz del usuario. De este proyecto se desprende el desarrollo del corpus lingüístico y de otras líneas de interés para el GIL.

¹ Baldinger K. 1970. Teoría semántica: Hacia una semántica moderna. Madrid: Ed. Alcalá.

La importancia del corpus radica en que permitirá concretar los estudios que se han venido realizando dentro de GIL, tales como:

- *Extracción automática de términos en un área de especialidad*². El corpus permitirá la extracción de términos mediante su consulta por internet en el área de ingeniería.
- *Identificación de contextos definitorios para extracción de conceptos de textos especializados*³. Para este trabajo se considera a la tipografía como una parte para desarrollar una metodología de extracción terminológica, ya que las etiquetas son parte del texto y pueden ser una base importante para empezar a buscar términos.
- *Elaboración de diccionarios especializados, tanto del tipo semasiológico como onomasiológico*⁴. Un sistema de búsqueda onomasiológica, significa un diccionario que permita la búsqueda de términos a partir de la descripción del concepto mediante el uso de lenguaje natural.
- *Identificación de paradigmas semánticos*.⁵

El corpus necesita y se alimenta de las dos primeras, en tanto que las dos últimas se derivan de las primeras. Debido a la importancia que presentan los contextos definitorios en el etiquetado del corpus (objetivo central a desarrollar en esta tesis), en el siguiente apartado se describen a más detalle lo relacionado con éstos.

Así mismo, el corpus en ingeniería dará la posibilidad, dada su importancia, de realizar diversos estudios no trabajados a la fecha, tal como:

² Reyes Pérez Antonio "Hacia una obtención computarizada de términos. (aplicación concreta al léxico de la física en el nivel bachillerato). Tesis. México 2002.

³ Sierra, G. Alarcón, R (2002). "Hacia la extracción automática de conceptos"; en *La Terminología: entre la globalización y la localización*, RI Term, Cartagena, Colombia, formato CD-Rom.

⁴ Sierra G. and McNaught John. "Design of an onomasiological search system: A concept oriented tool for terminology" *Terminology* Vol. 6(1), 2000, pp. 1-34.

⁵ Castillo Hernández Gabriel. "Algoritmo revisado para la extracción automática de agrupamientos semánticos". Tesis. México 2002.

- Elaboración de terminologías en el área de ingeniería.
- Identificación de redes semánticas en ingeniería.
- Elaboración de herramientas para redacción de documentos técnicos (artículos, informes, etc.).
- Corrección de ortografía de términos técnicos.
- Vacío en corpus de ingeniería: primer corpus lingüístico en el área de ingeniería.

Con el corpus en ingeniería se permitirá realizar análisis lingüístico a la información contenida en éste a través del Internet. El corpus está dirigido a investigadores, lingüistas y toda aquella persona que esté interesada en la información contenida en el corpus, pero sin hacer disponible éste, debido a las restricciones de derecho de autor. En aquellos casos en que no existan restricciones será posible tener acceso a los textos completos, con lo que la comunidad científica de México y del mundo podrán ver los avances en el campo de la ingeniería.

3.2.1. Descripción de contextos definitorios⁶

Como ya se ha mencionado anteriormente, los contextos definitorios dan pauta para el etiquetado del texto en el corpus de ingeniería, a continuación se describirá a más detalle la importancia de éstos para el corpus.

Un contexto definitorio es todo aquel fragmento textual donde se aporta la información necesaria para definir a un término. Los contextos definitorios constituyen un paso en la elaboración de una herramienta para la identificación

⁶ Alarcón Rodrigo y Sierra Gerardo "Hacia la extracción automática de conceptos" VIII Simposio Iberoamericano de terminología. Cartagena, Colombia, 2002.

automática de los posibles conceptos de un texto especializado, esto es, los términos y sus definiciones.

Es común encontrar, en los contextos definitorios, elementos estilísticos y sintácticos empleados por lo autores. Estos elementos sirven para resaltar los constituyentes de los contextos definitorios. Estos elementos pueden ser marcas tipográficas o bien predicaciones pragmáticas o predicaciones verbales. En algunos casos, las marcas tipográficas, al igual que las predicaciones verbales, funcionan como enlace entre el término y la definición.

Una predicación pragmática se refiere a otro tipo de información relevante para la comprensión del término que un contexto definitorio puede contener. Las predicaciones verbales sirven para unir directamente al término con su definición.

La tipografía de un texto sirve al lector como ayuda visual para identificar fácilmente algún elemento importante y diferenciarlo del resto del texto común, por ejemplo los términos y sus definiciones.

3.2.2 Requerimientos de etiquetado⁷

En este apartado se describirán tres tipos de requerimientos del corpus para el etiquetado de éste que permitirán extraer la información de contextos definitorios de los documentos. Estos tipos son: etiquetado de textos, etiquetado de partes de la oración y el etiquetado parser.

- **Etiquetado del texto.**- Puesto que la información conceptual es caracterizada típicamente por marcadores tipográficos, este primer tipo de

⁷ Sierra Gerardo, Medina Alfonso, Alarcón Rodrigo, Aguilar César A. & Martínez Ismael. "Towards the Extraction of Conceptual Information from Corpora". Corpus Linguistics 2003. Lancaster, 2003.

etiquetado señalará la ocurrencia de estos marcadores (por ejemplo, un término se encuentra en una viñeta y su definición en el párrafo siguiente).

- **Etiquetado de partes de la oración.**- El segundo tipo de etiquetas marcará las partes de discurso para identificar elementos constitutivos de predicaciones verbales y pragmáticas (por ejemplo, en español la predicación verbal "se definen como" consiste en el pronombre se, el verbo conjugado y un adverbio, que anuncian una definición).
- **Etiquetado parser.**- Finalmente, el tercer tipo indicará estructura de la frase a través del parsing para determinar modelos de la formación de términos y definiciones (por ejemplo, que un término puede consistir en una frase nominal más una frase preposicional, mientras una definición puede empezar con una frase nominal cuantificada).

Teniendo en cuenta estos tres tipos de etiquetas, es posible construir búsquedas complejas en las puede identificarse la información de la estructura de los conceptos en texto de especialidad (los términos y sus definiciones correspondientes).

El etiquetado de las partes de la oración y parser se ha trabajado mucho en el área de ingeniería lingüística⁸. Si bien el etiquetado del texto se ha dejado a los criterios de cada uno de los propietarios de los corpus, por las características específicas que presentan para la extracción de contextos definitorios, resulta necesario etiquetar partes del texto que usualmente son omitidas en otros corpus. Por ello, el objetivo de esta tesis es centrarse básicamente en el etiquetado del texto (capítulo 6).

⁸ Márquez Lluís, Padró Lluís & Rodríguez Horacio (1998), "Etiquetado Morfosintáctico de Corpus Textuales". Congreso Anual de la Asociación Española de Lingüística Aplicada (AESLA'98).

3.3 Clasificación del corpus de ingeniería

Una vez definidos los tipos de corpus y sus características, en el capítulo anterior, podremos ahora delimitar y dar la definición del corpus en ingeniería. La figura 3.1 muestra un cuadro sinóptico con la clasificación del corpus en ingeniería.

El corpus lingüístico en ingeniería es:

Origen de los textos: *Textual*.- Como solo se tomarán los textos de publicaciones en ingeniería y no se tomará nada hablado, el corpus en ingeniería es únicamente textual.

Especificidad en los textos: *corpus especializado o específico*.- Se especializa en el área de ingeniería únicamente, comprendiendo todas las ramas que comprenden a esta área.

Según el lenguaje: *corpus monolingüe*.- Esto se refiere a que el corpus en ingeniería sólo será para el español y no contemplará otro idioma.

Cantidad de texto: *corpus textual*.- El corpus en ingeniería se basará únicamente en textos, ya sea de libros, revistas, informes, memorias, y todo material escrito en ingeniería que proporcione información valiosa para el corpus.

Codificación y anotación: *corpus codificado o anotado*.- Es codificado ya que al corpus se añadirán etiquetas con el lenguaje XML para reconocer algunos de sus elementos en los documentos de interés que se hablará en los capítulos 5 y 6.

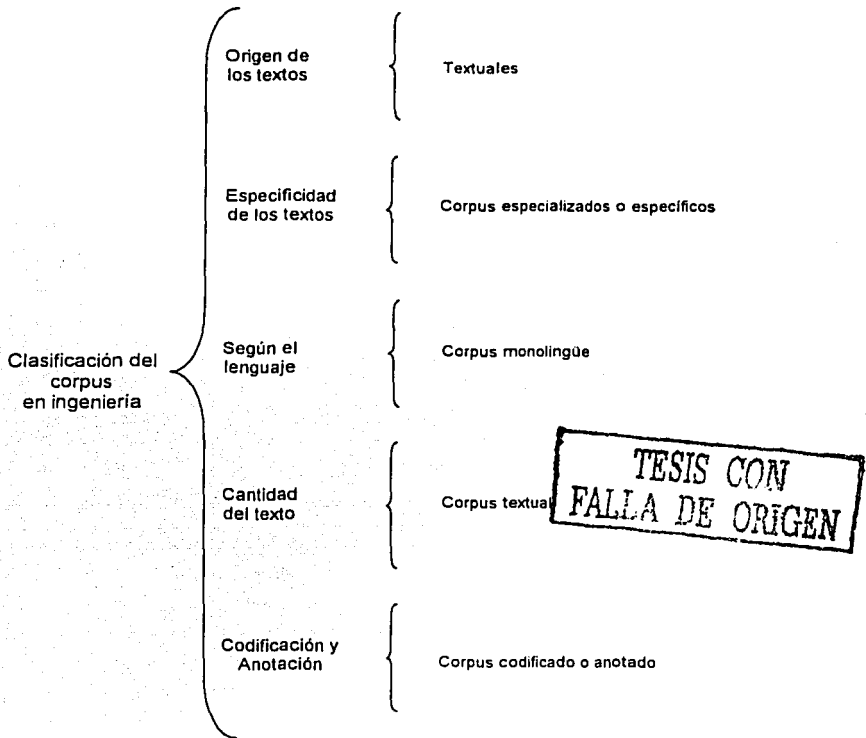


Figura 3.1. Cuadro sinóptico de la clasificación del corpus en Ingeniería

3.4. Aspectos técnicos del corpus de ingeniería

Como ya se ha mencionado, el corpus lingüístico se constituirá con documentos del área de ingeniería. En general, su arquitectura es de la siguiente forma: el usuario solicita la información de su interés por medio de Internet, una

vez hecha la solicitud se hará la petición al servidor web para dar respuesta al usuario. Los documentos almacenados que conforman el corpus están en archivos planos (tipo texto) debidamente etiquetados, mediante el motor de búsqueda y recuperación de la información, muestra los resultados de la consulta hecha por el usuario.

La figura 3.2 presenta el esquema general del funcionamiento del corpus lingüístico, cabe mencionar que en los apartados siguientes, se describirá a más detalle el funcionamiento del corpus.

Con lo que respecta a los documentos, la información que se almacenará, además del contenido del documento, se irá ampliando con anotaciones descritas anteriormente, todas ellas utilizando código SGML/XML. Para ello, se empleará el Corpus Encoding Standard (CES) propuesto por el EAGLES (The Expert Advisory Group on Language Engineering Standards), una iniciativa de la Comunidad Europea, dentro del DG XIII *Linguistic Research and Engineering programme*⁹.

3.5 Definición de descriptores y códigos

El corpus está etiquetado con el lenguaje XML, el cual permite dar formato para describir la información como datos estructurados, así mismo, permite ver, manipular y extraer los términos dentro del corpus a la web. De igual forma, facilita declaraciones de contenido precisas y resultados de búsquedas significativos en varias plataformas. En el capítulo siguiente se hablará más acerca de este lenguaje.

Se utilizará el lenguaje XML para el etiquetado y posterior extracción de términos dentro del corpus en ingeniería, debido a que XML es un lenguaje que

⁹ <http://www.cs.vassar.edu/CES>

ofrece un formato para la descripción de datos estructurados, facilitando declaraciones más precisas y resultados de búsquedas más significativos. XML

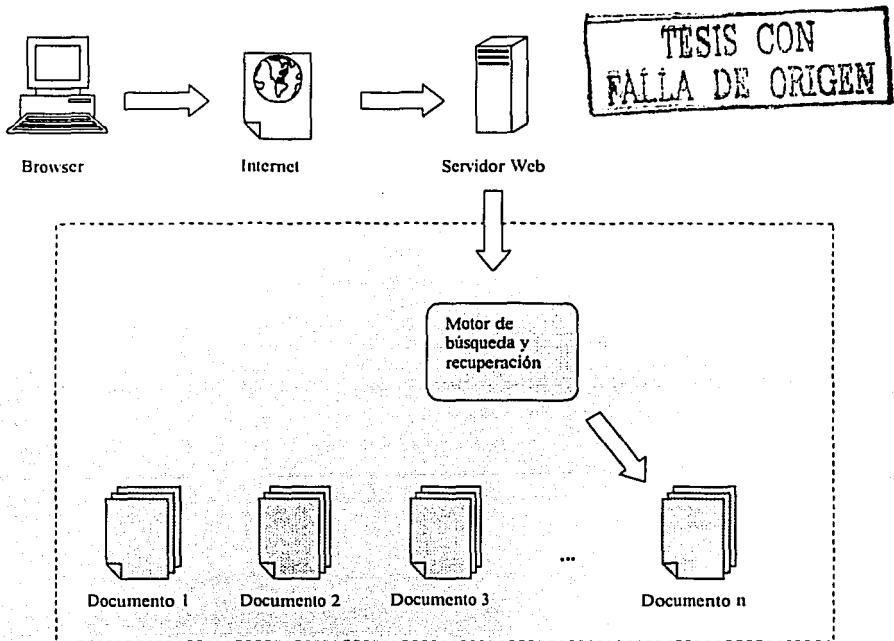


Figura 3.2. Arquitectura del corpus de ingeniería

XML es un metalenguaje, es decir, un lenguaje para definir lenguajes. Además de ser de bajo nivel, esto es, que solamente es a nivel de aplicación y no de programación como lo es HTML, pero se tiene la ventaja que XML permite el intercambio de información estructurada entre diferentes plataformas (Se puede

usar en bases de datos, editores de texto, hojas de cálculo, y casi cualquier cosa que podamos imaginar).

3.6 Criterios de selección de descriptores y códigos en la selección de información y la captura

Mediante XML se asigna una etiqueta única a cada uno de los rasgos lingüísticos del texto, pero también se proporciona un registro de la información metalingüística. Por ejemplo, la etiqueta "encabezado" puede contener sobre el tipo y procedencia del texto. De acuerdo con Atkins (1992)¹⁰, el encabezado de cada entrada del corpus contendría información sobre la fuente del texto, que identifica y diferencia un texto de otro, así como información de descripción del texto, que da cuenta del tipo de documento, la rigurosidad y prestigio, el área de la ingeniería, etc. Por otro lado, la codificación del texto mismo debe diferenciar los párrafos y los distintos tipos de párrafo, así como información sobre distintos tipos de texto, de formato y de fuentes. Esta diferenciación detallada obedece a uno de los usos que se va a dar al corpus. En el grupo de ingeniería lingüística interesa diferenciar entre tamaños, tipos y efectos de letra como, (negrita, Itálicas), el uso de viñetas, sangrías, etc, la definición de las etiquetas se describirán en los capítulos 5 y 6.

3.7 Diseño de búsqueda

El diseño de búsqueda se realiza en la interfaz usuario-sistema, entendiendo por sistema el corpus lingüístico, el cual se presenta de manera

¹⁰ Atkins, Sue, Jeremy Clear and Ostler Nicholas. 1992. "Corpus Design Criteria". pp.1-16, literary and linguistic computing, Volume 7, Number 1, Oxford University Press.

entendible y de fácil manejo para el usuario, ya que el corpus va dedicado no sólo a los ingenieros en particular, sino a todo aquel que se interese por el estudio del lenguaje en el área de la ingeniería. En éste caso, el usuario solamente se encargará de insertar la palabra o término que desea buscar, y seleccionar algunas características particulares de dicha palabra, como podría ser el tema, la rama de ingeniería, seleccionar también el lugar donde se desea realizar la búsqueda, o en qué tipo de textos (libros, revistas, artículos, etc.) se busca la palabra.

3.8 Diseño de salida

El corpus será capaz de mostrar al usuario los campos solicitados en la entrada, obteniendo él únicamente la información que de manera particular es de su interés.

De ésta manera, todo el procedimiento de búsqueda será transparente para el usuario, obteniendo única y exclusivamente, como salida, los campos que seleccionó al inicio de la búsqueda.

El corpus lingüístico lo realizará ordenadamente, de tal forma que el usuario pueda hacer uso de esa información permitiendo realizar un reporte de salida. Con esto se logra que el corpus lingüístico sea más amigable para el usuario.

3.9 Diseños y tipos de reportes

Los diseños y tipos de reportes es la presentación final de la información obtenida de la búsqueda, permitiendo al usuario imprimirlos o salvarlos en disco,

con el propósito de brindar mayor comodidad a las personas que usan el corpus lingüístico. Cabe resaltar que, por los derechos de autor del corpus, sólo podrán tenerse reportes selectivos de los textos, tales como los índices de frecuencias y otros datos estadísticos, tablas de concordancias con una ventana de 10 palabras a cada lado de la palabra seleccionada, etc.

3.10 Definición de procedimientos para la formulación de búsquedas

Se realizarán diferentes procedimientos de búsquedas. Por un lado, el usuario puede buscar por palabras determinadas y con ello obtener los datos de salida, tales como concordancias, colocaciones, datos estadísticos, etc. De las palabras puede también filtrar la información, con áreas temáticas, fechas, etc.

El sistema permitirá extraer el contexto del cual fue obtenido el término y la definición. Se ha observado que en los textos especializados, cuando se define un nuevo término, se utiliza una serie de operadores metalingüísticos (ya descrito anteriormente), tales como el uso de comillas, letras itálicas, viñetas, etc, así como frases que introducen el término que se va a definir y la definición. El conocimiento de estos operadores facilitan el trabajo terminológico. Por esta razón, que los contextos definitorios sean de gran importancia para el corpus. Con esto, no sólo se facilitará la recopilación de futuros términos en el corpus, sino, en general, para cualquier obra terminológica.

3.11 Diseño de las formas de interacción con el público usuario final

El ciclo de vida del corpus lingüístico será permanente ya que serán incorporados nuevos textos continuamente.

El corpus tendrá una interfaz amigable para el usuario. Esto se logrará haciendo que el corpus pueda ser visto con cualquier tipo de navegador de Internet. Se va a poder consultar el corpus y a ejecutar las herramientas desde cualquier tipo de plataforma (UNIX, Windows, Solaris, etc.). El usuario podrá hacer uso de la base sin ninguna complicación, ya que podrá estar regresando a las páginas vistas anteriormente sin perder la información, y tendrá la información necesaria en cada página para poder hacer las consultas que requiera. Para las dudas con el corpus o con el uso de las herramientas, se contará con un correo electrónico para que puedan mandar sus comentarios o sugerencias en cualquier momento y que los administradores de la base puedan resolverlas.

Por tanto, es suficiente que el usuario cuente únicamente con una PC que tenga conexión a Internet y un navegador de cualquier tipo (Netscape, Internet Explorer, etc.) para poder tener acceso al corpus lingüístico.

3.12 Descripción de la operación de corpus lingüístico

En lo correspondiente a la producción del corpus lingüístico:

- Se crearán las bases necesarias, con el espacio requerido para almacenar los textos que constituyen el corpus.
- Una vez creado el corpus, se podrán efectuar consultas.

A cada usuario se le va a proporcionar un login y un password para que puedan acceder a la base. Cada clave va a tener sus permisos necesarios dentro de la base, por lo que los usuarios no van a poder acceder si no tienen asignada una clave, así evitamos anomalías en la base. Estos permisos van a ser:

- Únicamente de consulta (para toda la comunidad ajena al desarrollo de la base),
- de consulta, agregar y modificar (para usuarios con derechos de dar de alta y modificar nuevos textos, pero sin derecho a eliminar), y
- de control total (para los administradores de la base de datos).

Para poder estar actualizando la base de datos, primero necesitamos tener los textos disponibles que se van a introducir. Después de tener los textos, los vamos a digitalizar y a etiquetar con códigos XML. Al tenerlos digitalizados y etiquetados vamos a proceder a introducirlos en el corpus y finalmente vamos a realizar los respaldos correspondientes en cintas magnéticas. Estos pasos se tienen que hacer de la misma manera, cada vez que se tenga un texto nuevo y se quiera colocar en el corpus.

Las consultas al corpus se van a estar haciendo por medio de scripts, que van a estar funcionando correctamente para evitar consultas innecesarias. Las consultas van a depender de qué acción requiere ejecutar el usuario. Lo que hacen las consultas es tomar el comando que lleva la instrucción del usuario, conectarse con el servidor, enviar el comando al servidor, el servidor ejecuta el comando y genera la respuesta, regresa la respuesta el servidor y se desconecta el servidor de la conexión que se hizo.

3.13 Definición de procedimientos de validación, control de calidad, mantenimiento y actualización del corpus lingüístico

Para tener un buen control de calidad en el corpus, éste debe contener cero errores, cero defectos, ofrecer el beneficio a los usuarios para la que está diseñada y, finalmente, cumplir con las expectativas de los usuarios para su plena satisfacción. Para verificar que es capaz de soportar fuertes cargas de trabajo, se

van a realizar pruebas exhaustivas en las que se someta la base a una sobrecarga como: accesos de varios usuarios a la base y hacer consultas al mismo tiempo con el fin de hacer que se sobrecargue el corpus.

La interacción de consulta entre el corpus y el usuario va a ser de una manera sencilla, haciendo que el corpus pueda ser visto con cualquier tipo de navegador de Internet y ejecutarse en cualquier plataforma. El usuario podrá regresar a las páginas vistas anteriormente sin perder la información, va a tener la información necesaria para poder hacer las consultas. Se va a contar con un correo electrónico para que puedan mandar sus comentarios o sugerencias.

Los scripts involucrados en los accesos al corpus serán programados de forma óptima para que no haya accesos que no sean necesarios y provoquen que se tarde en dar respuesta el servidor. Se pondrán a pruebas y revisiones exhaustivas para evitar código repetitivo y proporcionar el mejor servicio a los usuarios.

Para el control del corpus, el administrador del mismo debe estar enterado de quién está agregando, modificando o eliminando parte del corpus, por lo mismo tendrá un estricto control de quiénes tienen derecho de hacer este tipo de manejos en el corpus. Para la validación en el corpus se tienen dos fases:

- Al hacer la programación de los scripts, por cada uno que se tenga, siempre se corroborará la autenticación de la contraseña del usuario, la cual si no es válida, el usuario no podrá seguir haciendo uso de la base.
- El manejador del corpus tendrá su propio medio de autenticación. Con ayuda de éste, se darán los permisos necesarios, según el usuario que esté accediendo al corpus. El propio manejador impedirá el acceso a lugares o hacer ejecuciones no válidas para el usuario.

Cabe denotar que, como el tipo de búsquedas será por palabras, ésta tendrá un texto asociado y es lo único que los usuarios podrán consultar por Internet, por lo que los permisos que se pidan serán únicamente por el uso de los textos asociados únicamente y no del material completo.

El mantenimiento del corpus será periódico con el fin de mantenerlo en su forma óptima y así evitar que las búsquedas en el mismo sean muy tardadas y mermen la respuesta del servidor.

La actualización del corpus se irá haciendo conforme se vaya obteniendo más información. Los respaldos necesarios se irán haciendo en cinta magnética conforme se vaya agregando nueva información.

3.14 Criterios generales para seleccionar información para el corpus

La selección de documentos se basará en el análisis y propuesta de profesionales en el área de ingeniería. Las fuentes básicas de documentación son: informes, revistas, artículos, publicaciones. Se buscará en todo momento mantener un balance entre las diferentes áreas, con base en los criterios de representatividad y diversidad de la lingüística de corpus.

Deberá, además, hacerse una selección cuidadosa de las áreas que se cubrirán, este análisis debe considerar incluso las sub-áreas a incorporar al corpus.

Para la recolección de los textos para elaborar el corpus se realizará un proceso de selección de los mismos. Se recurrirá a todo documento impreso que se considere relevante en el área.

La documentación elegida se digitalizará por medio de un scanner, y se empleará un reconocedor óptico de caracteres (OCR, por sus siglas en inglés) para transformar las imágenes obtenidas en texto legible; posteriormente se procederá a realizar una revisión ocular de la información, de modo que los errores del OCR sean corregidos. De esta manera, se contará con un texto susceptible de almacenarse en un archivo plano. Dicho archivo se almacenará en un directorio con una estructura bien definida. Los recursos empleados para esta labor serán financiados por el presente proyecto.

Cabe mencionar que actualmente el Instituto de Ingeniería cuenta con un gran acervo bibliográfico y con un conjunto de publicaciones en el área de ingeniería (se estima que se cuenta con más de mil publicaciones propias). Estas publicaciones constituirán la base inicial del corpus.

La información almacenada en los archivos tipo texto debe contener las anotaciones pertinentes lingüísticas (morfológicas y sintácticas) y sobre el formato, por lo que se empleará, como ya se mencionó, el estándar empleado para el etiquetamiento, que se basará en XML.

Una vez hecha la codificación de la terminología se podrá vaciar la información para el corpus lingüístico.

CAPITULO 4

INTRODUCCIÓN AL ETIQUETADO CON XML

Este apartado sólo da una visión global al lenguaje XML con el objetivo de aportarle lo básico del lenguaje para apreciar las perspectivas de aplicación y las descripciones de herramientas analizadas en este trabajo. Se presentan así mismo, algunas ideas que definen el lenguaje evitando entrar en los detalles.

El lenguaje XML (XML, eXtensible Markup Lenguaje) sirve para la representación digital de los documentos y es actualmente el lenguaje más prometedor para almacenar y suministrar información a través de la World Wide Web.

Aunque el lenguaje de hipertexto HTML es actualmente el lenguaje más común para generar páginas web, posee capacidades limitadas para almacenar información. Por el contrario, XML tiene una sintaxis altamente flexible, que permite que lo utilicemos para describir virtualmente cualquier tipo de información,

desde una simple receta hasta complejas bases de datos (de ahí el término extensible). Además, un documento XML (junto con una hoja de estilo, o una página convencional HTML) se puede representar fácilmente en un explorador web. Debido a que los documentos estructuran y etiquetan la información que contienen de una manera tan efectiva, el explorador podrá buscar, extraer, filtrar, colocar y manipular esa información de muchas maneras diferentes.

4.1. Historia del XML¹

El XML proviene de un lenguaje que inventó IBM alrededor de los años 70. El lenguaje de IBM se llama GML (General Markup Language) y surgió por la necesidad que tenían en la empresa de almacenar grandes cantidades de información de temas diversos de las áreas en las que se trabajaba e investigaba.

Es por ello que necesitaban una manera de guardar la información y los expertos de IBM inventaron GML, un lenguaje con el que poder clasificarlo todo y escribir cualquier documento para que se pueda luego procesar adecuadamente.

Este lenguaje gustó mucho a la gente de ISO, entidad que se encarga de normalizar los procesos del mundo actual, de modo que por el año 1986 trabajaron para normalizar el lenguaje, creando el SGML, que no era más que el GML pero estándar (Standar en inglés).

SGML es un lenguaje muy trabajado, capaz de adaptarse a un gran abanico de problemas y a partir de él se han creado los siguientes sistemas para almacenar información.

Por el año 1989, para el ámbito de la red Internet, Tim Berners-Lee y Anders Berglund, dos investigadores del laboratorio europeo de física de partículas (CERN), crearon un lenguaje basado en etiquetas para marcar

¹ Morrison Michael, et al. XML al descubierto. Editorial Prentice Hall. España 2000. 4-6

documentos técnicos a fin de compartirlos en Internet. Este lenguaje fue finalmente ampliado en una aplicación simplificada de SGML llamada HTML, que supuso el primer formato de información estándar de la web. Este lenguaje fue adoptado rápidamente por la comunidad y varias organizaciones comerciales crearon sus propios visores de HTML y riñeron entre ellos para hacer el visor más avanzado, inventándose etiquetas como su propia voluntad les decía. Desde el 96 hasta hoy una entidad llamada W3C ha tratado de poner orden en el HTML y establecer sus reglas y etiquetas para que sea un estándar. Sin embargo el HTML creció de una manera descontrolada y no cumplió todos los problemas que planteaba la sociedad global de Internet.

El W3C presentó un equipo de expertos en SGML cuyo objetivo era el de crear una nueva tecnología de marcado con las ventajas nucleares de SGML (extensibilidad, estructura y validación) y con la relativa simplicidad de HTML. El mismo W3C en el 98 empezó y continúa en el desarrollo de XML (Extended Markup Language). En este lenguaje se ha pensado mucho más y muchas personas con grandes conocimientos en la materia están trabajando todavía en su gestación. Pretendían solucionar las carencias del HTML en lo que se respecta al tratamiento de la información. Problemas del HTML como:

- El contenido se mezcla con los estilos que se le quieren aplicar.
- No permite compartir información con todos los dispositivos, como pueden ser ordenadores o teléfonos móviles.
- La presentación en pantalla depende del visor que se utilice.

El código de HTML puede llegar a ser difícil de entender; por ejemplo, para extraer los datos que se necesiten dentro del HTML, se tiene que procesar en otras aplicaciones. Resulta muy difícil saber dónde está realmente la información que se busca, debido a que se encuentra siempre mezclada entre etiquetas , <TABLE>, <TD>, etc. Esto es una mala gestión de la información y el XML la soluciona.

TESIS CON
FALLA DE ORIGEN

4.2. Diferencias entre HTML y XML²

Para diferenciar XML y HTML se propone la siguiente tabla con las características que diferencian a estos lenguajes.

Tabla 4.1. Diferencias entre XML y HTML

El HTML está orientado a la presentación de datos. Se preocupa por formatear datos y para ello son las etiquetas que tiene el lenguaje, para formatear la información que se desea mostrar. Define un conjunto de etiquetas y atributos válidos, una utilización válida de estos elementos y un significado visual para cada elemento del lenguaje.	El XML orientado a los datos en sí mismos se preocupa por estructurar la información que pretende almacenar. La estructura la marca la lógica propia de la información. XML no define las etiquetas ni cómo se utilizan, sólo define unas pocas reglas sintácticas para crear documentos. Por eso XML es un metalenguaje (un lenguaje para definir otros lenguajes).
El desarrollo del HTML estuvo marcado por la competencia entre los distintos visores del mercado. Cada uno quería ser el mejor e inventaba etiquetas nuevas que a la larga entraban a formar parte del estándar del W3C, como la etiqueta <FRAME>.	El desarrollo del XML está siendo llevado a cabo con rigor, siempre ajustado a lo que marca el estándar que desarrolla el W3C, entidad que está desarrollando el XML con más diligencia que las empresas con intereses particulares.
Procesar la información en HTML es inviable, por estar mezclada con los	En XML se puede procesar la información con mucha facilidad,

² Charles F. Goldfarb, Manual de XML. Editorial Prentice Hall. España 2001.
Young J. Michael, Aprende XML ya. Editorial McGraw Hill. España 2001. pp 3-17.

estilos y las etiquetas que formatean la información.	porque todo está ordenado de una manera lógica, así mismo el formateo de la información para que se pueda entender bien por el usuario es viable a través de un pequeño procesamiento, a través de hojas de estilos o similares.
HTML sirve para presentar información en páginas web.	XML sirve para representar e intercambiar datos, independientemente de su presentación

Una de las preguntas que se hacen actualmente es que si ¿Sustituye XML a HTML? No, pues sirven para cosas distintas como ya se ha observado en la tabla anterior. XML y HTML son complementarios.

HTML sigue siendo el principal lenguaje utilizado para indicar a los navegadores cómo representar la información en la web. En lugar de sustituir a HTML, XML se utiliza, como ya se ha mencionado, de manera conjunta con HTML y amplía enormemente la capacidad de las páginas web para:

- Suministrar virtualmente cualquier tipo de documentos
- Ordenar, filtrar, reorganizar, localizar y manipular información de cualquier manera.
- Presentar información altamente estructurada

En otras palabras, XML fue diseñado para interoperar con HTML

4.3. Objetivos de XML

El lenguaje XML se creó para que cumpliera varios objetivos, dentro de los cuales se encuentran:

- Que fuera idéntico a la hora de servir, recibir y procesar la información que se tiene en HTML, para aprovechar toda la tecnología implantada para este último.
- Que fuera formal y conciso desde el punto de vista de los datos y la manera de guardarlos.
- Que fuera extensible, para que lo puedan utilizar en todos los campos del conocimiento.
- Que fuese fácil de leer y editar.
- Que fuese fácil de implantar, programar y aplicar a los distintos sistemas.

Los siguientes son los diez objetivos oficiales del diseño de XML, enunciados en la especificación oficial de XML, expuesta en el sitio web W3C³.

1. XML se debe utilizar directamente en Internet

XML fue diseñado principalmente para almacenar y suministrar información a través de la web.

2. XML debe admitir una gran variedad de aplicaciones

Aunque su principal objetivo consiste en proporcionar información a través de la web, mediante programas de servidor y navegadores, XML también está diseñado para ser usado con otros tipos de programas. Por ejemplo, XML ya se está empleando para compartir información entre programas financieros, para distribuir y actualizar software y para escribir scripts de voz que puedan enviarse a través del teléfono.

³ <http://www.w3.org/TR/REC-xml>

3. XML debe ser compatible con SGML

XML es un subconjunto de propósito especial de SGML. Una de las ventajas de esta funcionalidad es que las herramientas de software de SGML se pueden adaptar muy fácilmente para trabajar con XML.

4. Debe ser fácil crear programas que procesen documentos XML

Si XML ha de ser práctico, la creación de navegadores y demás programas que procesen documentos XML tendrá que ser muy simple. De hecho, la causa principal de que se formara el subconjunto XML de SGML fue la complejidad existente para generar programas que procesaran documentos SGML.

Los siguientes objetivos de diseño de esta lista están destinados a dar soporte a este objetivo fundamental.

5. El número de funcionalidades opcionales de XML deberá mantenerse en un mínimo absoluto, preferiblemente cero

La existencia de un mínimo número de funcionalidades opcionales en XML hace que sea más sencillo crear programas que procesen documentos XML. La abundancia de funcionalidades opcionales en SGML fue una de las principales causas por las que fue calificado como poco práctico para definir documentos web. Entre las funcionalidades opcionales de SGML estaban la redefinición de los caracteres delimitadores en los marcadores (normalmente < y >), y la omisión del marcador de fin cuando el procesador pudiera averiguar dónde terminaba un elemento. Un programa robusto que procesase documentos SGML tendría que tener en cuenta todas las funcionalidades opcionales, incluso aquellas que apenas se utilicen.

6. Los documentos XML deberán ser inteligibles para los humanos y razonablemente claros

XML está diseñado para convertirse en lengua franca para el intercambio de información entre los usuarios y los programas de todo el mundo. El que sea inteligible por los humanos permite alcanzar este objetivo, posibilitando a las personas (y a los programas de software especializados) la lectura y escritura de documentos XML. Su legibilidad distingue a XML de la mayoría de los formatos propietarios, utilizados en las bases de datos y en los documentos de los procesadores de texto.

Los humanos pueden fácilmente leer un documento XML, dado que está escrito en texto legible y tiene una estructura lógica de tipo árbol. Podemos incrementar la legibilidad de XML utilizando nombres significativos para los elementos, atributos y entidades de nuestros documentos y añadiendo comentarios adecuados (más adelante se explicará lo que son los elementos, atributos, entidades y comentarios).

7. El diseño de XML deberá prepararse rápidamente

XML será estándar viable únicamente si la comunidad de programadores y usuarios lo adopta. Este estándar necesita por tanto ser completado antes de que esta comunidad comience a adoptar estándares alternativos, los cuales tienden a producir las compañías de software a un ritmo vertiginoso.

8. El diseño de XML deberá ser formal y conciso

La especificación de XML está escrita en un lenguaje formal, utilizado para definir lenguajes informáticos y que se conoce como notación EBNF (Extended Backus-Naur Form). Este lenguaje formal, aunque difícil de leer a primera vista, resuelve las ambigüedades y en último término facilita la creación de documentos XML, especialmente del software de procesamiento de XML, con lo que se alienta aún más la adopción de XML.

9. Los documentos XML deberán ser fáciles de generar

Para que XML se convierta en un lenguaje práctico de marcado para los documentos web, no sólo debe ser fácil la creación de programas de procesamiento de XML, sino que también los propios documentos XML tendrán que poder crearse de manera sencilla.

10. La concisión en los marcadores XML tiene una importancia mínima

Para conseguir el objetivo 6, los marcadores XML no deberán ser tan concisos que lleguen a convertirse en crípticos.

4.4 Usos de XML⁴

El XML se puede usar para infinidad de trabajos y aporta muchas ventajas en amplios escenarios. La principal ventaja es que cualquier programa informático trabajará mejor con datos en XML. En esta sección se enumeran algunas de sus aplicaciones. Para ver una lista mucho más amplia de las aplicaciones XML actuales y propuestas, incluyendo descripciones detalladas de cada una de ellas, consulte la página web sobre SGML/XML de Oasis⁵.

- **Comunicación de datos.** Si la información se transfiere en XML, cualquier aplicación podría escribir un documento de texto plano con los datos que estaba manejando en formato XML y otra aplicación recibir esta información y trabajar con ella. La representación de los datos muy simple, fácil de transmitir por la red, estándar. En los últimos tiempos este uso se está haciendo muy popular con el surgimiento de los servicios web.
- **Migración de datos.** Si se tienen que mover los datos de una base de datos a otra sería muy sencillo si las dos trabajasen en formato XML.

⁴ Young J. Michael. Aprenda XML ya. Editorial McGraw Hill. España 2001. pp 3-17.

⁵ <http://www.oasis-open.org>

- **Aplicaciones web.** Hasta ahora cada navegador interpreta la información a su manera y los programadores de la web tienen que hacer unas cosas u otras en función del navegador del usuario. Con XML se tiene una sola aplicación que maneja los datos y para cada navegador o soporte podremos tener una hoja de estilo o similar para aplicarle el estilo adecuado. Si mañana la aplicación que se cree debe correr en WAP solo se tiene que crear una nueva hoja de estilo o similar. Permite separar contenido y presentación, y que los mismos datos se puedan mostrar de varias formas distintas sin demasiado esfuerzo.
- **Estructuración de documentos.** La estructura de árbol de los documentos XML convierte a XML en ideal para marcar la estructura de documentos como novelas, poesía y obras de teatro. Por ejemplo, se puede utilizar XML para marcar una obra de teatro en actos, escenas, oradores, renglones, acotaciones, etc. La marcación XML posibilita que el software represente o imprima el documento con el formato adecuado, localice, extraiga o manipule la información del documento, genere tablas de contenido, resúmenes y sinopsis y maneje la información de cualquier otra forma.
- **Almacenamiento de bases de datos.** Al igual que con los formatos propietarios de bases de datos, se puede utilizar XML para etiquetar cada uno de los campos de información dentro de cada registro de una base de datos. Por ejemplo, se podría etiquetar cada nombre, dirección y número de teléfono dentro de los registros de una base de datos de direcciones. Asignando un nombre a cada elemento de información, se puede representar los datos de diversas maneras y buscar, ordenar, filtrar y procesar los datos de muy distintas formas.

4.5 Fundamentos de la sintaxis de XML⁶

Enseguida se describirá brevemente la sintaxis de XML. El bloque de construcción básico de un documento XML es la entidad, que contiene datos analizados o no analizados sintácticamente. Los datos analizados sintácticamente están compuestos por datos o marcado de caracteres que son procesados por un procesador XML. Los datos no analizados sintácticamente se manejan como texto y no están procesados. En el siguiente ejemplo de datos analizados sintácticamente `<name>` y `</name>` son marcado, mientras que Maximiliano son datos de caracteres:

```
<name> Maximiliano </name>
```

El marcado se utiliza para proporcionar una descripción de la estructura de almacenamiento de un documento (entidades) y la estructura lógica (elementos). XML le permite imponer restricciones al diseño y estructura de un documento, especificando las relaciones que hay entre los componentes de marcado. La sintaxis XML describe esencialmente las construcciones empleadas para definir la estructura y diseño de los documentos, así como las restricciones que esto conlleva. Los documentos XML están diseñados para ser procesados por procesadores XML, razón por la que es imperativo que estos documentos se adhieran a una sintaxis muy rígida.

Un procesador XML es un módulo de software que lee un documento XML y que proporciona acceso a su contenido y estructura. Los procesadores XML normalmente procesan documentos XML en nombre de las aplicaciones. En otras palabras, una aplicación XML emplea un procesador XML para obtener acceso al contenido y a la estructura de los documentos XML. Para el usuario final, una aplicación XML y un procesador XML probablemente no se pueden distinguir. Un ejemplo de aplicación XML es Internet Explorer 5.0 que puede procesar y mostrar

⁶ Morrison Michael, et al. XML al descubierto. Editorial Prentice Hall, España 2000. 7-17.
Young J. Michael. Aprende XML ya. Editorial McGraw Hill. España 2001. pp 3-17.

documentos XML. Bajo la cubierta de Internet Explorer 5.0 hay un procesador XML que maneja el procesamiento de documentos XML en nombre del navegador.

A continuación se exponen los distintos componentes de marcado XML que se soportan en XML 1.0:

- Etiquetas de elemento
- Instrucciones de procesamiento
- Declaraciones de tipos de documento
- Referencias de entidades
- Comentarios
- Secciones marcadas

4.5.1 Etiquetas

Las etiquetas constituyen el componente más evidente de la sintaxis XML y se emplean para describir elementos. Por ejemplo, el elemento Maximiliano del ejemplo anterior está formado por las etiquetas `<name>` y `</name>`. Para mantener las cosas en orden, imagínese el término "elemento" como una pieza lógica de marcado, mientras que "etiqueta" hace referencia a una cadena de texto específica utilizada para representar un elemento de un documento XML.

Las normas que tiene XML son muy simples. Se escribe en un documento de texto ASCII, igual que HTML, y en la cabecera del documento se tiene que poner el texto

```
<?xml version="1.0"?>
```

En el resto del documento se deben escribir etiquetas como las de HTML, las etiquetas que nosotros queramos, por eso el lenguaje se llama XML, lenguaje de etiquetas extendido. Las etiquetas se escriben anidadas, unas dentro de otras.

```
<ETIQ1>...<ETIQ2>...</ETIQ2>...</ETIQ1>
```

Cualquier etiqueta puede tener atributos. Le podemos poner los atributos que queramos.

```
<ETIQ atributo1="valor1" atributo2="valor2"...>
```

4.5.2 Referencias de entidades

Las entidades son los bloques de construcción de los documentos XML, que son entidades en sí mismos y que suelen estar formados por otras entidades a través de referencias de entidades. Las referencias de entidades se usan en XML para asignar alias a piezas de datos. Esencialmente, una referencia de entidad sirve como nombre único para una pieza de datos XML. Por ejemplo, obsérvese lo siguiente:

```
<company> Frank&apos;s Ratchet Service </company>  
<company>Rosenberg&apos;s Shoes & Glass&apos; </company>
```

Las referencias de entidades están compuestas por un ampersand (&) y un punto y coma (;). En este ejemplo, ' y & Glass' son referencias de entidades que sirven, respectivamente, como alias de los caracteres ' y &. Normalmente, un analizador XML analizaría sintácticamente tales caracteres de forma distinta, debido a su papel estructural. No obstante, con las referencias de entidades puede usarlos sin que un analizador XML se interponga causando problemas.

4.5.3 Comentarios

Los comentarios se usan en un documento XML para presentar información que técnicamente no forma parte del contenido de ese documento. Al igual que ocurre con los comentarios en los lenguajes de programación, los comentarios XML se usan para proporcionar descripciones de datos de documentos para provecho del usuario. En otras palabras, los analizadores y aplicaciones XML suelen ignorar los comentarios.

Los comentarios se pueden utilizar en cualquier parte de un documento XML en la que aparezcan datos de caracteres analizados sintácticamente. Los comentarios empiezan con `<!--` y terminan con `-->`. La única limitación a los comentarios es que no se pueden incluir guiones altos (-) en un comentario, ya que entrarían en conflicto con la sintaxis de comentarios XML. Véase el siguiente ejemplo:

`<!-- Inicio de nombre -->`

La información contenida en estos comentarios no forma parte de los datos del documento XML.

4.5.4 Instrucciones de procesamiento

Las instrucciones de procesamiento son instrucciones especiales concebidas para ser usadas por la aplicación que está procesando un documento XML. Los analizadores XML no tienen que hacer nada con las instrucciones de procesamiento, sino que las tienen que pasar a la aplicación. Las instrucciones de procesamiento siempre comienzan con un signo menor que y un signo de interrogación (`<?`) y terminan con un signo de interrogación y un signo mayor que (`?>`). El ejemplo más obvio de instrucción de procesamiento es la instrucción de procesamiento XML:

```
<?xml version="1.0"?>
```

Esta instrucción de procesamiento indica que el documento se basa en la versión 1.0 de XML.

4.5.5 Declaraciones de tipos de documento

Las declaraciones de tipo de documento se emplean en XML para especificar información acerca de un documento, incluyendo el elemento raíz del mismo y la definición de tipo de documento (DTD). La declaración de tipo de documento es muy importante a la hora de establecer si éste es válido o si sólo está bien construido. A continuación se definen las tres tareas que lleva a cabo una declaración de tipo de documento:

- Especificar el elemento raíz del documento.
- Definir elementos, atributos y entidades específicas del elemento (DTD internas).
- Identificar una DTD externa en el documento.

A continuación se muestra un ejemplo que utiliza una declaración de tipo de documento:

```
<!DOCTYPE addressbook SYSTEM "AddressBook.dtd">
```

El elemento raíz del documento es el elemento `addressbook`, que está claramente especificado en la declaración de tipo de documento. La DTD externa del documento, `AddressBook.dtd`, también está claramente referenciada en la declaración de tipo de documento.

4.5.6 Secciones marcadas

Las secciones de datos de caracteres no analizados sintácticamente, o secciones CDATA, se emplean en los documentos XML para bloquear texto que tiene que ser puesto a un lado por un analizador XML. Más específicamente, las secciones CDATA de un documento XML contienen texto que no se quiere analizar sintácticamente como datos de caracteres XML. Se define una sección de código CDATA englobándola entre las cadenas `<![CDATA[y]]>`. A continuación se muestra un ejemplo de sección CDATA:

```
<![CDATA[  
<nombre> Juan Torres </nombre>  
<dirección> Sur 1 numero 15 </dirección>  
]]>  
<ciudad> Mexico </city>  
<estado> DF </estado>
```

En este ejemplo, los elementos nombre y dirección no se reconocen como marcado XML, y los datos que hay en ellos no se reconocen como datos de caracteres analizados sintácticamente, ya que las etiquetas se colocan dentro de una sección CDATA. Las etiquetas nombre y dirección se incluyen en una sección CDATA no analizada sintácticamente, por lo que nunca se analizan sintácticamente. Aunque este ejemplo muestra cómo colocar elementos XML normales en una sección CDATA, es normal usar secciones CDATA para citar una pieza de código XML.

4.6 Normalización de etiquetas para corpus lingüísticos⁷

El proyecto EAGLES se ocupa de la normalización de los recursos lingüísticos y de los sistemas de tecnología lingüística, por lo cual su ámbito de interés se circunscribe de forma más específica al trabajo de quienes aplican la tecnología informática al estudio del lenguaje y, en especial, de quienes tienen como objetivo el desarrollo de sistemas informáticos de procesamiento del lenguaje natural.

El principal objetivo de EAGLES es elaborar, mediante un amplio consenso, recomendaciones y especificaciones para áreas concretas de la tecnología lingüística a partir de los resultados de trabajos en curso en diversas organizaciones del ámbito comunitario y promover su adopción en futuros proyectos.

La propuesta en marcha de EAGLES ha sido posible gracias al compromiso asumido por expertos de más de 30 centros de investigación, empresas, consorcios y asociaciones profesionales de la Comisión europea de aportar su tiempo y esfuerzo al trabajo del grupo. El proyecto está coordinado por el Instituto de Lingüística Computacional de Pisa, y en su Consejo de Administración están representadas asociaciones y organismos de coordinación, también de ámbito europeo, como la Red Europea de Centros de Excelencia en Lenguaje y Habla (ELSNET), el capítulo europeo de la Asociación para la Lingüística Computacional (EACL), la Asociación Europea para la Comunicaciones del Habla (ESCA) y la Asociación Europea para la Lógica, el Lenguaje y la Información (FOLLI).

La normalización en la creación y explotación de corpus lingüísticos requiere, en primer lugar, la definición de un conjunto de parámetros para la clasificación y tipificación de corpus y de textos, ya que, para que un corpus sea

⁷ <http://www.cs.vassar.edu/CES>

realmente útil, es imprescindible que tanto los textos que contiene como el propio corpus puedan ser clasificados dentro de una tipología clara.

En lo referente a las normas de anotación lingüística de corpus, puesto que una normalización demasiado rígida de los sistemas de anotación morfosintáctica o etiquetado no resulta recomendable debido a las diferentes necesidades de cada proyecto, el trabajo de EAGLES se orienta hacia un marco general que permite diseñar esquemas concretos de anotación que sean compatibles. No obstante, además de este marco general, se proponen especificaciones por defecto que pueden ser adoptadas cuando no haya motivos especiales que requieran el desarrollo de esquemas específicos. Este marco general es susceptible de ser extendido para lograr la cobertura de fenómenos específicos de determinada lengua o lenguas. Además permite la adopción de diversos grados de especificidad en la anotación.

CAPITULO 5

ETIQUETADO DE DOCUMENTOS

En este capítulo se describen las etiquetas que identifican los documentos que conforman al corpus de ingeniería. Como ya se ha mencionado en capítulos anteriores, el etiquetado está basado en el lenguaje XML.

Debido a que el corpus de ingeniería estará conformado de documentos especializados en ésta área, (libros, informes, revistas, memorias) es relevante identificar de cada uno de ellos la información contenida en la portada y portadilla, como:

- título
- autor (es)
- editorial
- editores
- fecha y lugar de publicación

- copyright, etc.

Todo esto con el objetivo de que cada una de estas características contenga las etiquetas para hacer la extracción de la información que identifica a cada documento. Es importante resaltar que cada tipo de documento cuenta con información similar, como por ejemplo, un libro tiene autor al igual que una memoria, una revista o un informe, etc., pero cada uno de éstos tiene diferencias que los distingue.

A continuación se describirán a detalle cada una de las etiquetas de los documentos del corpus, para lo cual se ha expuesto en tablas para facilitar su entendimiento; en éstas se describe el dato y la etiqueta que le corresponde a cada uno de los documentos, de igual forma se acompaña a cada una de éstas con un ejemplo.

5.1. Etiquetado de libros

TESIS CON
FALLA DE ORIGEN

Para el etiquetado de los libros fue necesario identificar los datos del libro que se encuentran en la portada y portadilla, de los cuales se identificaron los siguientes:

Tabla 5.1. Etiquetas de libros

DATO	ETIQUETAS	DESCRIPCIÓN
Título	<TITLE> </TITLE>	Nombre del libro
Subtítulo	<SUBTITLE> </SUBTITLE>	Subtítulo del libro
Autor	<AUTHOR> </AUTHOR>	Nombre de cada uno de los autores del libro
Editorial	<EDITORIAL> </EDITORIAL>	Editorial que publicó el

		libro
Traducción	<TRADUCTION></TRADUCTION>	Si es el caso, se registrará quién llevó a cabo la traducción del libro
Título original	<TITLEORIG> </TITLEORIG>	Este se usará si el libro tenía un título original antes de la traducción
Lugar de publicación	<PLACEPUBLI> </PLACEPUBLI>	Lugar en donde se publicó el libro
Fecha de publicación	<DATE> </DATE>	Fecha en la que se publicó el libro
Copyright	<COPYRIGHT> </COPYRIGHT>	Derechos del libro

Por ejemplo, los datos del libro será etiquetados de la siguiente forma:

<BOOK>

<TITLE> Aprenda a programar en XML </TITLE>

<SUBTITLE> El libro que necesita para aprender a programar en XML

</SUBTITLE>

<AUTHOR> Michael J. Young </AUTHOR>

<EDITORIAL> McGraw Hill </EDITORIAL>

<TRADUCTION> VuelaPluma, S.L. </TRADUCTION>

<TITLEORIG> XML Step-by-Step </TITLEORIG>

<PLACEPUBLI> Madrid </PLACEPUBLI>

<DATE>2001</DATE>

<COPYRIGHT> McGraw Hill/Interamericana de España. S.A.U.

</COPYRIGHT>

</BOOK>

Como se puede observar, cada etiqueta identifica a cada elemento que distingue al libro. La etiqueta <book> indica que las etiquetas que se encuentran dentro de ella contiene los datos de los libros.

TESIS CON
FALLA DE ORIGEN

5.2. Etiquetado de informes

Los datos de los informes que interesan identificar para definir sus etiquetas son los siguientes:

Tabla 5.2. Etiquetas de informes

DATOS	ETIQUETAS	DESCRIPCIÓN
Nombre del informe	<TITLE> </TITLE>	Nombre del informe
Número del informe	<NUMBER> </NUMBER>	Número del informe
Nombre del proyecto	<PROJECT> </PROJECT>	Nombre del proyecto al que corresponde el informe
Número de etapa	<NUMBERSTAGE> </NUMBERSTAGE>	Etapa del proyecto
Nombre del autor	<AUTHOR> </AUTHOR>	Nombre de cada autor del proyecto
Datos del autor	<DATAAUTHOR> </DATAAUTHOR>	Datos del autor (es) del proyecto, como pueden ser: institución, puesto, e-mail, etc.
Lugar	<PLACEPUBLI> </PLACEPUBLI>	Lugar donde se publicó o elaboró el informe
Fecha	<DATE> </DATE>	Fecha en que se elaboró o publicó el informe
Patrocinador	<SPONSOR>	Nombre de quien patrocinó la

</SPONSOR>	realización del proyecto
------------	--------------------------

Ejemplo:

<REPORT>

<TITLE> Bases metodológicas y marco conceptual **</TITLE>**

<NUMBER> Informe No 4 **</NUMBER>**

<PROJECT> Estudios para mejorar la confiabilidad del funcionamiento del sistema Cutzamala **</PROJECT>**

<NUMBERSTAGE> Etapa 1 **</NUMBERSTAGE>**

<AUTHOR> Sierra G

<DATAAUTHOR> Ayudante de investigador **</DATAAUTHOR>**

</AUTHOR>

<AUTHOR> Gelman O

<DATA_AUTHOR> Investigador titular **</DATA_AUTHOR>**

</AUTHOR>

<AUTHOR> García E

<DATAAUTHOR> Becario de doctorado **</DATAAUTHOR>**

</AUTHOR>

<PLACEPUBLI> Instituto de Ingeniería UNAM **</PLACEPUBLI>**

<DATE> Octubre 1992 **</DATE>**

<SPONSOR> Comisión Nacional del Agua **</SPONSOR>**

</REPORT>

En este ejemplo, se observa que cada etiqueta, al igual que el de libros, tiene un principio y un fin, conteniendo en cada uno de éstas la información de un informe. Así mismo, este ejemplo contiene tres autores y por lo consiguiente tres datos de autor, por lo que se abre y cierra una etiqueta por cada uno de estos datos.

La etiqueta <report>, indica que lo que se encuentra dentro de ella son los datos de los informes.

TESIS CON
FALLA DE ORIGEN

5.3. Etiquetado de memorias

Los datos de interés de las memorias que se etiquetarán se muestran a continuación:

Tabla 5.3. Etiquetas memorias

DATOS	ETIQUETAS	
Nombre del autor	<AUTHOR> </AUTHOR>	Nombre de quien escribió el artículo
Nombre del artículo	<TITLE> </TITLE>	Nombre del artículo de la memoria
Título ó nombre del evento	<TITLEEVENT> </TITLEEVENT>	Nombre del evento en donde se recopiló la memoria
Nombre de la sociedad organizadora	<NAMEORG> </NAMEORG>	Nombre de la sociedad que organizó el evento
Fecha de impresión	<DATE> </DATE>	Fecha donde se realizó la impresión de la memoria
Lugar de impresión	<PLACEPUBLI> </PLACEPUBLI>	Lugar en la que se hizo la impresión de la memoria
Tomo	<NUMBER> </NUMBER>	Número de tomo de la memoria
Páginas	<PAG> </PAG>	Páginas donde está publicado el artículo

Lugar del evento	<PLACE> </PLACE>	Lugar en donde se realizó el evento
Fecha del evento	<DATEEVENT> </DATEEVENT>	Fecha en la que se llevó a cabo el evento
Editores	<PUBLISHING> </PUBLISHING>	Editores de la memoria
Copyright	<COPYRIGHT> </COPYRIGHT>	Derechos de la memoria

Ejemplo:

<PROCEEDING>

<AUTHOR> Fernando Torres </AUTHOR>
 <TITLE> Estado actual de la sismología </TITLE>
 <TITLEEVENT>X congreso nacional de ingeniería sísmica
 </TITLEEVENT>
 <NAMEORG> Sociedad Mexicana de Ingeniería Sísmica, A.C.
 </NAMEORG>
 <DATE> 1993 </DATE>
 <PLACEPUBLI> México </PLACEPUBLI>
 <NUMBER> </NUMBER>
 <PAG> 20 a 35 </PAG>
 <PLACE> Puerto Vallarte, Jalisco </PLACE>
 <DATEEVENT> 8-11 de octubre, 1993 </DATEEVENT>
 <PUBLISHING> Dr. Mario Chávez </PUBLISHING>
 <PUBLISHING> M.I. Belzay Martínez Romero </PUBLISHING>
 <COPYRIGHT> Sociedad Mexicana de Ingeniería Sísmica, A.C.
 </COPYRIGHT>

</ PROCEEDING >

Dentro de la etiqueta `<number>` se encuentran los datos etiquetados para identificar una memoria. Como puede observarse, la etiqueta `<number>` se encuentra vacía, ya que, para este ejemplo, la memoria no contiene número de tomo o volumen.

TESIS CON
FALLA DE ORIGEN

5.4. Etiquetado de revistas

Los datos de interés de las revistas para etiquetar y con esto identificarlas, se muestra en la siguiente tabla:

Tabla 5.4. Etiquetas revistas

DATOS	ETIQUETAS	DESCRIPCIÓN
Nombre del autor del artículo	<code><AUTHOR> </AUTHOR></code>	Nombre del autor que escribió el artículo
Título del artículo	<code><TITLE> </TITLE></code>	Título del artículo que está en la revista
Nombre de la revista	<code><NAMEJOURNAL> </NAMEJOURNAL></code>	Nombre de la revista
Lugar de publicación	<code><PLACEPUBLI> </PLACEPUBLI></code>	Lugar donde se hizo la publicación de la revista
Fecha de publicación	<code><DATE> </DATE></code>	Fecha de la publicación de la revista
Volumen o número	<code><NUMBER> </NUMBER></code>	Volumen o número de la revista
Año	<code><YEAR> </YEAR></code>	Año de la revista. Esto es debido a que la división anual de una revista se consigna en volúmenes o

		años. Si es mensual, doce números o fascículos, complementan un volumen o un año ¹
Páginas	<PAG> </PAG>	Páginas donde está publicado el artículo
Editorial	<EDITORIAL> </EDITORIAL>	Editorial que publica la revista
Editores	<PUBLISHING> </PUBLISHING>	Editores de la revista
Copyright	<COPYRIGHT> </COPYRIGHT>	Derechos de la revista

Ejemplo:

<JOURNAL>

<AUTHOR> Alejandro Castillo Morales </AUTHOR>

<TITLE> Remuneraciones de la industria </TITLE>

<NAMEJOURNAL> Obras </NAMEJOURNAL>

<PLACEPUBLI> México </PLACEPUBLI>

<DATE> Diciembre 2002 </DATE>

<NUMBER> 360 </NUMBER>

<YEAR> XXIX </YEAR>

<PAG>38,41,42,43,44,47,49 </PAG>

<EDITORIAL> Expansión </EDITORIAL>

<PUBLISHING> Arturo Villegas Rodríguez </PUBLISHING>

<COPYRIGHT> Expansión S.A. de C.V. </COPYRIGHT>

</JOURNAL >

¹ Olea Franco Pedro. Manual de técnicas de investigación documental para la enseñanza media. Esfinge, México 1993. pp 84-87.

Este ejemplo contiene los datos que identifican a una revista, los cuales se encuentran dentro de la etiqueta journal que se ha definido.

5.5 Estructura del etiquetado de los documentos

Como se mencionó en el capítulo anterior, un documento XML debe estar bien formado; por ello, se declaran las DTD correspondientes al etiquetado de los documentos. Una vez ya descritas arriba las características que se deben de etiquetar de los libros, revistas, memorias e informes, el paso siguiente es el de definir el encabezado del documento de éstos en XML.

El encabezado de cada documento depende del tipo de documento que se trate, ya sea libro, revista, memoria o informe. Primero, debe tener las DTD correspondiente al documento, en donde se haga referencia a un archivo externo, llamado DTD externo, que contenga las declaraciones específicas para dicho tipo de documento. Y segundo, una vez definidas las DTD, se procede a escribir las etiquetas descritas para los libros, informes, revistas y memorias, tal y como se ejemplifico en secciones pasadas de este capítulo.

Por ejemplo, para el caso particular de una revista, el encabezado del documento será:

Definición de la DTD → `<!DOCTYPE journal SYSTEM "Journal.dtd">`
 Etiqueta que indica el inicio del contenido de la revista → `<JOURNAL>`
 Cabeecera del documento { `<HEADJOURNAL>`
 `<AUTHOR> Alejandro Castillo Morales </AUTHOR>`
 `<TITLE> Remuneraciones de la industria </TITLE>`
 `<NAMEJOURNAL> Obras </NAMEJOURNAL>`
 `<PLACEPUBLI> México </PLACEPUBLI>`
 `<DATE> Diciembre 2002 </DATE>`

Cabecera
del
documento

```
<NUMBER> 360 </NUMBER>
<YEAR> XXIX </YEAR>
<PAG>38,41,42,43,44,47,49 </PAG>
<EDITORIAL> Expansión </EDITORIAL>
<PUBLISHING> Arturo Villegas Rodríguez
</PUBLISHING>
<COPYRIGHT> Expansión S.A. de C.V.
</COPYRIGHT>
</HEADJOURNAL>
```

Etiquetas que
indican el inicio y
fin del texto de la
revista

```
<CONTENT>
AQUI SE COLOCA EL TEXTO ETIQUETADO
</CONTENT>
```

Etiqueta que indica el fin
del contenido de la
revista

→ </JOURNAL>

Esta definición de la DTD irá en el encabezado del archivo XML. Por otro lado, en un archivo externo se escribirá el DTD externo, que muestra los contenidos de los archivos de entidad declarados en el encabezado.

Por ejemplo, el contenido del archivo de entidad Journal.dtd es :

```
<!ELEMENT JOURNAL (HEADJOURNAL, CONTENTJOURNAL)>
```

```
<!ELEMENT HEADJOURNAL (AUTHOR,TITLE,NAMEJOURNAL, PLACEPUBLI,
DATE, NUMBER, YEAR, PAG, EDITORIAL,
PUBLISHING, COPYRIGHT)>
```

```
<!ELEMENT AUTHOR(#PCDATA)*>
```

```
<!ELEMENT TITLE (#PCDATA)>
```

<!ELEMENT NAMEJOURNAL (#PCDATA)>

<!ELEMENT PLACEPUBLI (#PCDATA)>

<!ELEMENT DATE (#PCDATA)>

<!ELEMENT NUMBER (#PCDATA)>

<!ELEMENT YEAR (#PCDATA)>

<!ELEMENT PAG (#PCDATA)>

<!ELEMENT EDITORIAL (#PCDATA)>

<!ELEMENT PUBLISHING (#PCDATA)>

<!ELEMENT COPYRIGHT (#PCDATA)>

<!ELEMENT CONTENTJOURNAL (CONTENT)>

<!ELEMENT CONTENT (#PCDATA)>

Estas definiciones de DTD son las que se hacen referencia en el encabezado del documento XML para el etiquetado de los documentos. El asterisco en AUTHOR significa que un artículo puede tener más de un autor.

En el capítulo siguiente, se describirán ahora las etiquetas de los textos que, a diferencia de los documentos, identificarán los marcadores tipográficos (capítulo 3) que son de interés de identificar para el GIL.

**ESTA TESIS NO SALE
DE LA BIBLIOTECA**

CAPITULO 6

ETIQUETADO DEL TEXTO

En esta sección se describen las etiquetas a utilizar. Para esto, fue necesario realizar un análisis de los textos en ingeniería identificando los patrones metalingüísticos, tales como el uso de comillas, letras itálicas, viñetas, etc., así como frases que introducen el término que se va a definir y la definición, llegando así a la definición de las etiquetas diferenciando estos operadores.

Una vez teniendo los textos en forma digitalizada se etiquetarán con códigos XML. Como ya se ha mencionado, se utilizará el lenguaje XML para el etiquetado y posterior extracción de términos, debido a que XML es un lenguaje que ofrece un formato para la descripción de datos estructurados, facilitando declaraciones más precisas y resultados de búsquedas más significativos.

Cabe mencionar que para llevar a cabo el etiquetado del corpus, es necesario que el texto a etiquetar se guarde en archivo plano (.txt) para poder así

hacer posible que el etiquetado y la extracción de los términos se logre correctamente.

6.1. ETIQUETAS PARA EL TEXTO

Debido a que en el corpus se etiquetarán diversos documentos de ingeniería, como son informes, memorias, libros, revistas, etc., es importante recalcar que el uso de todas las etiquetas no se requerirá en todos los textos debido a que cada uno de los documentos tiene características propias.

Las etiquetas se han clasificado en cuatro secciones:

- **Estructura (elementos del documento).**- Esta clasificación se refiere a los distintos elementos y secciones que integran el documento, por lo que en esta categoría se encuentran: salto de párrafo, encabezamientos (desde 1 hasta n), resumen, bibliografía, texto especial, notas a pie de página, notas a fin de texto, figura, tabla, mapa, título de figura, título de tabla, título de mapa, título de fórmula.
- **Formato (énfasis).**- En ésta se encuentran el formato que presenta el texto del documento, como son: cambio de tipo de letra, cambio de espaciado de letras, letra más grande, letra más pequeña, itálicas, negritas, subrayado, mayúsculas, versales, indentación, viñetas.
- **Referencias.**- Se refiere a toda referencia que exista en el documento, dentro de esta clasificación encontramos: Referencias internas (apéndices, figuras, tablas, capítulos, páginas, etc.), referencias bibliográficas, llamado de pie de página, llamado a fin de documento.

- **Notaciones.**- Esta clasificación se refiere a todas aquellas partes del texto en donde aparecen fórmulas, subíndice, superíndice, notación de fórmula, numeración de fórmula, notación de unidad, siglas y abreviaturas.

Como el lenguaje XML no reconoce los acentos, al inicio del documento XML es necesario poner la palabra reservada `encoding` de la siguiente manera:

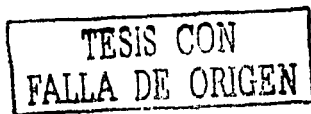
```
encoding="ISO-8859-1"
```

para que al momento de verla en la interfaz del usuario se muestre acentuada tal y como aparece en el texto del documento.

A continuación se da una descripción de las etiquetas que se utilizarán en el etiquetado del texto y un ejemplo de cada una de estas etiquetas.

6.2. Estructura (elementos del documento)

6.2.1. Salto de párrafo



Esta etiqueta indicará el inicio y fin de párrafo; se utiliza `<p>` para el inicio y `</p>` para el fin. Por ejemplo¹:

...forman el olor. Las plantas piloto muestran reducciones del 58 al 83 por ciento, eliminando al mismo tiempo el sabor que pueda tener el agua. El ozono es un bactericida efectivo. Con un contenido de ozono cr itico, las bacterias son prácticamente eliminadas (O'Donovan, 1965). `</p>`

`<p>` El ozono se emplea también para la oxidación de compuestos orgánicos complejos con el objeto de mejorar su adsorción y biodegradabilidad. Sin embargo, estos dos objetivos son incompatibles: la ozonación produce

¹ Vega Gonzalez Eduardo, et al. "Alternativas de tratamiento de aguas residuales" Volumen 1. Proyecto 2321. Informe Instituto de Ingeniería. UNAM. Agosto 1993

compuestos polares que se adsorben menos f ácil pero que tienen un menor peso molecular que los hace más biodegradables. ... </p>

Véase que las etiquetas <p> y </p> sólo marcan el inicio y término del párrafo respectivamente.

TESIS CON
FALLA DE ORIGEN

6.2.2. Encabezamientos (desde 1 hasta n)

El uso de la cabecera (en inglés llamada header) facilita el manejo y procesamiento de los títulos. Lo que redundará en la óptima de los diversos títulos que se manejan dentro de los textos en ingeniería.

Para etiquetar el texto, es necesario identificar el inicio y fin de cada título donde, respectivamente, se colocarán las etiquetas <header> para el inicio del título y </header> para el fin del título, las cuales tendrán un número consecutivo para indicar la descendencia de los header; por ejemplo, se iniciará un <header1> para identificar el primer título, si existe un subtítulo se etiqueta con el <header2> y así sucesivamente. Para ejemplificar² esto veremos un fragmento de texto etiquetado:

<header1>2. PROCESAMIENTO DE SISMOGRAMAS OBSERVADOS DURANTE
LA CAMPAÑA DE 1986 </header1>

<header2>2.1 Localización de eventos sísmicos. </header2>

El objetivo fundamental de la campaña de 1986 fue el de examinar la sismotectónica de la región del Istmo de Tehuantepec (Gaulon y Ponce, 1987, Ponce y Gaulon, 1987). Los eventos registrados tuvieron magnitudes $2.5 \leq M \leq 5.6$. Debido a que uno de los objetivos de.....

² Chávez M., Bravo M.A., Gaulon R., Padilla M.G., Ortega R., Pataú G., "Estimación del riesgo sísmico en el centro-sur de México". Proyecto 0751 Instituto de Ingeniería UNAM Abril 1991 p.1

6.2.3. Resumen

Esta etiqueta de resumen se refiere al resumen total del documento (si se requiere en cada caso), y no al resumen de cada capítulo del documento, ya sea informes, memorias y todo aquel que se utilice. La etiqueta que le corresponderá es: **<summary>** para iniciar el resumen y **</summary>** para finalizar.

Veamos el siguiente ejemplo³:

<header1> Resumen **</header1>**

<summary> Se presenta el análisis teórico para la instrumentación de un dispositivo cuya estructura propia se emplea como transductor de fuerzas, utilizando un mínimo de mediciones de deformación. La estructura en cuestión forma parte del sistema de enganche de tres puntos de un motocultor de alto despeje. **</summary>**

Obsérvese que la etiqueta de **<summary>** abarca solamente el resumen, pero no el encabezado que, como ya se mencionó anteriormente, va etiquetado con **<header>** que en este caso le corresponde el número 1.

TESIS CON
FALLA DE ORIGEN

6.2.4. Bibliografía

Esta etiqueta de estructura se refiere a la bibliografía que es mencionada en el documento. La bibliografía se distinguirá con la siguiente etiqueta: **<bib>** para iniciar la bibliografía y **</bib>** indica que termina la bibliografía. Veamos el siguiente ejemplo⁴:

³ Academia nacional de ingeniería "Memoria del XVII congreso" Monterrey, N.L., México 18-20 de septiembre de 1991

⁴ Barriga Villanueva Rebeca, et al. "La Lingüística en México". 1980-1996. UCLA, México, 1998, pp. 554-581.

BIBLIOGRAFÍA

OBRAS GENERALES



<bib>1. ABELLAN GIRAL, CONCEPCIÓN, "La edición de textos". *Memorias de las Jornadas Filológicas* 1994. UNAM, México, 1995, pp. 405 -408. (*Ediciones Especiales*, 1). **</bib>**

<bib>2. ANTUNEZ, ERASTO, "El eterno descubrimiento de América". *Homenaje a Leonardo Manrique Castañeda*. Coords. Martha Muntzel y Bruna Radelli. Instituto Nacional de Antropología e Historia, México, 1993, pp. 47 -56. (*Colección Científica*, 269). **</bib>**

...
<bib>262. FANDRYCH, CHRISTIAN Y ULRIKE TALLOWITZ PRADE, "Das CELEProjekt". **</bib>**

En este ejemplo podemos observar que la etiqueta **<bib>** y **</bib>** solo abarcan el inicio y fin de la bibliografía (según corresponda), en este caso se observa que la bibliografía consta de 262 libros.

Puede darse el caso que la bibliografía se encuentre dividida en un texto, como sucede en algunas publicaciones en donde la bibliografía viene acompañada de observaciones del autor, de forma que aparece una parte de la bibliografía, luego observaciones del autor y luego otra parte de la bibliografía. En tal caso, se marcará con las etiquetas **<bib>** y **</bib>** en la parte correspondiente de bibliografía, tantas veces como sea necesario. Por ejemplo:

... **<bib>** 395. SILVA GALEANA, LIBRADO, "Un discurso en náhuatl". *Estudios de Cultura Náhuatl*, México, 16 (1993), 219-224. **</bib>**

Véanse, además: III *Gramática*: 309; V. *Linguística Histórica*: 169; IX. *Etnolingüística*: 100 -105, 107 -109, 112, 115, 116, 118, 125, 126; XI. *Sociolingüística*: 501

Como podemos observar, la bibliografía en este caso, viene acompañada de notas del autor, y las etiquetas solo cubrirán la parte de la bibliografía.

TESIS CON
FALLA DE ORIGEN

6.2.5. Texto Especial

El texto especial se refiere a los casos en que se presenten en el texto listas, citas, énfasis, definiciones, etc.; la etiqueta para indicar el inicio de este texto será **<texesp>** y para finalizar **</texesp>**. Por ejemplo⁵:

Ejemplos de trabajos lexicográficos de obras dedicadas a vocablos con marcación diacrítica o diafásica:

<texesp> W. Beinhauer, *Spanische Umgangssprache*, 2ª ed., Bonn, 1958; versión española: *El español colloquial*, 3ª ed., Madrid, 1978.

M. Criado de Val, *Diccionario de español equivoco*, Madrid, 1981.

...

P. M. De Usandizaga y Mendoza, *el chigol* és. *Primer diccionario del lenguaje popular mexicano*, 2ª ed., Méjico, 1973. **</texesp>**

Observemos que el inicio de la etiqueta **<texesp>** se coloca después de la indicación del enlistado de los ejemplos, concluyendo con la etiqueta **</texesp>**.

6.2.6. Notas a pie de página

Éstas se refieren a las notas que aparecen al final de la misma página en la que hace la referencia. Esta etiqueta se distingue de la siguiente forma: **<notefp>** para iniciar la nota y **</notefp>** para cerrar la nota al pie de página. En la parte de abajo aparece un ejemplo⁶ de un texto con una nota al pie de página:

La expresión relativamente reciente de industrias de la lengua' es una denominación político-comercial que sirve para designar un vasto campo de actividad industrial ...

⁵ G. Haensch, et al. "La Lexicografía". Editorial Gredos, Madrid, 1982, pp 144.

⁶ Cabré M. Teresa, "La terminología". IULA, Barcelona, 2000, pp 251

“El término *industrias de la lengua* surge en 1986 en una reunión de tipo político: la primera cumbre de jefes de estado de países...”

Cabe mencionar que para etiquetar las notas al pie de página en el texto se hace de la siguiente forma: al momento de aparecer la marca de la nota al pie de página inmediatamente aparece el texto de la nota que aparece al final de la página.

Ahora veamos el mismo ejemplo pero ahora etiquetado:

La expresión relativamente reciente de *industrias de la lengua* **<notefp>**El término *industrias de la lengua* nace en 1986 en una reunión de tipo político: la primera cumbre de jefes de estado de países... **</notefp>** es una denominación político-comercial que sirve para designar un vasto campo de actividad industrial ...

Como se puede observar ahora la etiqueta **<notefp>** aparece en lugar de la marca de la nota al fin de página y **</notefp>** para finalizar esta nota.

6.2.7. Notas a fin de texto

Esta etiqueta es similar a la anterior. La diferencia es sólo la nota que aparece al final del texto y no de la página. Para indicar el inicio de la nota se utiliza la etiqueta **<endnote>** y para finalizarla **</endnote>**.

Ejemplo⁷ de una nota al fin de texto:

... en cambio la organización de las tres primeras manifiesta la “polisemia categorial” que el término presenta en la terminología (Skujina, 1993:52)’.
...

⁷ IV Simposio Iberoamericano de Terminología. “Terminología y desarrollo”. Buenos Aires Argentina, 17 al 20 de octubre de 1994. pp 107-111.

<p style="text-align: center;">TESIS CON FALLA DE ORIGEN</p>
--

Según esta autora, la "polisemia categorial" es el fenómeno por el cual el contenido de una palabra incorpora ...

Ahora veremos este mismo ejemplo etiquetado:

... en cambio la organización de las tres primeras manifiesta la "polisemia categorial" que el término presenta en la terminología (Skujina, 1993:52) **<endnote>** Según esta autora, la "polisemia categorial" es el fenómeno por el cual el contenido de una palabra incorpora ... **</endnote>**.

Como observamos, la etiqueta **<endnote>** reemplaza la marca de la nota al final del texto indicando el inicio de esta, y **</endnote>** para indicar el final.

6.2.8. Figura

Esta etiqueta indica la aparición de una figura en el texto; debido a que en el desarrollo del corpus no son de importancia las figuras, tablas, fórmulas, (éstas últimas se explicarán más adelante) sino únicamente texto, éstas no se tomarán en cuenta.

Las etiquetas que le corresponden a figura son: **<fig>** indica el inicio de la existencia de una figura dentro del texto y **</fig>** que se refiere al final de la figura. Para ejemplificar esto, véase lo siguiente⁸:

... Se conocen alrededor de 1,500 especies que se clasifican en relación con criterios tales como: tamaño, forma y agrupamiento de células; características de la colonia; reacción a la tinción; requerimientos de crecimiento; movilidad y reacciones químicas específicas. Se encuentran formas aerobias, anaerobias y facultativas (FIG 3.1).

<fig> </fig>

FIG 3.1 Tipos de bacterias

⁸ Vega González Eduardo, et al. Art. Cit., pp 38,39.

TESIS CON FALLA DE ORIGEN

Hongos

Los hongos son protistas eucariontes aerobios, multicelulares, no fotosintéticos y heterotrofos. Algunos hongos son...

Como podemos observar las etiquetas `<fig>` `</fig>` sustituyen la figura que aparecía entre el texto.

6.2.9. Tabla

Similarmente, la etiqueta para indicar el inicio de la tabla es `<table>` y el fin lo indica `</table>`.

Ejemplo⁹:

La TABLA 2.1 muestra algunas de las enfermedades infecciosas, en cuya incidencia puede influir el agua. La causa de estas enfermedades puede tener su origen en bacterias, protozoarios o gusanos. Su control y detención...

TABLA 2.1 PRINCIPALES ENFERMEDADES RELACIONADAS CON EL AGUA
`<table>` `</table>`

Aunado a lo anterior, es importante tener presente que todas las aguas naturales contienen varios contaminantes que provienen de la erosión, la lixiviación ...

Como podemos observar, las etiquetas `<table>` `</table>` abren y cierran enseguida, esto es debido a que en XML cuando se abre una etiqueta es necesario que esta se cierre, ya que de otra forma, no lo reconoce y manda un error conteniendo un atributo nulo; como se había mencionado anteriormente, no interesa al corpus el contenido que tenga la tabla.

⁹ Vega González Eduardo, et al. Art. Cit., pp 16,17.

TESIS CON FALLA DE ORIGEN

6.2.10. Mapa

Se refiere a la aparición de un mapa en el documento. Las etiquetas que le corresponden son: **<map>** para iniciar el mapa y **</map>** para finalizar.

Por ejemplo¹⁰:

... Comenzamos analizando la posible amplificación por efecto de sitio de las estaciones y posteriormente utilizamos las máximas amplitudes para obtener mediante un análisis de mínimos cuadrados la forma de la curva de atenuación.

<map></map>

Mapa2. Localización del sismo ocurrido el 14 de mayo de 1993. El asterisco ...

Las etiquetas correspondientes a la existencia de un mapa para ejemplificar algo descrito en el texto, sustituyen al mapa.

6.2.11. Gráfica

En un documento puede aparecer una gráfica para ejemplificar el tema; por ello se requiere de una etiqueta para que sustituya la aparición de esto. La etiqueta que le corresponde a la gráfica es: **<graphic>** para el inicio de ésta y **</graphic>** para el término.

Veamos el siguiente ejemplo¹¹:

Las gráficas 1 y 2 presentan las curvas de temperatura para los nodos localizados en ...

¹⁰ Sociedad Mexicana de Ingeniería Sísmica A.C. "X Congreso Nacional de Ingeniería Sísmica", Memoria. 8-11 octubre 1993, Puerto Vallarta Jalisco, México. Pp 67

¹¹ Academia Nacional de Ingeniería, A.C. art. cit., pp. 392,393

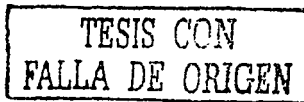
<graphic> </graphic>

Gráfica 1. Curvas de temperatura

<graphic> </graphic>

Gráfica 2. curvas de temperatura

... útil para el tratado de problemas de frontera m ovil. La reformulación de la ecuación ...



Como en este ejemplo existen dos gráficas, se colocan entonces las etiquetas <graphic> </graphic> donde hacen la sustitución de las gráficas.

6.2.12. Título de figura

El título de la figura es el que aparece debajo de una figura; todos los títulos de figura, tabla, mapa, fórmula, y gráfica (éstos últimos se explicarán más adelante) se tomarán en cuenta debido a que es texto y esto es de gran importancia para el corpus.

Las etiquetas que corresponden a título de figura son: <titlefig> para el inicio del título y </titlefig> para el término del título. Veamos el siguiente ejemplo¹²:

... Se conocen alrededor de 1,500 especies que se clasifican en relación con criterios tales como: tamaño, forma y agrupamiento de células; características de la colonia; reacción a la tinción; requerimientos de crecimiento; movilidad y reacciones químicas específicas. Se encuentran formas aerobias, anaerobias y facultativas (FIG 3.1).

<fig> </fig>

<titlefig> FIG 3.1 Tipos de bacterias </titlefig>

¹² Vega González, Eduardo, et al. Art. Cit., pp 38,39.

Como se observa las etiquetas `<fig>` y `</fig>` indican (como ya se explicó anteriormente) la existencia de una figura en el texto, y enseguida está el título que va a acompañado ahora de las etiquetas correspondientes a éste.

6.2.13. Título de tabla

Esta etiqueta marca el título de la tabla, la etiqueta de ésta es: `<titletable>` para el inicio del título y `</titletable>` para el final de éste. Veamos el siguiente ejemplo¹³:

La TABLA 2.1 muestra algunas de las enfermedades infecciosas, en cuy a incidencia puede influir el agua. La causa de estas enfermedades puede tener su origen en bacterias, protozoarios o gusanos. Su control y detención...

`<titletable>` TABLA 2.1 PRINCIPALES ENFERMEDADES RELACIONADAS CON EL AGUA
`</titletable>`

`<table>` `</table>`

Aunado a lo anterior, es importante tener presente que todas las aguas naturales contienen varios contaminantes que provienen de la erosión, la lixiviación ...

En este ejemplo se puede observar que primero está el título de la tabla acompañado de las correspondientes etiquetas, y después las etiquetas de tabla ya que por lo general las tablas aparecen después del título correspondiente.

6.2.14. Título de mapa

El título de un mapa irá acompañado de las etiquetas `<titlemap>` para el inicio del título y `</titlemap>` para el término de este. Ejemplo¹⁴:

¹³ Vega González Eduardo, et al. Art. Cit., pp 16.17.

¹⁴ Sociedad Mexicana de Ingeniería Sísmica A.C. art. Cit., p 67

... Comenzamos analizando la posible amplificación por efecto de sitio de las estaciones y posteriormente utilizamos las máximas amplitudes para obtener mediante un análisis de mínimos cuadrados la forma de la curva de atenuación.

<map></map>

<titlemap> Mapa2. Localización del sismo ocurrido el 14 de mayo de 1993.
El asterisco ... **</titlemap>**

6.2.15. Título de fórmula

El título de fórmula es el que aparece debajo de ésta; las etiquetas para identificar el título son: **<titleform>** para el inicio y **</titleform>** para el final.

Por ejemplo, supongamos que en un texto aparece la siguiente fórmula:

La fórmula para calcular el volumen de un prisma cualquiera tendrá que calcularse a partir de...

$$

<titleform> Fórmula de un prisma **</titleform>**

En este ejemplo, se utilizan dos etiquetas, una de las cuales corresponde al título de la forma que indica el inicio y fin del título. La otra corresponde a la existencia de una fórmula en el texto (esta etiqueta se explicará más adelante).

3.2.16. Título de gráfica

Las etiquetas que indican el inicio y fin de un título de gráfica son: **<titlegraphic>** y **</titlegraphic>** según corresponda:

Ejemplo¹⁵:

Las gráficas 1 y 2 presentan las curvas de temperatura para los nodos localizados en ...

<graphic> **</graphic>**

<titlegraphic> Grafica 1. Curvas de temperatura **</titlegraphic>**

<graphic> **</graphic>**

<titlegraphic> Grafica 2. curvas de temperatura **</titlegraphic>**

... útil para el tratado de problemas de frontera móvil. La reformulación de la ecuación ...

Como se observa en este ejemplo existen dos gráficas y por ello se colocan las etiqueta **<titlegraphic>** y **</titlegraphic>** solo en el inicio y fin del título (según el caso).

6.3. Formato (énfasis)

6.3.1. Cambio de tipo de letra

Esta etiqueta se refiere a cuando en el texto se presenta un cambio de letra, ya sea para resaltar alguna definición o cualquier otro tipo de indicación. Las etiquetas que corresponden al cambio de tipo de letra son: **<changeFont>** para el inicio del cambio y **</changeFont>** para el fin del cambio.

Ejemplo¹⁶:

¹⁵ Academia Nacional de Ingeniería, A.C. art. Cit., pp. 392,393

¹⁶ G. Haensch, et al. Obra citada, p 453

... Este sistema se encuentra, por ejemplo, en le ya mencionado Diccionario Anaya, del que sacamos la siguiente muestra:

<changeFont> tierra (lat. terra) s.f. 1. Planeta que habitamos. 2. Superficie del planeta. 3. Suelo, piso (tropezó y cayó en tierra). 4. País, región **</changeFont>**

Como se muestra en el ejemplo, el cambio del tipo de letra se da cuando se hace referencia al Diccionario Anaya, esto para ejemplificar su texto; por ello, las etiquetas **<changeFont>** **</changeFont>** solo abarcan este cambio.

6.3.2. Cambio de espaciado de letras

Este tipo de etiqueta se aplicará cuando el texto contenga letras espaciadas en su contenido. La representación de esta etiqueta es la siguiente: **<spacing>** para indicar el inicio del cambio de letra y **</spacing>** para el término del cambio de letra. Veamos un ejemplo¹⁷.

...Los dos tipos más usuales de diccionarios con diferenciación cronológica son **<spacing>** el diccionario histórico y el diccionario etimológico **</spacing >**. El primero estudia la trayectoria de una palabra...

Donde la etiqueta **<spacing>** abarca únicamente el inicio y final del cambio de espaciado de las letras.

TESIS CON
FALLA DE ORIGEN

¹⁷ G. Haensch, et al. Obra citada.

6.3.3. Letra más grande

Esta etiqueta identificará la aparición de letra más grande que aparezca en el texto; en ocasiones, la existencia de este tipo de letra puede indicar: ejemplos, resaltar algo, definiciones, y otras más.

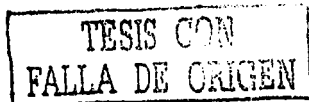
Las etiquetas que indicarán la letra más grande son: **<bigfont>** para el inicio de la letra mas grande y **</bigfont>** para el término.

Ejemplo¹⁸:

... hasta ahora de "escuela", por carecer de fundamentaciones y criterios **<bigfont> propios </bigfont>** en el manejo de la...

Como se observa, las etiquetas que corresponden a la letra más grande, indican el inicio y fin de ésta.

6.3.4. Letra más pequeña



A veces en un texto llega a presentarse un cambio a letra más pequeña, para ejemplificar o hacer referencia a algo, etc.; por ello, es necesario identificar estos tipos de cambio. Las etiquetas que identifican esto son: **<smallfont>** y **</smallfont>**. Ejemplo¹⁹:

... La combinación del orden alfabético con una agrupación por familias de palabras se puede hacer de dos maneras:

<smallfont> Las palabras que forman parte de una familia se ordenan por orden alfabético, después de la "entrada de familia". A esta y cada palabra de la familia corresponde una entrada aparte con un lema puesto **</smallfont>**

¹⁸ Fedor de Diego, Alicia. "La Terminología, teoría y práctica" Equinoccio Ediciones de la Universidad Simón Bolívar, Venezuela, 1995, p.21

¹⁹ G. Haensch, et al obra citada, p. 453

En este ejemplo se observa que la etiqueta `<letterlike>` `</letterlike>` inicia y termina solo en donde aparece la letra más pequeña.

6.3.5. Itálicas

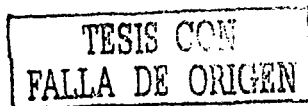
Las letras que estén en itálicas también serán etiquetadas; para diferenciarlas se usa la etiqueta `</i>` para el inicio y `</i>` para el fin de la itálica.

Veamos el siguiente ejemplo²⁰:

... para interrumpir cada una de las ramas identificadas en el DRF, sea a través de las medidas y las actividades orientadas a la `</i>` *prevención*`</i>` de algunas causas, así como por medio de las que buscan la `</i>` *mitigación*`</i>` de sus efectos...

Observamos que las etiquetas `</i>` y `</i>` solo inician y terminan cuando solo aparece la palabra en cursiva.

6.3.6. Negritas



Cuando en un texto existe una letra en negritas, la etiqueta que hace referencia a esto es: `` para el inicio y `` para el fin de las negritas.

Por ejemplo²¹:

1. `` CANTIDAD Y CALIDAD DEL AGUA RESIDUAL ``

²⁰ Oysei Gelman, et al. " Metodología de protección y rescate de obras de almacenamiento: Un caso práctico". Memoria del XIV Congreso Nacional de Ingeniería Civil Sociedades Técnicas. México, diciembre de 1987.

²¹ Vega Gonzalez, Eduardo, et al. Art. Cit., p I.

** 1.1 GENERALIDADES **

El requerimiento Fisiológico básico de agua de una persona es de 2.5 L/día, aunque la carga de trabajo y las condiciones climáticas pueden aumentar bastante esta cifra, más que nada debido a la necesidad de reemplazar el agua perdida por la transpiración. A medida que el nivel de vida mejora, aumenta el uso del agua; esto trae como consecuencia una ...

Como se observa las etiquetas **** y **** sólo se colocan en donde aparece el formato de negritas.

6.3.7. Subrayado

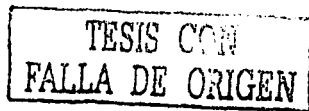
En ocasiones en un texto aparece una o varias palabras subrayadas, lo cual es importante identificar en el corpus; para esto se utiliza las etiquetas **<u>** y **</u>** para el inicio y fin del subrayado según corresponda.

Veamos el siguiente ejemplo²²:

<u> COMPARACIÓN DE ELEMENTOS MECÁNICOS </u>

TRABES

Las tablas 4.20 y 4.21 contienen los momentos flexionantes y torsionantes de algunas vigas tipo de los niveles 1, 5 y 8, para el sismo actuando en las direcciones X y Y, respectivamente. Se hacen comparaciones con y sin los efectos de la interacción suelo-estructura, y de los de las excentricidades torsionales de diseño para las diferentes posiciones de los centros...



Se observa en este ejemplo que la etiqueta **<u>** y **</u>** indican la existencia de palabras subrayadas, las cuales abre y cierran para hacer referencia a esto.

²² Ávila Jorge A., et al. "Ejemplos de aplicación de las normas técnicas complementarias para diseño por sismo. DDF." Ejemplo 2. Proyecto 2527. Informe Instituto de Ingeniería, UNAM. Noviembre, 1994 pp 8/36.

6.3.8. Mayúsculas

En varios textos existen palabras completas e incluso frases enteras con puras letras en mayúsculas, que para el desarrollo de este corpus es necesario identificarlas; por esto, se utiliza la etiqueta **<cap>** para iniciar el texto con mayúscula y **</cap>** para el fin de la mayúscula.

Veamos nuevamente el ejemplo mencionado en la etiqueta anterior:

<u> </u> COMPARACIÓN DE ELEMENTOS MECÁNICOS </u></u>

</c/ >TRABES </c/ >

Las tablas 4.20 y 4.21 contienen los momentos flexionantes y torsionantes de algunas vigas tipo de los niveles 1, 5 y 8, para el sismo actuando en las direcciones X y Y, respectivamente. Se hacen comparaciones con y sin los efectos de la interacción suelo-estructura, y de los de las excentricidades torsionales de diseño para las diferentes posiciones de los centros...

Como se muestra en este ejemplo, la etiqueta de mayúsculas se coloca antes y al terminar las letras en mayúsculas, en este caso, como existieron dos renglones con palabras en mayúsculas, se colocan una por cada renglón.

6.3.9. Versales

Las letras en versales no hay que confundirlas con las mayúsculas, digamos que son un tipo de mayúsculas pequeñas. Éstas comúnmente se utilizan para hacer alguna referencia o para enlistar bibliografía. La etiqueta que distinguirá a las versales es: **<smallcap>** para el inicio y **</smallcap>** para el fin de las versales.

TESIS CON
FALLA DE ORIGEN

Por ejemplo²³:

... 79. **<smallcap>**ERIVIS DURNE, ERIKA**</smallcap>**, "el uso de la preposición a ante objeto directo en el habla popular de la ciudad de México". *Actas del II Congreso Internacional sobre el Español de América, ciudad de México, 27-31 de enero de 1968*. Ed. José G. Moreno de Alba. UNAM, México, 1986, pp. 404-406.

80. **<smallcap>**ELORDUY, MARIA ESTHER**</smallcap>**, "Paralelismos en la distribución de las formas verbales en el sistema temporal. La aplicación de dos formas de descripción al español y al alemán". *Estudios de Linguística Aplicada, México, 1991, núm. 13, 66-77*.

...

En este ejemplo las letras versales las encontramos en una descripción bibliográfica, por ello se colocan las etiquetas correspondientes al inicio y final de donde ocurren.

6.3.10. Cambio de margen

La palabra indentación se tomó de la traducción en inglés indentation. Según el diccionario VOX²⁴ significa: sangrar, sangría. Para este trabajo lo tomamos como cambio de margen, que se presenta para hacer citas, un énfasis, una definición.

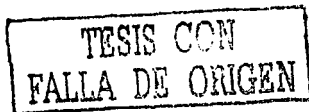
Una característica de la indentación se tiene que el margen es diferente al resto mostrado en un documento. La etiqueta que se usa para la indentación es: **<indentation>** y **</indentation>** para iniciar y finalizar la indentación según corresponda. Por ejemplo²⁵:

... La indicación de equivalentes en el español de América es indispensable en los casos siguientes:
<indentation>

²³ Barriga Villanueva Rebeca, et al, obra citada, p. 79.

²⁴ www.vox.es

²⁵ Haensch G., et al. Obra citada, p. 515



Cuando al equivalente del español peninsular corresponde otro valor denominativo en un área del español de América. Ejemplo: franc es cigarette- español peninsular cigarrillo, pitillo (en Colombia y Venezuela pitillo significa 'paja para sorber bebidas'). **<indentation>**

En este ejemplo se observa que las etiquetas de indentación abarcan en donde comienza y termina ésta.

6.3.11. Viñetas

La etiqueta correspondiente a las viñetas es: **<item>** para iniciar la viñeta y **</item>** para finalizar.

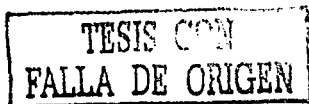
Por ejemplo²⁶:

... lo relaciona y delimita respecto a otros y se manifiesta en el sistema. Estos principios son los siguiente:

<item>

1. la definición es la base para la selección de la denominación de un concepto.
2. la definición dependerá del sistema de conceptos donde se ubique el concepto a definir.
3. las definiciones de los conceptos de un sistema deben ser consistentes entre sí.
4. todos los conceptos utilizados en una definición deben ser conceptos anteriormente definidos... **</item>**

Cabe mencionar que las viñetas se pueden presentar en numeración o, en su caso, con una figura. Se observa en este ejemplo que la etiqueta **<item>** y **</item>** inicia y finaliza donde corresponde para indicar la viñeta.



²⁶Fedor de Diego, Alicia. Obra citada, p.54

6.4. Referencias

6.4.1. Referencias internas

Una referencia interna es aquella que hace referencia a un apéndice, figura, tabla, capítulos, páginas, etc., dentro del texto. Esta etiqueta es: **<reference>** para el inicio y **</reference>** para el final de la referencia.

Ejemplo²⁷:

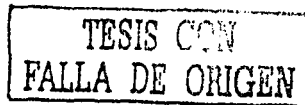
En las **<reference>** FIG 4.1 y 4.2 **</reference>** se presentan algunos ejemplos de diagramas de flujo para el tratamiento del agua residual. Los correspondientes a la **<reference>** FIG 4.1 **</reference>** se emplean comúnmente para pequeñas comunidades, mientras que los de la **<reference>** FIG 4.2 **</reference>** son para grandes comunidades.

En este ejemplo observamos que la etiqueta de referencia abre y cierra en donde en el texto se menciona alguna referencia, en este caso las de unas figuras.

6.4.2. Referencias bibliográficas

La referencia bibliográfica es la que sólo se hace para la bibliografía, y la etiqueta que identificará a este tipo de referencia es: **<refbib>** para el inicio y **</refbib>** para cerrar la referencia.

Por ejemplo²⁸:



²⁷Vega González Eduardo, et al. Art. Cit., p.50

²⁸Fedor de Diego. Alicia. Obra citada, p.66

... o varias de estas etapas a tres niveles de análisis diferentes: el análisis terminológico, el análisis sintagmático y el análisis sintáctico **<refbib>** (Goffin, 1977) **</refbib>**.

En la frase de asimilación, la terminología ayuda al traductor a esclarecer ciertas formas opacas...

En este ejemplo se observa como la etiqueta **<refbib>** marca la referencia que se hace para indicar que es bibliográfica.

6.4.3. Llamado a pie de página

El llamado de pie de página es el número, símbolo o cualquier otro carácter que aparezca en la parte superior del texto para indicar un pie de página. La etiqueta que identificará a este llamado es: **<footnote>** para el inicio y **</footnote>** para el final. Para el mismo ejemplo utilizado anteriormente se tiene²⁹:

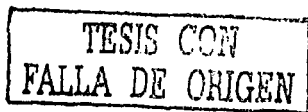
La expresión relativamente reciente de industrias de la lengua **<footnote>**₁**</footnote>****<notefp>**El término industrias de la lengua nace en 1986 en una reunión de tipo político: la primera cumbre de jefes de estado de países ... **</notefp>** es una denominación político-comercial que sirve para designar un vasto campo de actividad industrial ...

Como se observa, el carácter esta acompañado de las etiquetas de llamado a pie de página y después viene el pie de página correspondiente.

6.4.4. Llamado a fin de texto

El llamado a fin de texto son las notas que aparecen al final del documento; estas se pueden utilizar, por ejemplo, para notaciones del autor u otros. Las

²⁹ Cabré M. Teresa., obra citada. p. 251



etiquetas que marcaran a este llamado son: `<notend>` para el inicio del llamado y `</notend>` para el fin del llamado.

... en cambio la organización de las tres primeras manifiesta la "polisemia categorial" que el término presenta en la terminología (Skujina, 1993:52) `<notend>`₁`</notend>` `<endnote>` Según esta autora, la "polisemia categorial" es el fenómeno por el cual el contenido de una palabra incorpora ... `</endnote>`.

6.5. Notaciones

6.5.1. Fórmulas

Para el corpus no interesa la notación matemática de una fórmula, sino simplemente es necesario marcar la existencia de una fórmula en el texto, y la etiqueta es: `<f>` para el inicio y `</f>` para el final de la fórmula. Como por ejemplo³⁰:

Al realizar un balance térmico en la cámara de combustión y considerando la mezcla (aire-productos de la combustión-vapor) sea un gas ideal, se obtiene la ecuación para calcular el calor específico a presión constante de la mezcla.
`<f>` `</f>`

En la figura 7 se ve que al disminuir la carga parcial el calor específico...

En este ejemplo observamos que donde aparece en el texto original la fórmula la sustituimos por la etiqueta `<f>` `</f>` para indicar la existencia de una fórmula en el texto.



³⁰ Academia Nacional de Ingeniería, A.C. art. Cit., p.66

6.5.2. Subíndice

EL subíndice es el número, símbolo o cualquier otro carácter que aparezca en la parte superior de una palabra o una notación en el texto. La etiqueta que le corresponde es: `_{` para el inicio del subíndice y `}` para el final. Por ejemplo³¹:

`u` = desplazamiento horizontal de la base del edificio, debido a la interacción, relativo al movimiento horizontal del terreno en campo libre ...

En este ejemplo observamos que el subíndice aparece en el texto en una notación de fórmula, donde las etiquetas correspondientes marcan al subíndice.

6.5.3. Superíndice

El superíndice será identificado por `^{` para el inicio y `}` para el final; estas etiquetas se colocaran en donde aparezcan una letra, número u otro carácter que sea superíndice.

Por ejemplo³²:

`x` = $u + u_0 + h_j \langle \sup \rangle^\theta \langle \sup \rangle + x_j$, es el de splazamiento horizontal total del nivel j con respecto a un eje vertical

Como se puede observar en este ejemplo aparecen dos superíndices, los cuales van marcados al inicio y final por la etiqueta que le corresponde.

TESIS CON FALLA DE ORIGEN

³¹ Mendoza Otero Enrique. "Influencia de la interacción suelo-estructura en la respuesta sísmica de edificios". Informe del proyecto 6704. Instituto de Ingeniería UNAM, julio 1989, México p 9

³² Mendoza Otero Enrique, "art. Cit., p.9

6.5.4. Numeración de fórmula

La numeración de fórmula se refiere a la aparición a un costado de una fórmula de una numeración, como puede ser solo por números, letra acompañada de un número, u otro que indique el consecutivo de una fórmula.

La etiqueta que hará referencia a esto es: **<fnum>** para el inicio de la numeración y **</fnum>** para el final de la numeración.

Ejemplo³³:

- a) Por equilibrio dinámico de fuerzas horizontales en los n niveles de la superestructura:

$$MR \quad +CA+Kx=0$$

<fnum> (1a) **</fnum>**

Como podemos observar en este ejemplo, la numeración de la fórmula está acompañada de una letra, donde la etiqueta correspondiente inicia y termina, en cada caso, sólo en donde aparece la numeración.

6.5.5. Notación de unidad

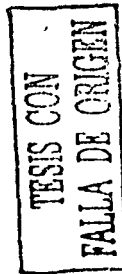
La notación de unidad se refiere a: km, litros, seg, bps, etc., y la etiqueta que la distinguirá es: **<unit>** para iniciar y **</unit>** para terminar la notación.

Ejemplo³⁴:

Si se disminuye la TGET de 1000**<unit>** °c**</unit>** a 800**<unit>** °c**</unit>** y se mantiene constante la potencia en 3000 **<unit>** KW**</unit>**, la cantidad de vapor que se tiene que inyectar es de 0.579 **<unit>** kg/seg**</unit>** (16,

³³ Mendoza Otero Enrique, art. cit., p.7.

³⁴ Academia Nacional de Ingeniería, A.C. art. Cit., p.66



675, 200 **<unit>** kg/año **</unit>**). Este flujo es el 4.8 **<unit>** % **</unit>** del flujo de aire que circula en la TG. Para una TGET de 1000 **<unit>** °C **</unit>**, el flujo de aire es de 12.064 **<unit>** kg/seg **</unit>**.

La etiqueta **<unit>** indica la notación de unidad en el texto.

6.5.6. Notación de fórmula

Este se utilizará para describir los elementos que contiene una fórmula, la etiqueta que identificarán a notación de fórmula es, **<fnotation>** y **</fnotation>** por ejemplo:

$$C_x = \rho V_s A$$

CR ° - - 4 0 PVS1 n (1 - v) 3

donde

<fnotation> CR **</fnotation>** rigidez equivalente a la traslación horizontal de la base de la estructura

Este es un ejemplo dónde aparece una fórmula y después de ésta, aparece la explicación del contenido de ésta, donde, CR es la notación de fórmula.

6.5.7. Siglas

La etiqueta indicará la aparición en el texto de siglas, que se identificará como **<acronym>** para el inicio y **</acronym>** para el final. Ejemplo³⁵:

... al diccionario más extenso de la lengua francesa que se está publicando actualmente, el *Trésor de la langue française* (**<acronym>** TLF **</acronym>**)...

³⁵ G. Haensch, et al. Obra citada., p.154

CAPITULO 7

ADMINISTRACIÓN INTERNA DE LOS DOCUMENTOS DEL CORPUS

Para llevar a cabo el control de los documentos que se van digitalizando para el corpus, se desarrolló una base de datos. Por los intereses del GIL, en este capítulo se muestra la interfaz del usuario y no la elaboración de la base de datos que no es tema de esta tesis.

7.1. Desarrollo de la base de datos del GIL

El objetivo del desarrollo de esta base de datos para el Grupo de Ingeniería Lingüística, es el de almacenar la información de los documentos que se solicitan a los diferentes autores y de las personas del GIL que los están procesando, con el fin de llevar un mejor control de éstos para evitar su duplicidad y conservar su información bibliográfica dentro del Grupo de Ingeniería Lingüística.

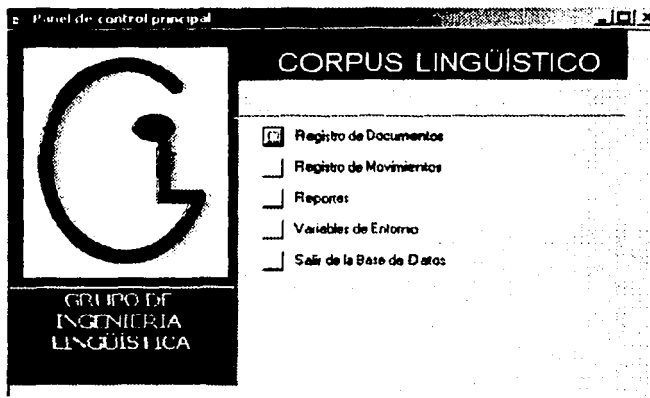
La información de interés para el GIL, como ya se ha mencionado anteriormente, va desde el nombre del libro, memoria, informe ó revista hasta quien patrocina el documento (memoria, informe), así como toda la información contenida en la portada y portadilla de éste. La base de datos fue desarrollada en Microsoft Access XP.

7.2. Contenido de la base de datos

La base de datos tiene un menú principal (figura 7.1) Este menú contiene diferentes opciones, como son:

- Registro de documentos
- Registro de movimientos
- Reportes
- Variables de entorno
- Salir de la base de datos

Cabe mencionar que en las diferentes opciones de toda la base de datos, se tiene un botón con la opción de regresar al menú principal.



TESIS CON
 FALLA DE ORIGEN

Fig. 7.1 Panel de control principal de la base de datos

7.3. Registro de documentos

Esta opción se utiliza para dar de alta ya sean libros, informes, memorias ó revistas y la opción de regresar al menú principal (ver figura 7.2).

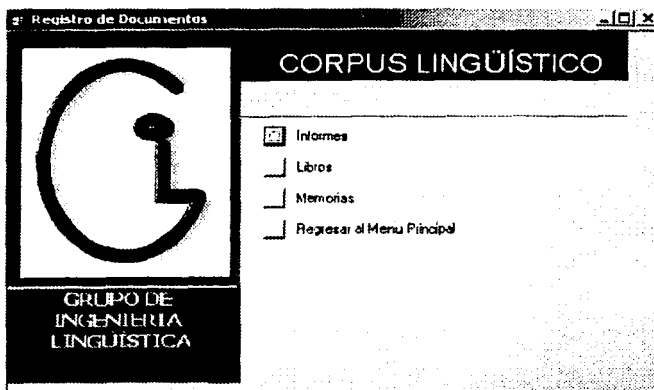


Figura 7.2. Menú de registro de documentos

7.3.1. Llenado de registro de informes

En esta opción, se darán de alta cada uno de los informes que se hayan recopilado para el corpus. Este contiene los siguientes campos (en el capítulo 5 también se han descrito esta tabla):

Tabla 7.1. Llenado de informes

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de registro será automático y dependerá, como su nombre lo indica, del número de informes que contenga
Nombre del informe	Se registrará el nombre del informe
Número de informe	Se ingresa el número del informe
Nombre del proyecto	Nombre del proyecto al que corresponde el informe
Número de etapa	Etapa correspondiente al proyecto
Nombre del autor	Nombre de cada autor del proyecto
Datos del autor	Datos del autor (es) del proyecto, como pueden ser: institución, puesto, e-mail, etc.
Lugar	Lugar donde se publicó o elaboró el informe
Fecha	Fecha en que se elaboró o publicó el informe
Patrocinador	Nombre de quien patrocinó la realización del proyecto

La figura 7.3 muestra la interfaz para el registro de un informe con los datos correspondientes.

TESIS CON
FALLA DE ORIGEN

Informes	
Registro	3
Nombre del informe	Número del informe
Bases metodológicas y marco conceptual	4
Nombre del proyecto	Número de etapa
Estudios para mejorar la confiabilidad del funcionamiento del sistema Cutzamala	1
Nombre del autor (es)	Datos del autor (es)
Sierra G, Gelman O, García E	Ayudante de investigador, Investigador titular, Becario de doctorado
Lugar	Fecha
Instituto de Ingeniería UNAM	Octubre 1992
Patrocinador	Comisión Nacional del Agua

Registro: 1 | 4 | 1 | 1 | de 1

Figura 7.3. Datos de los informes

**TESIS CON
FALLA DE ORIGEN**

7.3.2. Llenado de registro libros

Esta opción se utiliza para llevar a cabo el alta de los diversos libros que se han recopilado para el corpus, y los campos para el llenado de la opción se muestran en la tabla 7.2 (en el capítulo 5 también se han descrito esta tabla).

Tabla 7.2. Llenado de libros

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de registro será automático y dependerá, como su

	nombre lo indica, del número de libros que contenga
Título	Se registrará el nombre del libro
Subtítulo	Subtítulo del libro
Autor	Nombre (s) del autor (es) que publicaron el libro
Editorial	Editorial que publicó el libro
Traducción	Si es el caso, se registrará quien llevó a cabo la traducción del libro
Título original	Se llenará este campo si el libro tenía un título original antes de la traducción
Lugar de publicación	Lugar en donde se publicó el libro
Fecha de publicación	Fecha en donde se hizo la publicación del libro
Copyright	Derechos del libro

Los campos correspondientes a libro se ilustran en la figura 7.4.

7.3.4. Llenado de registro de memorias

TESIS CON
 FALLA DE ORIGEN

Este registro tiene los campos que se muestran en la tabla 7.3 (en el capítulo 5 también se han descrito esta tabla).

Tabla 7.3 Llenado de memorias

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de registro será automático y dependerá, como su nombre lo indica, del número de memorias que contenga
Nombre del autor	Nombre de quien escribió el artículo que se va a digitalizar
Nombre del artículo	Nombre del artículo que se va a digitalizar

Título ó nombre del evento	Nombre del evento en donde se recopiló la memoria
Nombre de la sociedad organizadora	Nombre de la sociedad que organizó el evento
Lugar de impresión	Lugar donde se realizó la impresión de la memoria
Fecha de impresión	Fecha en la que se hizo la impresión de la memoria
Tomo	Número de tomo de la memoria
Páginas	Páginas donde aparece el artículo en la memoria
Lugar del evento	Lugar en donde se realizó el evento
Fecha del evento	Fecha en la que se llevó a cabo el evento
Editores	Editores de la memoria
Copyright	Derechos de la memoria

Libros

Registro

Título Aprenda a programar en XML

Subtítulo El libro que necesita para aprender a programar en XML

Autor (es) Michael J. Young

Editorial McGraw Hill

Traducción VuelaPluma, S.L.

Título Original Step-by-Step

Lugar de publicación Madrid

Fecha de publicación 2001

Copyright McGraw Hill/Interamericana de Esp

Figura 7.4 Registro de libros

TESIS CON FALLA DE ORIGEN

La figura 7.5 muestra los campos que contiene las memorias.

The screenshot shows a window titled 'Memorias' with a list of fields on the left and their corresponding values on the right:

- Registro:** 14
- Nombre del autor:** Fernando Torres
- Nombre del artículo:** Estado actual de la sissr
- Título o nombre del evento:** X congreso nacional de ingeniería sísmica
- Nombre de la sociedad organizadora:** Sociedad Mexicana de Ingeniería Sísmica, A.C.
- Lugar de impresión:** México
- Fecha de impresión:** 1993
- Tomo:** 0
- Páginas:** 20-35
- Lugar del evento:** Puerto Vallarta, Jalisco
- Fecha del evento:** 8-11 octubre 1993
- Editores:** Dr. Mario Chávez, M.I. Belzay Martínez Romero
- Copyright:** Sociedad Mexicana de Ingeniería

At the bottom, there is a status bar that reads 'Registro: 14 de 1'.

Figura 7.5. Campos de las memorias

7.3.5. Llenado de registro de revistas

TESIS CON FALLA DE ORIGEN

Para llevar a cabo el alta de una revista, se requiere el llenado de los campos que se muestran en la tabla 7.4.

Tabla 7.4 llenado de revistas

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de

	registro será automático y dependerá, como su nombre lo indica, del número de revistas que contenga
Nombre del autor del artículo	Nombre del autor del artículo
Título del artículo	Título del artículo a digitalizar
Nombre de la revista	Nombre de la revista
Lugar de publicación	Lugar donde se hizo la publicación de la revista
Fecha de publicación	Fecha de la publicación de la revista
Volumen o número	Volumen o número de la revista
Año	Año de la revista. Esto es debido a que la división anual de una revista se consigna en volúmenes o años. Si es mensual, doce números o fascículos, complementan un volumen o un año ¹
Páginas	Páginas donde se encuentra el artículo
Editorial	Editorial que publica la revista
Editores	Editores de la revista
Copyright	Derechos de la revista

La figura 7.6 muestra la pantalla para dar de alta una revista a la base de datos.

7.4. Registro de Escaneos

La opción de registro de escaneos señala cuándo se da de alta los escaneos y cuando se termina de digitalizar un mismo documento; es decir, cuando la persona quien es la encargada de digitalizar un documento debe

¹ Olea Franco Pedro. Manual de técnicas de investigación documental para la enseñanza media. Esfinge, México 1993, pp 84-87.

registrarlo, para esto, tendrá que ir a esta opción para hacer el llenado del formulario correspondiente la figura 7.7 muestra la pantalla de registro de documentos. Cabe mencionar que cinco de las seis opciones que corresponden a registro de escaneos, contienen la misma información ya sea para registrar un libro, revista, informe o memoria, esto, con el objetivo de facilitar el alta de documentos a el usuario. La última opción corresponde al regreso al menú principal.

Revista	
Registro	
Autor del artículo	Alejandro Castillo Morales
Título del artículo	Remuneraciones de la industria
Nombre de la revista	Obras
Lugar de publicación	México
Fecha de la publicación	diciembre 2002
Volumen o número	360
Año	29
Páginas:	38, 41, 42, 43, 44, 47, 49
Editorial	Expansión
Editores	Arturo Villegas Rodriguez
Copyright	Expansión S.A. de C.V.

TESIS CON
 FALLA DE ORIGEN

Figura 7.6. Formulario de revistas

Los campos que contiene el formulario para dar de alta un nuevo documento a digitalizar se muestran en la tabla 7.5.

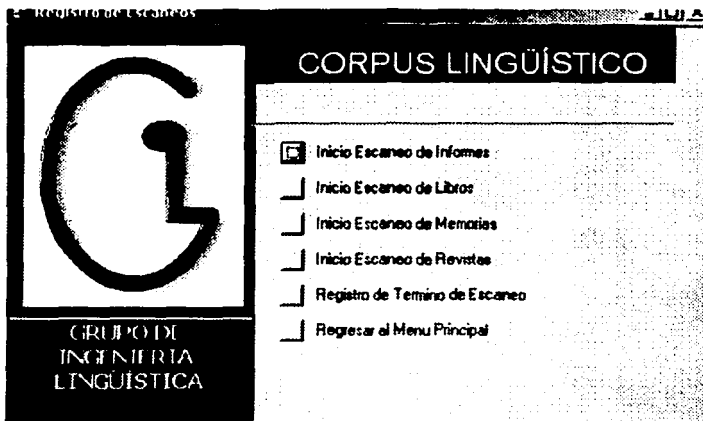


Figura 7.7. Registro de movimientos

Tabla 7.5. Alta de escaneo de documento

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de registro será automático y dependerá, como su nombre lo indica, del número de documentos que contenga
Nombre	Nombre de la persona que va a digitalizar el documento. En esta opción automáticamente aparecerán los nombres de las personas dadas de alta en la base de datos
Fecha de inicio	Fecha de inicio del digitalizado del documento
Nombre del documento	Nombre del documento a digitalizar. Esta opción contendrá automáticamente los diferentes nombres de los documentos que hasta el momento contiene la base de datos, donde el usuario seleccionará el

	nombre correspondiente
Formato del documento	El usuario especificará en que formato se encuentra el documento (pdf, texto, electrónico)

La figura 7.8 muestra la interfaz del llenado del registro para dar de alta un nuevo documento, ya sea libro, revista, memoria o informe.

Figura 7.8 Pantalla de alta de un inicio de escaneo

La tabla 7.6. muestra los campos a llenar para el término de la digitalización de un documento, cabe resaltar que esta opción se utilizará para registrar el término de escaneado para cualquier tipo de documento.

Tabla 7.6. Término de escaneado de un documento

CAMPO	DESCRIPCIÓN
Registro	Cabe mencionar que para esta opción el número de registro será automático y dependerá, como su nombre lo indica, del número de documentos que contenga

Nombre	Nombre de la persona que va a digitalizar el documento. En esta opción automáticamente aparecerán los nombres de las personas dadas de alta en la base de datos
Fecha de término	Fecha de término del digitalizado del documento
Nombre del documento	Nombre del documento que se digitalizó
Formato del documento	Formato del documento digitalizado
Total de páginas digitalizadas	Número total de las páginas que se han digitalizado del documento
Estado del documento	Esta contiene dos opciones: limpio y sucio, donde el usuario seleccionará una de ellas. Limpio se refiere a que el documento se encuentra revisado y listo para ser incorporado al corpus y sucio cuando el usuario no ha hecho ningún cambio y lo ha dejado tal como se quedó después del digitalizado mediante el OCR

Dónde el usuario responsable de ese documento tendrá que dar en esta opción la fecha de término del scaneo del documento. La figura 7.9. Muestra la pantalla de término de digitalización de un documento.

7.5. Reportes

La opción de reportes consiste en presentar el status de los documentos que contiene la base de datos. Este contiene dos opciones: reporte de status de scaneo y regresar al menú principal (figura 7.10). La pantalla que mostrará al seleccionar esta opción se muestra en la figura 7.11.

Registro del Término de Escaneo

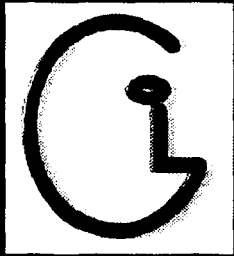
Registro	
Nombre	Karla Ivette Ortega Hernandez
Fecha de inicio	01/06/2002
Fecha Término	16/06/2002
Nombre del documento	Obras
Formato del documento	Texto
Total Pag. escaneadas	120
Estado del documento	suicio

Registro: 14 de 5

Figura 7.9. Formulario término de escaneo de un documento

TESIS CON FALLA DE ORIGEN

Reportes



GRUPO DE INGENIERIA LINGÜÍSTICA

CORPUS LINGÜÍSTICO

- Reporte de status de scaneo
- Regresa al Menu principal

Figura 7.10 Reportes

REPORTE DEL STATUS DEL MATERIAL ESCANEADO

Registro	Nombre	Fecha inicio	Fecha fin	Nombre del documento	Formato documento	Total paginas escaneadas	Estado
1	Karla La Torre Ortega Hernandez	01/06/2002	16/06/2002	Obras	Texto	120	si cob
2	Karla La Torre Ortega Hernandez	01/06/2002	16/06/2002	El arte metodológico y marco c	Texto	120	si cob
3	Karla La Torre Ortega Hernandez	17/06/2002	01/07/2002	Aplicación programar en XML	Texto	110	in po
4	Ambra Estada Treb	14/07/2002	10/08/2002	X congreso internacional de	Electrónico	154	in po
5	Ambra Estada Treb	05/07/2002		El arte metodológico y marco c	Texto	0	

**TESIS CON
FALLA DE ORIGEN**

Figura 7.11 Reporte del status de un documento

Esta pantalla del "Reporte del status del material digitalizado" contiene la información que se presenta en la tabla 7.7.

Tabla 7.7. Reporte de status de digitalizado

CAMPO	DESCRIPCIÓN
Registro	Número de registro correspondiente
Nombre	Nombre de la persona que digitalizó el documento
Fecha de inicio	Fecha de inicio del digitalizado del documento
Fecha de término	Fecha de término del digitalizado del documento
Nombre del documento	Nombre del documento digitalizado
Formato del documento	El usuario especificará en que formato se encuentra el documento (pdf, texto, electrónico)
Total de páginas scaneadas	Número total de paginas que se han digitalizado
Estado	Esta contiene dos opciones: limpio y sucio, según el estado del documento

7.6. Opción de Variables de entorno

En esta opción se pueden realizar las siguientes operaciones:

- Alta de usuarios
- Alta de estado del documento
- Alta de formatos del documento
- Regresar al menú principal

**TESIS CON
FALLA DE ORIGEN**

Estas opciones se muestran en la figura 7.12.

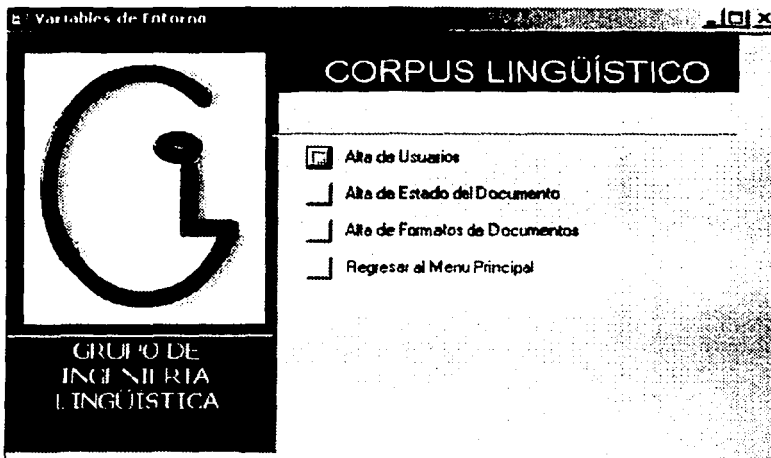


Figura 7.12. Opciones de variables de entorno

TESIS CON
FALLA DE ORIGEN

7.6.1. Alta de Usuarios

El contenido del formulario permite dar de alta a un nuevo usuario que vaya a digitalizar los documentos. Esto se va a realizar cada vez que haya un nuevo usuario para que realice el scaneo. Los datos del formulario de usuario son los que se presentan en la tabla 7.8.

Tabla 7.8. Alta de usuarios

CAMPO	DESCRIPCIÓN
El número consecutivo	El cual asignará automáticamente el número al nuevo usuario que se registre

Nombre completo	Se introducirá el nombre completo del nuevo usuario que va a digitalizar
-----------------	--

La figura 7.13 muestra la ventana del formulario.

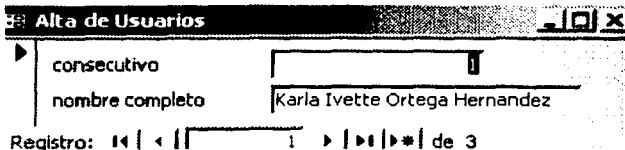


Figura 7.13. Formulario de alta de nuevo usuario

TESIS CON
 FALLA DE ORIGEN

7.6.2. Alta de estado del documento

Este formulario se usa cuando se va a dar un nuevo tipo del estado del documento que se va a dar de alta en la base de datos. Por ejemplo, hasta ahora este formulario cuenta con dos opciones: si el documento esta limpio y la de sucio. Si en un futuro se puede requerir otro estado del documento, en este formulario se dará de alta con el fin de llevar un control de los estados del documento (ver figura 7.14).

Cabe mencionar que el numero consecutivo (como en todos los casos), se asignará automáticamente cuando se realice una alta.

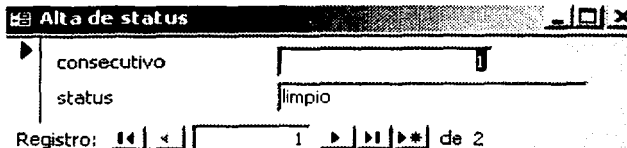


Figura 7.14. Formulario de estado de

7.6.3. Alta de formatos de documentos

Se llena este formulario para dar de alta algún nuevo formato; hasta este momento la base de datos cuenta con los siguientes formatos (fig.7.15):

- PDF
- Texto
- Electrónico

Alta de Formatos

consecutivo

formato PDF

Registro: 1 de 3

Figura 7.15. Alta formatos de documentos

TESIS CON
FALLA DE ORIGEN

CAPITULO 8

CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS

Una vez que se ha presentado el cuerpo básico de la tesis, conviene hacer unas reflexiones que resultan del desarrollo de la misma, así como algunas observaciones que deben considerarse para continuar con el trabajo en el corpus lingüístico en ingeniería.

8.1 CONCLUSIONES DE LA TESIS

En el marco de este trabajo de tesis de licenciatura, se observó que en la formación profesional de la carrera de licenciatura en informática, es desconocida la ingeniería lingüística, la cual constituye un área de aplicación de los conocimientos informáticos adquiridos durante la carrera. Por eso, esta tesis es una buena oportunidad de dar a conocer que hay otra área de aplicación de la informática, que con su constante crecimiento, permite crear nuevas herramientas para el análisis lingüístico.

Para dar pie al desarrollo de esta tesis, se investigó, primeramente, las generalidades del área de ingeniería lingüística y algunos ejemplos de corpus existentes, para que con ello, se conociera todo lo relacionado con ésta y los corpus lingüísticos.

El corpus de ingeniería, dentro de la clasificación y tipos de corpus que existen, se define como un corpus: textual, especializado o específico, monolingüe y codificado o anotado.

De las necesidades fundamentales del GIL para el desarrollo del primer corpus en ingeniería se definió, principalmente, la extracción automática de términos y definiciones en ingeniería mediante la identificación de contextos definitorios. El corpus de ingeniería permitirá hacer tareas que hoy en día no se han realizado, tales como el desarrollo de terminología en el área de ingeniería, redacción de documentos e informes técnicos y corrección de ortografía.

La identificación de los contextos definitorios en los documentos, dan pauta para el etiquetado del texto de los libros, revistas, informes y memorias.

Para llevar a cabo la anotación del corpus, se utiliza el lenguaje XML por ser el lenguaje de marcado, diseñado específicamente para almacenar y

suministrar información a través de la World Wide Web. XML cuenta con una sintaxis muy flexible. Los nombres de los elementos de un documento (como por ejemplo, LIBRO, CUENTO, NOVELA) no forman parte de la definición de XML, sino que se generan al crear un documento particular y se puede elegir cualquier nombre válido para los elementos. Un documento XML está estructurado mediante una jerarquía tipo árbol, con elementos totalmente anidados dentro de otros elementos, y con un único elemento de nivel superior que contiene a todos los demás documentos.

Los documentos especializados en el área que conformarán el corpus de ingeniería son: libros, revistas, memorias e informes. De los cuales se identificaron los elementos que caracterizan a cada uno de éstos, como:

LIBROS	INFORMES	MEMORIAS	REVISTAS
Título	Nombre del informe	Nombre del autor	Nombre del autor
Subtítulo	Número del informe	Nombre del artículo	Título del artículo
Autor	Nombre del proyecto	Nombre del evento	Nombre de la revista
Editorial	Número de etapa	Nombre de la	Lugar de publicación
Traducción	Nombre del autor	sociedad organizadora	Fecha de publicación
Título original	Datos del autor	Fecha de impresión	Volumen o número
Lugar de publicación	Lugar	Lugar de impresión	Año
Fecha de publicación	Fecha	Tomo	Páginas
Copyright	Patrocinador	Páginas	Editorial
		Lugar del evento	Editores
		Fecha del evento	Copyright
		Editores	
		Copyright	

Una vez definidos los elementos a etiquetar de cada uno de los documentos, que permitirá su identificación, para conocer de que documento se extrajo la información, se procedió al análisis, diseño y desarrollo del etiquetado.

En lo que corresponde al etiquetado de los textos, se realizó el análisis correspondiente de dónde se identificaron los patrones metalingüísticos, de interés y necesidad particular para el GIL. El etiquetado de textos se clasificó en cuatro categorías:

Categoría	Elementos
Estructura (elementos del documento)	salto de párrafo, encabezamientos (desde 1 hasta n), resumen, bibliografía, texto especial, notas a pie de página, notas a fin de texto, figura, tabla, mapa, título de figura, título de tabla, título de mapa, título de fórmula
Formato (énfasis)	cambio de tipo de letra, cambio de espaciado de letras, letra más grande, letra más pequeña, itálicas, negritas, subrayado, mayúsculas, versales, indentación, viñetas
Referencias	Referencias internas (apéndices, figuras, tablas, capítulos, páginas, etc.), referencias bibliográficas, llamado de pie de página, llamado a fin de documento
Notaciones	fórmulas, subíndice, superíndice, numeración de fórmula, notación de unidad, siglas y abreviaturas

En el transcurso del desarrollo de esta tesis, se vió la necesidad de almacenar los documentos que se recopilaban para su digitalización; por ello, se realizó una base de datos para llevar a cabo esta administración interna de libros, revistas, memorias e informes. La base de datos fue elaborada en Access XP, la cual permite saber qué documentos se han digitalizado, el nombre de la persona encargada de llevar a cabo la digitalización, fecha de inicio y término de este proceso, el nombre del documento digitalizado, formato de documento (pdf,

electrónico, texto) y el estado en que se encuentra dicho documento (limpio, sucio).

8.2. LINEAS DE TRABAJO FUTURAS

Un trabajo como el realizado en esta tesis es una fuente de líneas de trabajo en el futuro. Se van a destacar estas líneas con el objetivo fundamental de culminar el desarrollo del primer corpus en ingeniería:

- Recabar documentos para el corpus.
- Digitalizar todos los documentos necesarios que permitan contar con una base de datos lo suficientemente amplia y debidamente etiquetados con base en las etiquetas presentadas en esta tesis.
- Desarrollar el diseño de búsqueda, usuario-sistema, que permita al usuario insertar la palabra o término que desea buscar, y seleccionar algunas características particulares de dicha palabra, como el tema, la rama de ingeniería, así como el de seleccionar también el lugar donde se desea realizar la búsqueda, o en qué tipo de textos se buscará la palabra.
- Diseño de salida de la información que el usuario solicitó permitiendo a este poder guardar la información en disco.
- Desarrollar la interfaz con el usuario, esto es, que el corpus pueda ser visto en Internet sobre cualquier plataforma.
- Analizar y desarrollar los scripts necesarios involucrados para los accesos al corpus, los cuales deben ser programados de forma óptima para que no haya accesos innecesarios y que provoquen que se tarde en dar respuesta

el servidor. Para que permitan estar enterados de quién está agregando, modificando o eliminando parte del corpus, mantener un estricto control de quiénes tienen derecho de hacer este tipo de manejos en el corpus.

- La actualización del corpus realizarla conforme se vaya obteniendo más información.

BIBLIOGRAFÍA

Referencias

Atkins, Sue, Jeremy Clear and Ostler Nicholas. 1992. "Corpus Design Criteria". pp.1-16, literary and linguistic computing, Volume 7, Number 1, Oxford University Press.

Bleuca José Manuel, Clavería Gloria, et al "Filología e Informática" ed. Nuevas tecnologías en los estudios filológicos. Seminario de filología e informática. Barcelona 1996. pags. 45,46.

Castillo Hernández Gabriel. "Algoritmo revisado para la extracción automática de agrupamientos semánticos". Tesis. México 2002.

De Yzaguirre, LI. (1996) "Ingeniería lingüística y terminología", *Terminómetro. Monográfico: La terminología en España*, págs. 69-71, Unión Latina-IULA, París.
Morrison Michael, et al. "XML al descubierto". Editorial Prentice Hall, España, 2000.

Goldfarb & Prescod Paul. "Manual de XML". Editorial Prentice, Hall, España, 1999.

Lara Luis Fernando, et al. "Investigaciones lingüísticas en lexicografía". El Colegio de México, Jornadas 89. México 1979. pp. 7-83.

Baldinger K. 1970. Teoría semántica: Hacia una semántica moderna. Madrid: Ed. Alcalá.

Márquez Lluís, Padró Lluís & Horacio Rodríguez (1998), "Etiquetado Morfosintáctico de Corpus Textuales". Congreso Anual de la Asociación Española de Lingüística Aplicada (AESLA'98).

McEnery Tony, Wilson Andrew, "Corpus Linguistics" Edinburg University Press
Koinonia, Manchester 2001. Publicación electrónica:
<http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>

Olea Franco Pedro. Manual de técnicas de investigación documental para la enseñanza media. Esfinge, México 1993, pp 84-87.

Reyes Pérez Antonio "Hacia una obtención computarizada de términos. (aplicación concreta al léxico de la física en el nivel bachillerato). Tesis. México 2002.

Sierra, G, Alarcón, R (2002), "Hacia la extracción automática de conceptos"; en *La Terminología: entre la globalización y la localización*, RITerm, Cartagena, Colombia, formato CD-Rom.

Sierra G. and John McNaught. "Design of an onomasiological search system: A concept oriented tool for terminology" *Terminology* Vol. 6(1), 2000, pp. 1-34.

Sierra Gerardo, Medina Alfonso, Alarcón Rodrigo, Aguilar César A. & Martínez Ismael. "Towards the Extraction of Conceptual Information from Corpora". *Corpus Linguistics* 2003. Lancaster, 2003.

Young Michael J. "*Aprenda XML YA*". Editorial Mc Graw Hill, España, 2000.

Fuentes usadas en los ejemplos

Academia nacional de ingeniería "Memoria del XVII congreso" Monterrey, N.L., México 18-20 de septiembre de 1991 pags. 66, 121,392,393.

Avila Jorge A., et al. "Ejemplos de aplicación de las normas técnicas complementarias para diseño por sismo, DDF." Ejemplo 2, Proyecto 2527. Informe Instituto de Ingeniería, UNAM. Noviembre, 1994 p 8/36.

Barriga Villanueva Rebeca, et al, "La Lingüística en México". 1980-1996. UCLA, México, 1998, pp. 79, 554-581.

Cabré M. Teresa., "*La terminología*". IULA, Barcelona, 2000, pp 251, 251

Chavez M., Bravo M.A., Gaulon R., Padilla M.G., Ortega R., G. Patau, "Estimación del riesgo sísmico en el centro-sur de México" Proyecto 0751 Instituto de Ingeniería UNAM Abril 199, p.1

Fedor de Diego, Alicia. "La Terminología, teoría y práctica" Equinoccio Ediciones de la Universidad Simón Bolívar. Venezuela, 1995, pp. 21, 54, 66

Haensch G., et al., "La Lexicografía". Editorial Gredos, Madrid, 1982, pp 144, 154, 216,453,515,

Mendoza Otero Enrique, "Influencia de la interacción suelo-estructura en la respuesta sísmica de edificios". Informe del proyecto 6704. Instituto de Ingeniería UNAM, julio 1989, México pp. 9, 7

Ousei Gelman, et al. " Metodología de protección y rescate de obras de almacenamiento: Un caso práctico". Memoria del XIV Congreso Nacional de Ingeniería Civil Sociedades Técnicas. México, diciembre de 1987.

Sociedad Mexicana de Ingeniería Sísmica A.C. "X Congreso Nacional de Ingeniería Sísmica", Memoria. 8-11 octubre 1993, Puerto Vallarta Jalisco, México. p 67

Vega Gonzalez Eduardo, et al. "Alternativas de tratamiento de aguas residuales" Volumen 1, Proyecto 2321. Informe Instituto de Ingeniería, UNAM. Agosto 1993, pags, 1,16,17,38,39,50.

IV Simposio Iberoamericano de Terminología, "Terminología y desarrollo". Buenos Aires Argentina, 17 al 20 de octubre de 1994. pp 107-111.

Páginas de Internet consultadas

<http://www.w3c.org/XML/>

<http://www.vox.es>

<http://www.rae.es/>

<http://www.sintx.usc.es/>

<http://www.hltcentral.org/projects/CRATER>

<http://www.iula.upf.es/corpus/corpus.es.htm>

<http://www.hltcentral.org/>

<http://cvc.cervantes.es/obref/agle/prologo/>

<http://www.etde.org/etdeweb/>

<http://iling.torreingenieria.unam.mx/>

http://terminotica.upf.es/membres/DE_YZA/PUBLI/INGE.HTM

http://cvc.cervantes.es/obref/anuario/anuario_98/parte2/cap3//listerri_01.htm

<http://www.cs.vassar.edu/CES>

<http://www.w3.org/TR/REC-xml>

<http://www.oasis-open.org>

<http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>

ANEXO A. ETIQUETADO DE UN DOCUMENTO

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE report SYSTEM "Report.dtd">
<REPORT>
<HEADREPORT>
  <TITLE> Efectos de interacción suelo-estructura en la respuesta sísmica de un
    edificio instrumentado </TITLE>
  <NUMBER> </NUMBER>
  <PROJECT> </PROJECT>
  <NUMBERSTAGE> </NUMBERSTAGE>
  <AUTHOR> David Muria Vila
    <DATAAUTHOR> Investigador Instituto de Ingeniería UNAM
  </DATAAUTHOR>
</AUTHOR>
  <AUTHOR> Ricardo González Alcorta
    <DATA_AUTHOR> Técnico académico Instituto de Ingeniería UNAM
</DATA_AUTHOR>
  </AUTHOR>
  <PLACEPUBLI> Instituto de Ingeniería UNAM </PLACEPUBLI>
  <DATE> Marzo 1992 </DATE>
  <SPONSOR> </SPONSOR>
</HEADREPORT>
<CONTENT>
```

```
<b><header1>1. <cap>INTRODUCCIÓN</cap></header1</b>
<p>En el diseño sísmico de edificios es práctica común aplicar los movimientos
sísmicos a nivel de cimentación y suponer que la estructura descansa sobre un suelo
infinitamente rígido. Sin embargo, se sabe que la deformabilidad del suelo modifica
significativamente las propiedades dinámicas del sistema suelo-estructura, además de
que el suelo funciona como un disipador de energía a través de los amortiguamientos
por radiación y del propio material <reference>(<abrev>ref</abrev>
1)</reference>. </p>
<p>A partir de 1987, diversas instituciones de investigación iniciaron una campaña para
instrumentar edificios desplantados sobre suelo blando y suelo firme, con el fin de
analizar su comportamiento estructural al someterlos a la acción de sismos. Entre los
aspectos importantes por analizar en estas estructuras destaca la posibilidad de definir la
influencia de la interacción suelo-estructura en la respuesta estructural. </p>
<p>En esta investigación se presenta el estudio del comportamiento dinámico del
edificio de la Escuela Secundaria <abrev>N<sup>o</sup></abrev> 3, desplantado en
suelo blando e instrumentado con cuatro acelerógrafos digitales instalados por el Centro
de Instrumentación y Registro Sísmico <acronym>(GIRES)</acronym>. Se analiza la
respuesta de dicho edificio ante cinco eventos sísmicos ocurridos entre 1987 y
1990. </p>
<b><header1>2. <cap>EL EDIFICIO </cap></header1</b>
<p>El edificio consta de tres niveles y se localiza en la zona de terreno compresible. Su
estructuración es a base de marcos formados por columnas y traves de acero, con las
primeras embebidas en concreto reforzado. Se distingue la presencia de muros de
concreto de 15 <unit>cm</unit> de espesor, ubicados en las dos direcciones principales
de la planta de la estructura. El sistema de piso está constituido por traves de acero
```


sobre las cuales se apoya una losa plana de concreto de 10 <unit>cm</unit> de espesor, la cual forma una sección compuesta con la trabe mediante conectores de cortante.</p><p>El edificio está compuesto por dos estructuras, identificadas como cuerpo A y cuerpo B, que se encuentran separados por una junta constructiva del 0 <unit>cm</unit> de espesor. Las dimensiones en planta del cuerpo A son 8 por 26 <unit>m</unit> y las del cuerpo B, 8 por 29 <unit>m</unit>, con una altura de entrepiso de 3.05 <unit>m</unit> en ambos cuerpos. Una vista isométrica y la planta tipo del edificio se muestran en la <reference><abrev>fig</abrev> 1</reference>. Cada uno de los cuerpos se apoya sobre un cajón de cimentación desplantado a una profundidad de 2.50 <unit>m</unit>. Las resistencias nominales de los materiales son 4200 <unit>kg /cm ²</unit> como límite de fluencia del acero de refuerzo y 200 <unit>kg /cm ²</unit> como resistencia del concreto en compresión. El subsuelo contiene arcillas altamente deformables del valle de México.</p>

<header>3. <cap>CARACTERÍSTICAS DINÁMICAS</cap></header><p>Para determinar experimentalmente las propiedades dinámicas de los edificios es común recurrir a pruebas de vibración ambiental, las cuales han mostrado ser aplicables en diferentes tipos de estructuras. Sin embargo, en estructuras rígidas desplantadas en suelos blandos se ha comprobado que dichas pruebas tienen ciertas limitaciones <reference> (<abrev>ref</abrev> 1) </reference>. Por ello, se decidió estudiar en la estructura seleccionada las diferencias que se pueden encontrar en sus modos fundamentales y coeficientes de amortiguamiento correspondientes al aplicar diferentes tipos de pruebas.</p>

<header>3.1 <cap>METODOLOGÍA EXPERIMENTAL</cap></header></p>

<p>Los tres tipos de pruebas que se aplicaron son:</p><indentation>

<p>- Pruebas de vibración ambiental, que consisten en medir las vibraciones ambientales en el edificio y realizar un análisis espectral estadístico de los registros según los criterios establecidos para análisis de señales aleatorias <reference> (<abrev>ref</abrev> 2), </reference> con el fin de lograr espectros típicos del movimiento de la estructura.</p>

<p>Se calcularon espectros de potencia de cada señal, con promedios de 32, 64 y 128 eventos; la duración de cada evento es 12.5 ó 25 <unit>s</unit>.</p>

<p>

<item>

- Pruebas de impulsos, que consisten en medir las vibraciones generadas manualmente por dos o tres personas al jalar con cierta frecuencia una cuerda atada a la azotea del edificio <reference> (<abrev>ref</abrev> 1), </reference>

</p>

</item>

<p>Se sabe que el espectro de potencia de una señal de impulsos periódica se caracteriza por una ordenada espectral significativa asociada a la frecuencia de los impulsos, además de las de sus armónicos cuyas amplitudes disminuyen sucesivamente <reference> (<abrev>ref</abrev> 3) </reference>. Entonces, cuando las pruebas de impulsos se aplican a una estructura, algunas de estas ordenadas crecen al acercarse a la frecuencia de un modo de vibrar y decrecen al alejarse de la misma, lo cual permite identificar sus frecuencias naturales. Se efectuaron pruebas solamente en la dirección transversal.</p>

<item><p>

- Pruebas de tracción, que consisten en proporcionar a la estructura energía potencial, la cual se libera instantáneamente para que se generen movimientos de vibración libre que se atenúan en función de la capacidad de disipación de energía del sistema. La información se registró en el dominio del tiempo mediante un convertidor analógico-digital.

</p></item>

<p>Para estas pruebas se diseñó y construyó un nuevo dispositivo, formado por poleas, cables, yugos y un disparador, el cual permite liberar súbitamente la fuerza de tracción de manera eficiente y segura. En la <reference><abrev>fig</abrev> 2</reference> se muestra el disparador <reference> (<abrev>ref</abrev> 4) </reference> y en la <reference><abrev>fig</abrev> 3</reference>, el dispositivo completo. Para transmitir adecuadamente la fuerza al edificio, el sistema de yugos y cables se fijó en uno de sus extremos a los elementos resistentes de la estructura y en el otro, a los elementos estructurales externos, según se indica en la <reference><abrev>fig</abrev> 4</reference>. Al sujetar los cables y yugos se tuvo el cuidado de mantener la integridad del inmueble; por tanto, se emplearon yugos apropiados y rozaderas. La fuerza se indujo a través de los cables por medio de un tensor mecánico acoplado a un sistema de poleas para aumentar la magnitud de la fuerza sobre la estructura. Se aplicaron siete pruebas de tracción variando la magnitud de la fuerza y su punto de aplicación <reference> (<abrev>figs</abrev> 4 y 5). </reference></p>

</indentation>

<p>En estas pruebas, las señales se captan con acelerómetros localizados en los puntos de medición elegidos y, a través de cables blindados, se transmiten a unos acondicionadores donde se amplifican; se filtran las frecuencias mayores de 30 <unit>Hz</unit>. </p>

<p>Las señales acondicionadas se envían a un analizador de espectros de dos canales, el cual procesa las señales emitidas mediante la transformada de Fourier, lo que permite determinar los espectros de potencia, la función de transferencia (en fase y amplitud) y la coherencia correspondiente. La adquisición de la información procesada se realiza por medio de la tarjeta <abrev><acronym>IEEE</acronym>-488</abrev>, la cual permite transferir los datos del analizador a una microcomputadora tipo <abrev>PC</abrev> <reference> (<abrev>ref</abrev> 5). </reference> Durante las mediciones se colocaron los acelerómetros en puntos diferentes de cada cuerpo y uno en el terreno circundante, como se ilustra en la <reference><abrev>fig</abrev> 6</reference>. </p>

....

<header2>Pruebas de impulsos</header2>

<p>Para identificar la frecuencia natural de vibrar de la estructura en la dirección T se recurrió a las pruebas de impulsos. Los espectros obtenidos se presentan en la <reference><abrev>fig</abrev></reference> 18, donde se observa que la frecuencia fundamental se localiza en el intervalo de 2.6 a 2.9 <unit>Hz</unit>, por estar asociado a las amplitudes máximas y a coherencias mayores que 0.8. <p>

<header2>Pruebas de tracción</header2>

<p>En la última etapa de medición se efectuaron las pruebas de tracción solo en el cuerpo B del edificio. En las <reference><abrev>figs</abrev> 19 a 21</reference> se presentan los registros obtenidos de las pruebas de tracción correspondientes a diferentes puntos de aplicación de la fuerza y los espectros de Fourier asociados a los mismos. La elección de los puntos de prueba tuvo como finalidad excitar los primeros

</f>
<f> KR
1+6H I; 1+2R
(1+0. 7-)</f> <fnum>(2)</fnum>

<f> <fnum>3</fnum>
CR ° ~ - 4 0 P V S I n (1-v)
</f>

<p>donde</p>

<texesp><indentation>
<p><fnotation>KK</fnotation>rigidez equivalente a la traslación horizontal de la base de la estructura</p>
<p><fnotation>KR</fnotation>rigidez equivalente al cabeceo de la base de la estructura</p>
<p><fnotation>CX</fnotation>coeficiente de amortiguamiento de radiación equivalente asociado a la traslación horizontal de la base de la estructura</p>
<p><fnotation>CR</fnotation>coeficiente de amortiguamiento de radiación equivalente asociado al cabeceo de la base la estructura</p>
<p><fnotation>G</fnotation>módulo de rigidez del suelo</p>
<p><fnotation>v</fnotation> relación de Poisson</p>
<p><fnotation>p</fnotation>densidad de masa</p>
<p><fnotation>Vg</fnotation>velocidad de las ondas de cortante</p>
<p><fnotation>A</fnotation>área de la superficie neta de cimentación</p>
<p><fnotation>I</fnotation> I nercia de la superficie neta de cimentación con respecto a un eje trasversal a la dirección de análisis</p>
</texesp></indentation>

.....

</CONTENT>
</REPORT>

CONTENIDO DE LA DTD EXTERNA DE REPORT

<!ELEMENT REPORT (HEADREPORT, CONTENTREPORT)>

<!ELEMENT HEADREPORT (TITLE, NUMBER,
PROJECT,NUMBERSTAGE,AUTOR,PLACEPUBLI,
DATE,SPONSOR)>

<!ELEMENT TITLE (#PCDATA)>

<!ELEMENT NUMBER (#PCDATA)>

<!ELEMENT PROJECT (#PCDATA)>

<!ELEMENT NUMBERSTAGE (#PCDATA)>

<!ELEMENT AUTHOR (#PCDATA | DATAAUTHOR)*>

<!ELEMENT PLACEPUBLI (#PCDATA)>

<!ELEMENT DATE (#PCDATA)>

<!ELEMENT SPONSOR (#PCDATA);>

<!ELEMENT CONTENTREPORT (CONTENT)>

<!ELEMENT CONTENT (#PCDATA)>

ANEXO B. ENCABEZADO DE UN LIBRO

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>  
<!DOCTYPE book SYSTEM "book.dtd">
```

```
<BOOK>
```

```
<HEADBOOK>
```

```
  <TITLE> Aprenda a programar en XML </TITLE>
```

```
  <SUBTITLE> El libro que necesita para aprender a programar en XML
```

```
  </SUBTITLE>
```

```
  <AUTHOR> Michael J. Young </AUTHOR>
```

```
  <EDITORIAL> McGraw Hill </EDITORIAL>
```

```
  <TRADUCTION> VuelaPluma, S.L. </TRADUCTION>
```

```
  <TITLEORIG> XML Step-by-Step </TITLEORIG>
```

```
  <PLACEPUBLI> Madrid </PLACEPUBLI>
```

```
  <DATE>2001</DATE>
```

```
  <COPYRIGHT> McGraw Hill/Interamericana de España S.A.U.
```

```
  </COPYRIGHT>
```

```
</HEADBOOK>
```

```
<CONTENT>
```

```
DENTRO DE ESTA ETIQUETA IRÁ EL TEXTO DEL LIBRO
```

```
</CONTENT>
```

```
</BOOK>
```

CONTENIDO DE LA DTD EXTERNA DE LIBRO

<!ELEMENT BOOK (HEADBOOK, CONTENTBOOK)>

<!ELEMENT HEADBOOK (TITLE, SUBTITLE, AUTHOR,
EDITORIAL, TRADUCTION, TITLEORIG,
PLACEPUBLI, DATE, COPYRIGHT)>

<!ELEMENT TITLE (#PCDATA)>

<!ELEMENT SUBTITLE (#PCDATA)>

<!ELEMENT AUTHOR (#PCDATA)*>

<!ELEMENT TRADUCTION (#PCDATA)>

<!ELEMENT TITLEORIG (#PCDATA)>

<!ELEMENT PLACEPUBLI (#PCDATA)>

<!ELEMENT DATE (#PCDATA)>

<!ELEMENT COPYRIGTH (#PCDATA)>

<!ELEMENT CONTENTBOOK (CONTENT)>

<!ELEMENT CONTENT (#PCDATA)>

ANEXO C. ENCABEZADO DE UNA REVISTA

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>  
<!DOCTYPE journal SYSTEM "Journal.dtd">
```

```
<JOURNAL>
```

```
<HEADJOURNAL>
```

```
<TITLE> Obras </TITLE>
```

```
<NUMBER> 360 </NUMBER>
```

```
<YEAR> XXIX </YEAR>
```

```
<PLACEPUBLI> México </PLACEPUBLI>
```

```
<DATE> Diciembre 2002 </DATE>
```

```
<EDITORIAL> Expansión </EDITORIAL>
```

```
<PUBLISHING> Arturo Villegas Rodríguez </PUBLISHING>
```

```
<COPYRIGHT> Expansión S.A. de C.V. </COPYRIGHT>
```

```
</HEADJOURNAL>
```

```
<CONTENT>
```

```
ESTA ETIQUETA INDICA EL INICIO Y FIN DEL TEXTO DE LA REVISTA
```

```
</CONTENT>
```

```
</JOURNAL>
```


CONTENIDO DE LA DTD EXTERNA DE UNA REVISTA

<!ELEMENT JOURNAL (HEADJOURNAL, CONTENTJOURNAL)>

<!ELEMENT HEADJOURNAL (AUTHOR, TITLE, NAMEJOURNAL,
PLACEPUBLI, DATE, NUMBER, YEAR, PAG,
EDITORIAL, PUBLISHING,COPYRIGHT)>

<!ELEMENT AUTHOR (#PCDATA)*>

<!ELEMENT TITLE (#PCDATA)>

<!ELEMENT NAMEJOURNAL (#PCDATA)>

<!ELEMENT PLACEPUBLI (#PCDATA)>

<!ELEMENT DATE (#PCDATA)>

<!ELEMENT NUMBER (#PCDATA)>

<!ELEMENT YEAR (#PCDATA)>

<!ELEMENT PAG (#PCDATA)*>

<!ELEMENT EDITORIAL (#PCDATA)>

<!ELEMENT PUBLISHING (#PCDATA)>

<!ELEMENT COPYRIGHT (#PCDATA)>

<!ELEMENT CONTENTJOURNAL (CONTENT)>

<!ELEMENT CONTENT (#PCDATA)>

ANEXO D. ENCABEZADO DE UN INFORME

<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>

<!DOCTYPE proceeding SYSTEM "Proceeding.dtd">

<PROCEEDING>

<HEADPROCEEDING>

<AUTHOR> Fernando Torres </AUTHOR>

<TITLE> Estado actual de la sismología </TITLE>

<TITLEEVENT>X congreso nacional de ingeniería sísmica

</TITLEEVENT>

<NAMEORG> Sociedad Mexicana de Ingeniería Sísmica, A.C.

</NAMEORG>

<DATE> 1993 </DATE>

<PLACEPUBLI> México </PLACEPUBLI>

<NUMBER> </NUMBER>

<PAG> 20 a 35 </PAG>

<PLACE> Puerto Vallarta, Jalisco </PLACE>

<DATEEVENT> 8-11 de octubre, 1993 </DATEEVENT>

<PUBLISHING> Dr. Mario Chávez </PUBLISHING>

<PUBLISHING> M.I. Belzay Martínez Romero </PUBLISHING>

<COPYRIGHT> Sociedad Mexicana de Ingeniería Sísmica, A.C.

</COPYRIGHT>

</HEADPROCEEDING>

<CONTENT>

TEXTO DEL CONTENIDO DE UN INFORME

</CONTENT>

</PROCEEDING>

CONTENIDO DE LA DTD EXTERNA DE UN INFORME

<!ELEMENT PROCEEDING (HEADPROCEEDING, CONTENTPROCEEDING)>

<!ELEMENT HEADPROCEEDING (AUTHOR, TITLE, TITLEEVENT,
NAMEORG,DATE,
PLACEPUBLI, NUMBER,PAG,PLACE,
DATEEVENT,PUBLISHING,COPYRIGHT)>

<!ELEMENT AUTHOR (#PCDATA)*>

<!ELEMENT TITLE (#PCDATA)>

<!ELEMENT TITLEEVENT (#PCDATA)>

<!ELEMENT NAMEORG (#PCDATA)>

<!ELEMENT DATE (#PCDATA)>

<!ELEMENT PLACEPUBLI (#PCDATA)>

<!ELEMENT NUMBER (#PCDATA)>

<!ELEMENT PAG (#PCDATA)*>

<!ELEMENT PLACE (#PCDATA)>

<!ELEMENT DATEEVENT (#PCDATA)>

<!ELEMENT PUBLISHING (#PCDATA)>

<!ELEMENT COPYRIGHT (#PCDATA)>

<!ELEMENT CONTENTPROCEEDING (CONTENT)>

<!ELEMENT CONTENT (#PCDATA)>